



基于种子地址的 IPv6 地址探测技术综述

李果^{1,2}, 何林^{1,2}, 宋光磊^{1,2}, 王之梁^{1,2}, 杨家海^{1,2}, 李子木^{1,2}

(1. 清华大学网络科学与网络空间研究院, 北京 100084;

2. 北京信息科学与技术国家研究中心, 北京 100084)

摘要: 现有的扫描技术能够在非常短的时间内对整个 IPv4 空间完成扫描, 但是这些方法都不适合庞大的 IPv6 网络空间。因此近年来, 许多研究学者提出了各种启发式的 IPv6 地址探测算法。这些算法使用收集的 IPv6 种子地址作为输入, 输出最可能活跃的 IPv6 地址列表作为扫描目标, 大大缩小了活跃地址区域的扫描范围。对这些基于种子地址的技术进行了分类、梳理和总结, 然后详细分析了各个算法的优缺点, 并且讨论了这些技术面临的多种挑战。从参考文献公开的 IPv6 数据集和中国教育科研计算机网的北京节点两个来源收集了总共 7 300 万种子地址。通过实验, 比较了 4 种基于种子地址的地址探测算法的命中率和时间性能。最后给出了对本领域的思考和未来研究方向。

关键词: IPv6; 种子地址; 网络扫描; 生成算法

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019296

Survey on IPv6 address scanning technology based on seed sources

LI Guo^{1,2}, HE Lin^{1,2}, SONG Guanglei^{1,2}, WANG Zhiliang^{1,2}, YANG Jiahai^{1,2}, LI Zimu^{1,2}

1. Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China

2. Beijing National Research Center for Information Science and Technology, Beijing 100084, China

Abstract: Nowadays, the state-of-the-art technologies can spend a very short time to scan the whole IPv4 space, but these methods cannot be applied to the huge IPv6 space easily. Therefore, many researchers propose different heuristic algorithms for the sake of IPv6 scanning. The common way of these algorithms is to input collected IPv6 seed addresses and output new most likely active IPv6 addresses as candidates for later scanning. These methods greatly reduce the scanning range of the active address area. These technologies based on seed addresses were classified, analyzed and summarized, and detailed analysis of the advantages and disadvantages of each method was given. And the several challenges faced by the methods were discussed. 73M seed addresses were collected in total from two sources, including published IPv6 datasets in papers and Beijing Node of China Education and Research Network. Through the proposed experiments, time performance and hit rate of four IPv6 address scanning technologies based on seed addresses was compared. Finally, the own thoughts on this field and some future research directions were proposed.

Key words: IPv6, seed address, network scanning, generation algorithm

收稿日期: 2019-11-19; 修回日期: 2019-12-10

通信作者: 杨家海, yang@cernet.edu.cn

基金项目: “十三五”国家重点研发计划资助项目 (No.2017YFB0803004)

Foundation Item: The National Key Research and Development Program during the 13th Five-year Plan Period (No.2017YFB0803004)

1 引言

无论是在安全研究、拓扑发现还是网络诊断方面,互联网扫描都有着非常重要的意义。最先进的针对 IPv4 网络空间的扫描技术已经能够在非常短的时间内对整个网络空间进行扫描。Masscan^[1]能够在 6 min 内对整个 IPv4 地址空间指定端口完成扫描。Durumeric 等^[2]设计的 ZMap 能够在 45 min 内对整个 IPv4 地址空间完成扫描。

随着 IPv4 的地址分配耗尽^[3],下一代互联网协议 IPv6 也在全球范围内开始加速推广和部署。从 1996 年开始,一系列关于 IPv6 的定义和规范 RFC 文件逐步发表,构筑了 IPv6 协议的体系。2012 年,多家知名网站开始永久性支持 IPv6,如 Facebook、Google 等。2019 年最新的 APNIC (Asia-Pacific Network Information Centre, 亚太互联网络信息中心) 统计数据显示 IPv6 部署率最高的两个大国是美国和印度,排名第一的是印度,IPv6 部署率已经达到 64.01%,第二名的美国是 56.44%。国内也有对 IPv6 网络部署的现状统计的报告,2019 年 7 月,推进 IPv6 规模部署专家委员会发布的白皮书《中国 IPv6 发展状况》显示我国 IPv6 活跃用户数已达 1.3 亿、我国基础电信企业分配 IPv6 地址的用

户数达到了 12 亿。

随着近年 IPv6 的迅猛发展,对 IPv6 网络的现状测绘、拓扑发现和安全研究都成为了非常关键的研究问题。其中,最核心的也是最基础的研究手段就是网络探测。相比于 IPv4 的 2^{32} 的地址空间,IPv6 有 2^{128} 的地址空间。IPv6 具有庞大的地址空间的特点使一些传统的探测方法在 IPv6 空间无法使用,比如全网暴力扫描。在万兆链路条件下,ZMAPv6 采用最高的配置模式进行 IPv6 全网空间扫描,也至少需要上亿年的时间,这是无法实现的事情。

近年,已经有许多研究者提出了有效的 IPv6 地址探测手段(非暴力扫描),如图 1 所示。主动探测技术是指使用的技术手段会涉及发送数据分组的操作,主要包括基于 DNS (domain name system, 域名系统) 数据的方法和基于种子地址的方法,两者最重要的区别在于是否会进一步对收集的地址分析并生成新的地址用于主动探测。

早期的工作^[4]指出通过 DNS 收集 IPv6 地址是有效的一种手段。基于 DNS 数据的方法也是主要通过主动发送 DNS 查询数据分组来收集 IPv6 地址。Strowes^[6]提出了一种通过 IPv4 地址进行反查然后进行 AAAA 查询得到 IPv6 地址的方法。随

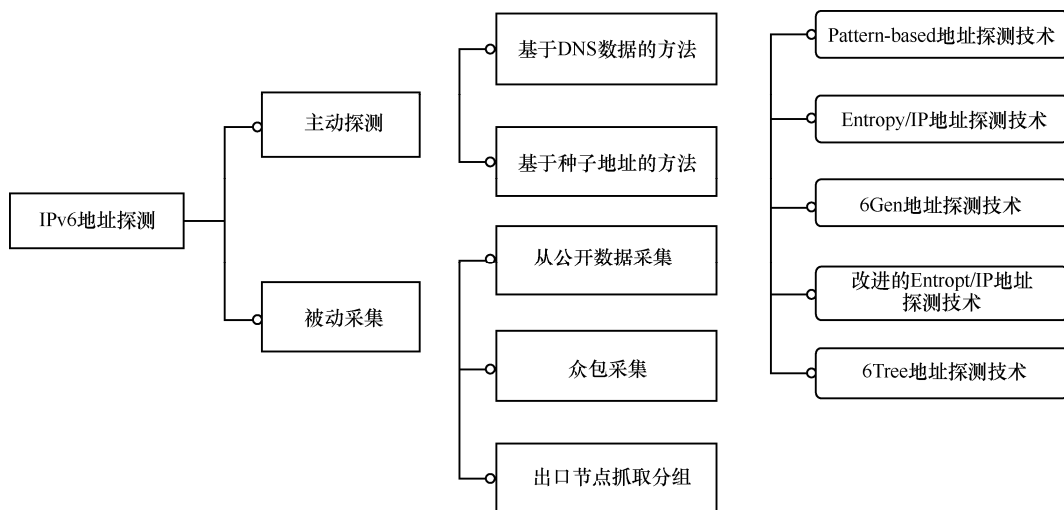


图1 近年 IPv6 地址探测技术分类



后 Fiebig 等^[7-8]通过进一步利用 DNS 反向查询 (PTR 记录) 收集了 2.8×10^6 个 IPv6 地址。通过 DNS 收集 IPv6 地址方式的优点在于其中很多地址都是提供服务的设备地址, 但是缺点是 IPv6 地址的发现数量相当有限。此外, Borgolte 等^[9]提出了结合 DNSSEC (domain name system security extensions, 域名系统安全扩展) 的区域文件收集 IPv6 地址的方法。

另一个探测思路是基于种子地址的探测方法, 这也是本文重点讨论的对象。种子地址指的是研究者收集的曾经或者一直存活的 IPv6 地址。这类探测技术通过挖掘这些已有的 IPv6 地址分配的内在规律, 选择性地生成新的 IPv6 地址作为主动扫描的输入。Ullrich 等^[10]在 2015 年提出了利用规则从种子地址中生成新地址的技术。Foremski 等^[11]在 2016 年提出的 Entropy/IP 使用香农熵研究种子地址结构, 并且使用贝叶斯模型去生成新的地址集合。随后, Murdock 等^[12]在 2017 年提出基于密度聚类的 6Gen 算法, 也是利用种子地址作为输入, 然后聚类后生成最可能存活的新地址。2018 年, Gasser 等^[13]的工作是阶段性总结工作, 首次指出种子地址的不平衡问题, 对 Entropy/IP 工作做出了复现和微小的改进, 同时在 6Gen 工作的基础上提出了新的别名检测算法, 并比较 Entropy/IP 和 6Gen 算法的地址生成好坏。Gasser 等^[13]的工作公开了全部数据集和代码, 希望未来的研究者能以这些数据和方法为起点继续探索 IPv6 探测领域。2019 年 Liu 等^[14]提出了基于分裂层次聚类的 6Tree 算法, 和 6Gen 相比, 其探测技术在对种子地址进行空间建模部分有着线性时间复杂度。

除了主动探测, 被动采集也是网络研究领域非常常见的手段。比如, 从网络上公开的 IPv6 地址源^[13,15]进行数据收集。在 IPv6 骨干网络的核心节点网卡监听并抓包收集 IPv6 地址也是非常有效的方法, 但是对于很多研究者来讲使用条件较为苛刻甚至不可得。此外, 参考文献[16-18]提出了

通过众包平台来采集 IPv6 地址的方法。

基于种子地址的探测方法往往需要依靠其他方法来提供 IPv6 种子地址作为算法的输入, 这部分的方法可以看作其他方法的延伸。同时, 无论是 DNS 数据解析的方法还是被动采集的方法都存在很大的局限性——不仅收集的地址数量有限, 而且很多情况下对于研究者来说并没有足够的条件来进行被动采集。相比之下, 结合了主动探测的、基于种子地址的探测方法能够在庞大的 IPv6 地址空间探测到大量存活 IPv6 地址, 故而成为近年的研究热点。

本文将重点讨论 5 种基于种子地址的探测技术, 总结这些技术的核心思想和各自的贡献, 分析这些技术的算法优缺点, 并且对这些方法从多个角度做出综合对比。然后, 本文使用从参考文献公开的数据集和自己采集的数据集共计 7 300 万种子地址集合作为输入, 通过试验对这些探测技术的命中率和时间性能进行比较。最后, 本文会给出对该领域的思考和未来值得继续研究的方向。

2 背景

这部分会介绍本文涉及的一些基础概念, IPv6 地址的编址原理以及 IPv6 的 IID (interface identifier, 接口标识) 的分配原则, 并通过对 CERNET (China Education and Research Network, 中国教育和科研计算机网) 持续采集的一个月 IPv6 地址分析, 来了解最近 IPv6 地址的 IID 分配情况。

2.1 基础概念

本文涉及的一些基础概念如下。

- IPv6 标准表示法: 用 32 个半字节表示, 比如: 2001:0db8:85a3:08d3:1319:8a2e:0370:7344。
- IPv6 缩略表示法: 对于连续为 0 的部分可以压缩, 使用 :: 表示, 比如 2001:0DB8::

1428:57ab。

- 命中率：评估地址生成算法的一种指标，是探测存活的地址占全部探测地址数量的比值。
- 别名前缀（aliased prefix）：参考文献[13]在研究中指出别名前缀的问题。在该 IPv6 地址前缀下，存在大量 IPv6 地址指向同一个主机接口。
- 别名地址：多个 IPv6 地址指向同一个主机接口。

2.2 IPv6 编址原理

通常一个 IPv6 地址是由路由前缀、子网 ID 和接口标识组成。和 IPv4 不同，在常用的 IPv6 环境，没有可变长度子网前缀的表示法。根据 IPv6 地址用途，编址类型可以划分为单播地址、任播地址和多播地址三类。其中全球单播地址是大规模 IPv6 探测工作主要的目标，因为该类地址是全球路由可达的。本文对 Gasser 等^[13]在互联网公开的 IPv6 数据集的有效地址部分进行了统计分析，见表 1。

从表 1 中可看到，全球单播地址占比高达 99.98%，并且从 Gasser 等^[13]公开的日探测集合中本文只发现少量的非全球单播地址。这也非常符合本文的预期，因为在大规模探测中，能够被探

测到的地址必然是全球可达的地址。

表 1 IPv6 编址类型占比

IPv6 编址类型	占比
全球单播地址	99.98%
站点本地地址	0.01%
多播地址	0.01%
其他地址	0

2.3 IID 接口标识分配方式

接口标识为 64 位，用于标识链路上的接口，在每条链路上接口 ID 必须唯一。在无状态配置中也是主要设置 IPv6 后 64 位取值。近年对 IPv6 的研究工作很少关注接口标识的分配方式，但是在本文看来，IPv6 全球单播地址之所以难以探测，重要原因是 IID 接口分配方式的随机性。IPv6 前缀 64 位绝大部分 bit 位的取值都可以由公开的 BGP 和分配的地址信息获取，但是对于后面 64 位的取值往往是未知的。IID 分配方式相对前缀较为复杂，因为有许多人为决定的因素，并没有完全统一的标准。本文整理了参考文献和 RFC 相关资料总结出的 IID 分配类型见表 2。

如果 IPv6 地址的分配方式都是 Low-Bytes 类型，那么只需要将对地址末尾的取值进行递增暴力扫描就能够探测出这种类型的全部 IPv6 地址。

表 2 IID 分配类型

IID 分配类型	描述	示例
Random ^[4]	随机设定取值	2402:f000:6:8601:ca70:72df:75b3:c5d8
Low-Byte ^[4]	连续零值和低位零值	2001:da8:a0:f015::1
IPv4 嵌入 ^[22]	IPv4 地址嵌入	2001:db8:122:344::192.0.2.33
Teredo 隧道 ^[22]	过渡采用的隧道技术	2001:0000:4136:e378:8000:63bf:3fff: added
EUI-64 嵌入 ^[24]	MAC 地址嵌入	2402:f000:6:8601:221:2FFF:FEB5:6E10
Modified EUI-64 ^[24]	嵌入组织唯一标识符	2001:db8::1234:56ff:fe00:ffff
Wordy ^[4]	嵌入英语词汇	2001::cafe:face
Ports ^[4,25]	嵌入服务端口	2001:db8::21



事实上,早期的针对 IPv6 网络的探测思考都是从这些分配方式是否易于扫描来考虑的^[2,19]。但是从本文对 CERNET 北京节点抓取的一个月数据分析中显示,随机类型的地址是绝大多数地址的分配方式,平均占比 96%,如图 2 所示。特别地,笔者注意到 EUI-64 (extended unique identifier, 扩展的唯一标识符) 的分配方式由于隐私问题使得其数量较小并且在不断下降。

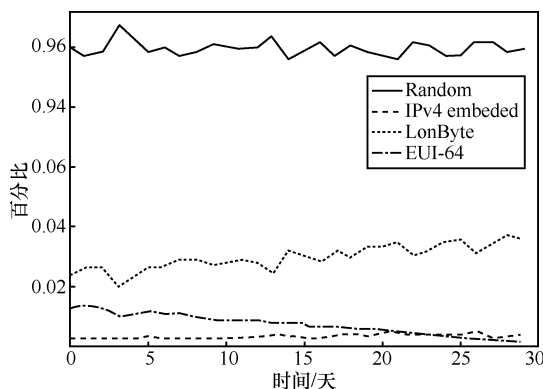


图2 一个月内 IID 分配类型占比变化趋势图

3 地址探测技术

本文重点是讨论基于种子的地址探测技术,这类技术的共性就是需要已有的 IPv6 种子地址作为输入,通过建模和分析进而生成新的可能存活的 IPv6 地址,从而缩小搜索空间,然后通过主动探测的方式去验证这些新的 IPv6 地址是否存活。这部分将从探测算法原理和探测技术优缺点进行阐述。

3.1 Pattern-based 地址探测技术

pattern 可以理解为 IPv6 地址集合中的一种规律,比如表 2 列举的 IPv6 接口标识分配方式都可以算作 pattern。早期的 IPv6 探测工作主要是通过分析 IPv6 中易于被猜测的规律来生成新的地址^[2,19]。期间许多研究学者提交了 RFC 设计草案,提出了很多有规律的 IPv6 地址生成方式,比如 wordy 这种嵌入的 face、café 英文单词的方案,这些有规律的分配方式也给扫描带来了便利。Gont^[4]在

2015 年开发的 IPv6 工具通过结合一些人为设定的规律来做 IPv6 探测扫描,这些规律包括 Low-Bytes、EUI-64 嵌入、IPv4 嵌入等。早期的 IPv6 地址探测背后思想都是挖掘 IPv6 地址 128 bit 本身的规律特性,基于这些规律猜测最可能被使用或者分配的 IPv6 地址。

但是早期的对 IPv6 地址规律的挖掘研究很多都是根据经验人为猜测,并且人为设定新的规律来发现新的 IPv6 地址。因此,Ullrich 等^[10]在 2015 年提出了 Pattern-based 地址探测技术。这项技术能够自动分析种子地址集合的 pattern,并且依据这些 pattern 生成新的 IPv6 地址。

Pattern-based 地址探测技术的核心是递归的地址生成算法。它从最初设置的 bit 开始搜索,此后每次递归新增一个比特加入 pattern,最终未列入 pattern 的 bit 位到达设定的阈值的时候程序停止。整个算法主要包括 4 个环节,具体如下。

- 提取 pattern: 最终提取的 pattern 是所有候选 pattern 中符合的种子地址数量最多的。在每次递归搜索过程中根据种子地址在该位的取值情况决定 pattern 中的取值(0 或 1)。
- 反向规则: 整个递归搜索类似二叉搜索树。
- 初始化: 初始化需要指定 X 个 bit 位, X 为正整数且取值至少为 1。算法会从 2 的 X 次方取值情况开始搜索,以保证涵盖全部可能取值的搜索空间。
- 终止条件: 如果未被加入 pattern 的 bit 位数达到设置的阈值,算法停止。

此外,虽然这项技术尽可能地使用递归搜索的方式自动挖掘种子地址的 pattern,但是笔者还是人为地添加了一些大家熟知的也是常用的 pattern,比如 Low-Byte、Ports 和 Modified EUI-64。这项研究作为启发式的 IPv6 地址探测技术开创了先河。

Pattern-based 地址探测技术主要具有 2 个优

点。首先该技术是能够自动地从种子地址挖掘地址每一个 bit 取值规律, 避免了人为分析的缺陷, 并且能够成功地将其方法运用到新的种子地址集合中。第二, 从作者的实验结果可以看出(Figure 1^[10]), 这项技术发现的地址数量高于暴力扫描, 证实了对于种子地址的启发式探测方法优于暴力扫描的结论。

作为发表时间偏早的基于种子地址的启发式探测算法^[10], 这项技术探测 IPv6 地址能力远远不及后面的探测技术。其次, 这项技术对于种子地址集合的要求比较严格。如果种子地址随机化的现象比较严重, pattern 的挖掘结果不够准确, 大大降低了命中率。图 2 中, 在笔者近期观测的 IPv6 地址集合中随机分配的方式较大比例, 这会影响 Pattern-based 地址探测技术效果。另外, 参考文献[10]并没有充分考虑各种已有的 IID 分配类型, 只是人为选定了一些常见的类型。其最大的贡献是引发研究者开始思考使用启发式的方法分析种子地址, 从已知的地址中寻找规律去探测未知而庞大的 IPv6 地址空间。

3.2 Entropy/IP 地址探测技术

Foremski^[11]在 2016 年提出这项地址探测技术, 一种自动化挖掘种子地址结构的方法。这项技术结合了信息论的香农熵、机器学习和概率图模型, 并且给出了可视化 IPv6 地址结构的多种方案。这些可视化技术能够帮助研究者很好地从整体了解 IPv6 地址集合的特点。

Entropy/IP 地址探测技术的地址生成算法可以分为 4 个步骤。

步骤 1 香农熵

香农熵被提出是用于解决对信息的量化度量问题。香农熵的计算原理如式(1)所示, 香农熵的取值范围是[0, 1], 熵值越大表示该变量不确定性越大, 其随机性就越强。

$$H(X) = -\sum_{i=1}^k P(x_i) \lg P(x_i) \quad (1)$$

基于 Entropy/IP 的地址探测技术会计算 IPv6

集合的每个半字节的熵值。

步骤 2 地址分段

一个 IPv6 地址总共有 32 个半字节(128 位), 该步骤将熵值相近的半字节合并成段。相近的阈值不能设置过大或者过小, 不然会严重影响分段的正确性, 以至于后面的贝叶斯建模效果低下。

步骤 3 分段信息挖掘

这个部分的工作主要是对步骤 2 划分的段做一些统计分析, 包括该段最多的取值情况, 取值最密集的范围、最小值、最大值等。这些统计信息会作为贝叶斯建模的输入。

步骤 4 贝叶斯网络建模

根据段与段之间的取值条件概率构建整个地址的贝叶斯图模型。这个图模型描述的是每个段的取值以及每个段取值与之前段的条件概率。

最终该技术可以从贝叶斯网络图中随机地遍历整个贝叶斯概率模型描述的取值空间, 选取用户设定地址数量作为待探测目标。该工作首次将贝叶斯和信息论的知识应用在 IPv6 探测领域。

Entropy/IP 技术的主要贡献是提出了以上 4 步骤的启发式地址生产模型并给出了可视化 IPv6 地址结构的有效方法。通过这些可视化地址集合的熵和概率图, 研究者可以很清楚地了解 IPv6 地址集合在每一个半字节上的取值情况。除此之外, 参考文献[11]通过可视化 IPv6 地址每个半字节的熵值情况, 分析比较了服务地址、客户端地址等不同地址类型的地址取值差异。

整个算法的执行时间随输入的种子地址增大而增大, 但是耗时仍在可接受范围内, 例如在处理 100 万种子地址的时候耗时只需要 112 s (如图 3 所示)。该技术在持续探测的场景也有良好的应用, 因为 Entropy/IP 一旦完成贝叶斯建模, 后续的地址生成部分无需再重新建模。

虽然这项技术很好地解决了 IPv6 地址结构可视化问题, 并且也具有从全局探索了种子地址的特点, 但是算法生成的地址命中率较低。并且随

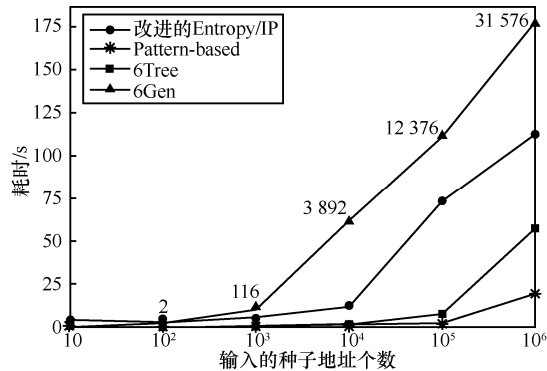


图3 4种探测算法在不同数量种子地址上的耗时

着现在的地址 IID 分配方式随机化, 此算法对于 IPv6 地址后 64 位部分的半字节计算很容易出现熵值接近 1 的情况, 这样后面 64 位都可能作为一个整个大段或者几个段被贝叶斯建模处理, 严重影响算法效果。

本文认为 Entropy/IP 算法本身至少在两个方面可以有所改进。一是算法处理地址的细粒度可以有所调整, 不一定只在半字节 (4 bit) 分析地址的熵值特点。二是贝叶斯网络生成地址时候不一定要采用随机方式来遍历贝叶斯网路的取值空间, 可以适当的加入启发式条件。

3.3 6Gen 地址探测技术

Murdock 等^[12]提出的方法 6Gen 是从种子地址集合中识别地址取值空间中密集的区域并在这些密集区域选择新的地址取值, 进而生成需要扫描的候选地址。该方法主要基于一个很直观的假设: 已有的种子地址在不同半字节取值最密集的区域也是未知地址最可能的取值区域。6Gen 地址探测技术的地址生成算法核心步骤如图 4 所示。

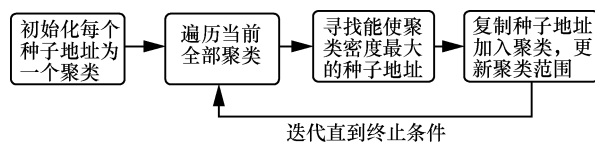


图4 6Gen 算法核心步骤

6Gen 对于输入的 IPv6 种子地址集合采用了 DBSCAN 聚类算法, 将每一个种子地址初始作为一个聚类, 并在迭代过程中使多个聚类代表

的范围不断增大, 直到达到阈值停止。6Gen 的地址生成算法有 2 个重要的概念——聚类的范围和密度。

- 聚类的范围: 如图 5 所示, 聚类的范围由该聚类集合的全部种子地址在不同的半字节位置确定。范围的定义有二种模式: Tight 和 Loose, 区别在于处理变化的半字节是采用最大值和最小值作为范围还是 [0-f] 全部可能取值作为范围。

种子地址	Tight	Loose
2001::2 2001::b 2001::e	2001::[2-e] ↓ 范围 13的1次方	2001::[0-f] ↓ 范围 16的1次方

图5 Tight 和 Loose 范围对比图

- 密度: 该聚类的密度等于该聚类集合中种子地址数量与该聚类范围的比值

在参考文献[12]的实验中, 聚类的范围选择 Tight 还是 Loose 模式对于最终发现的地址差异并不明显。

6Gen 地址探测技术 (以下简称 6Gen) 是将聚类算法应用在 IPv6 探测领域中比较成功的技术。该技术通过刻画 IPv6 种子地址集合的相互关系, 来研究 IPv6 地址空间中潜在的地址取值。参考文献[12]提出的核心假设也非常合理, 因为在现实世界, 同一个组织或者同一个网络下, 网络管理者很可能会给这片区域的网络设备使用相同的分配策略进行分配, 比如最典型的是连续的 IP 地址。因此, 6Gen 地址探测技术首先去聚类已有的 IPv6 地址来寻找这些地址所在的“区域”, 然后在这些区域中去寻找未被发现的 IPv6 地址。Murdock 等^[12]的工作指出, 和 2016 年的 Entropy/IP^[11]地址探测技术相比, 6Gen 能够发现更多的 IPv6 地址, 并且 6Gen 在较小的地址集合上生成的 IPv6 地址探测的命中率高于 Entropy/IP

地址探测技术。

6Gen 另一个贡献就是首次指出了 IPv6 生成地址的别名地址和别名前缀问题。在种子集合中大量存在别名地址会严重影响地址生成算法的普适性,会使得算法生成的地址具有“偏见”,即很可能探测到的大部分存活地址都是指向同一个接口。虽然最精确检测地址别名的方法应该是指纹识别,但是扫描代价往往是不可接受的。为此,6Gen 提出了一种简洁的随机探测发现别名前缀的方法,通过这种算法识别 206 个别名前缀。其设计思想很简洁直观——既然 IPv6 别名地址大量指向同一个接口,那么在别名前缀下随机生成一定数量的地址,如果都被检测为存活的,那么这些地址很可能就是别名地址,这个前缀就是别名前缀。因为在 6Gen 看来,一个 64 位前缀下的 IPv6 地址空间如此庞大,IPv6 地址又是如此稀疏,随机生成的这些地址不可能都存活。

但是在时间性能方面,6Gen 耗时随着种子地址集合的增大迅速增大,呈指数增长。这是因为 6Gen 的地址生成算法本身使用了多个 for 循环,不仅迭代查找种子还会迭代查找聚类。复杂的算法性能使得 6Gen 的使用者不得不考虑做一些代码上的优化或者并行运行之类的操作。

控制 6Gen 地址生成算法耗时的一种方法是控制终止条件,但是这会牺牲聚类的效果和生成的地址数量。另一个缺陷就是,6Gen 算法本身是不断增大各个聚类的集合,各个聚类集合存在大量的交集,使得整个算法对内存的消耗也比较大。

3.4 改进的 Entropy/IP 地址探测技术

Gasser 等^[13]在 2018 年的工作可以算是 IPv6 探测工作中一个阶段性的成果,其公开了全部实验数据和代码,希望未来的研究者能以此为基础继续探索 IPv6 探测领域。

这项探测技术的主要贡献是解决了输入的种子地址不平衡问题和提出了新的别名检测算法,并复对比了改进之后的 Entropy/IP 和 6Gen 算法,

同时也从端口扫描和 ICMPv6 探测 2 个方面研究了 IPv6 地址探测中的存活性问题。此外,在可视化方面,又进一步给出描绘整个 IPv6 地址空间的方法。

首先,参考文献[13]最大的亮点就是全面地整理了网上公开的 IPv6 地址数据源,并且详细分析了这些地址的平衡性问题——地址集合在不同 AS 的分布差异过大,得出了之前的 IPv6 地址探测技术由于对种子地址没有合理的处理,或多或少地存在生成“偏见”的 IPv6 地址问题。在参考文献的工作中,发现大量 IPv6 地址都集中于主要的几个 AS 域,并且不同来源的种子地址集合数量和所覆盖的 AS 广度的差异比较大。大量同一个 AS 的地址会使地址探测技术生成的地址多数集中于这些 AS,甚至是大量的别名地址。

其次,参考文献[13]提出了新的别名检测算法。这种方法是在 6Gen 工作的基础上改进而来,通过迭代步进的方式不断探测多组 IPv6 地址,以确定是否是别名问题。这种启发式的别名检测方法也是基于 6Gen 提出的别名检测类似的假设。所不同的是,参考文献[13]的工作利用指纹识别技术对这些别名地址做出了验证,确认了大部分通过启发式算法生成的地址都是别名地址。

这项技术在对 Entropy/IP 地址生成算法做出了较小的改动,主要是将原本随机遍历贝叶斯网络的地址生成方式改为全局遍历。同时比较了 Entropy/IP 和 6Gen 的地址生成能力和命中率,通过实验发现两者在生成亿级的新地址时,重复的地址也只有百万级别。参考文献[13]得出结论在实际运用过程中需要同时采用这二个地址探测技术。

该探测技术还首次发现 IPv6 存活性地址的特点。通过在 53、443 端口上 TCP/UDP 探测和 ICMPv6 探测来研究地址的响应情况,发现有少部分 ICMPv6 未响应的地址会响应 53 或者 80 端口。



因此要想全面探测地址的存活性, 需要结合端口扫描的方法。但是其工作也指出这部分情况的地址数量比例较小, ICMPv6 仍然可以作为判断 IPv6 是否存活的主要依据。

3.5 6Tree 地址探测技术

Liu 等^[14]提出了一种分裂聚类的 IPv6 地址探测技术 6Tree, 不仅解决了 6Gen 算法的时间效率问题, 还给出了一种基于反馈的别名去除方法。

6Tree 地址探测技术模型如图 6 所示。

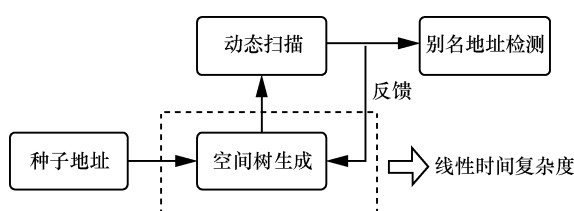


图 6 6Tree 地址探测技术模型

6Tree 算法的核心思想是分析已知的活动地址作为种子来学习它们的分布特征, 将 IPv6 地址理解为高维向量, 并在相应的种子向量上执行分裂层次聚, 以生成称为空间树的数据结构, 该数据结构表征不同维度中的半字节值变化。Liu 等^[14]认为活动地址向量分布不均匀, 并且某些维度比其他维度具备更多的可变值, 即它们的信息熵是不同的。扫描器可以优先搜索这些可变维度, 因为未发现的活动地址向量可能在这些维度中具有不同的值, 但在其他稳定的维度中具有相同的值。6Tree 的地址生成阶段和扫描是同时进行的, 该技术会根据扫描的结果动态调整生成地址的空间, 并且在这过程中进行别名地址检测。

和之前的工作^[11-12]相比, 6Tree 不仅能处理较大的种子地址输入还能够探测到更多的存活地址, 实验结果^[14]显示 6Tree 能够基于 274 万种子地址, 生成 3 亿待探测地址, 并且发现 636 万存活地址, 其中去除别名的地址占有 167 万。同时参考文献^[14]提供了别名问题一种新的结论——99.5% 的别名地址聚集在别名前

缀下面。

该探测技术对于 IPv6 的研究给出了数学化的定义, 式 (2) 将 IPv6 的 λ bit 地址按照一定 bit 划分映射为多维的向量空间。

$$v_{(2^m)}[\delta] = \frac{N \bmod 2^{m \left(\frac{a+1-b}{m} \right)} - N \bmod 2^{m \left(\frac{a-b}{m} \right)}}{2^{m \left(\frac{a-b}{m} \right)}} \quad (2)$$

其中, N 表示 IPv6 地址中 a 比特二进制转换为十进制的整数, m 表示将这 a 比特按照 m 比特划分, b 表示向量第几个维度。这种映射思路在前面的研究工作中也都有体现, 但是 6Tree 算法^[14]第一次将其用数学的方式表达出来。

6Tree 的地址探测技术优势就是与 6Gen 相比, 6Tree 具有更快的线性时间复杂度。并且和 6Gen 不同, 6Tree 是将整个种子集合不断划分形成空间树结构, 不会出现叶子节点有重复的地址情况, 内存消耗优于 6Gen。在别名检测方面, 6Tree 通过扫描反馈解决别名问题是比较接近问题本质的一种解决方案。

该技术的缺点就是由于扫描和地址生成是反馈进行的, 总体的耗时会有所提高。而且这种交互的行为让 6Tree 很难应用于一些高效的扫描技术^[2]。但是 6Tree 的工作给研究者一个新的思考角度, 是否可以根据反馈信息动态调整算法本身, 甚至可以结合近年热门的强化学习。

3.6 5 种地址探测技术对比

上文对各种算法的原理和优缺点做了分析, 表 3 比较了这 5 种基于种子地址的探测技术。对现有探测技术的综合比对能够帮助人们更好地分析各个探测技术的特点, 并且为之后的研究方向提供参考。

表 3 中处理较大地址耗时结果可以参考图 3 在 100 万种子地址作为输入时的结果。各个算法的可视化能力是按照论文中对其技术的可视化的贡献进行的比较。扫描是否有反馈是指扫描的结果是否会影响探测技术的模型本身。

表3 5种基于种子地址的探测技术对比

对比项目	Pattern-based ^[10]	Entropy/IP ^[11]	6Gen ^[12]	改进的 Entropy/IP ^[13]	6Tree ^[14]
年份	2015 年	2016 年	2017 年	2018 年	2019 年
处理较大地址耗时	最快	中等	最慢	中等	快
关键技术	递归	贝叶斯网络	聚类	贝叶斯网络	分裂树
是否考虑别名	否	否	是	是	是
可视化能力	较弱	较强	较弱	较强	较强
扫描是否有反馈	无	无	无	无	有
是否公开数据集	否	否	否	是	否
是否开源代码	否	否	否	是	否
论文中发现地址量级	10 ³	10 ⁶	10 ⁶	10 ⁸	10 ⁸

虽然 6Tree 工作使用了 2018 年改进的 Entropy/IP 工作^[13]公开的 IPV6 地址数据集,但是由于他们并没有公开从中提取和处理过的数据,因此在表格中标注 6Tree 未公开数据集。特别的,2018 年的改进 Entropy/IP 工作和 2016 年的 Entropy/IP 有共同作者,虽然他在 2018 年公开了 Entropy/IP 的代码,但是由于 2016 年的工作^[11]本身没有公开代码,因此在表格中仍然标注为否。

4 评估

4.1 实验说明

虽然前面讨论了 5 种基于种子地址的地址探测技术,但是在这部分的实验只比较 Pattern-based、6Gen、改进的 Entropy/IP 和 6Tree 4 种地址生成算法的效果。因为 2018 年改进的 Entropy/IP 探测技术主要贡献在于解决种子地址不平衡问题和提出新的别名检测算法,对 2014 年的 Entropy/IP 的地址生成算法模型改动微小(见第 3.4 节)。此外,在本文的实验中也没有考虑对生成的地址作别名检测,故只需要比较这 4 种探测技术的地址生成算法。

这部分实验没有考虑别名检测原因有 2 个。首先,Pattern-based 和 Entropy/IP 都没有考虑别

名问题的情况,而 6Gen、改进的 Entropy/IP 和 6Tree 都采用了启发式的别名检测算法,均存在一定程度的误报和漏报。其次,最精确的别名检测需要使用指纹识别的技术,但是对上百万的地址作详细的指纹识别需要很长的时间周期。在未来的工作中,笔者会深入研究别名检测问题。

4.2 实验环境

在近年的 IPv6 地址探测工作中,命中率是一个非常重要的考查指标。此外,耗时也能反映算法的时间复杂度,同时也是衡量该算法是否易于在生产环境部署的重要考虑指标。本文将比较这些探测技术时间性能和命中率的好坏。

实验环境如下。

- 种子地址: 2018 年开源的 IPv6 数据集和 CERNET 北京节点采集的数据共 7 300 万^[13]。
- 扫描主机: 2 台 64 G 内存单核 Intel(R) Xeon(R) CPU E5-2640v4 2.40 GHz 的 Ubuntu16.04 服务器。
- 复现代码: Python 实现 6Gen、6Tree 和 Pattern-based 算法;改进的 Entropy/IP 使用^[13]公开的代码。
- 探测存活技术: ICMPv6, 使用工具 ZMAPv6^[21]。



4.3 实验结果

和参考文献[13]的工作类似,为了解决种子地址在不同 AS 的分布不平衡问题,本文尽可能从每个 AS 下的种子地址随机抽取相同的种子数量作为本文的种子地址集合。通过这种方式本文构造了数量分别为 10、100、1 000、10 000、100 000、1 000 000 的种子地址作为 4 种地址生成算法的输入。

其中,种子地址输入为 10、100、1 000 的时候,由于种子地址集合 AS 种类数(在本文的 7.3×10^7 种子地址集合中涵盖 9 378 个 AS)大于采样个数,随机选择一定数量的 AS 作为被采样对象。

为了消除采样不完全带来的误差,本文重复实验 10 次计算各个算法的耗时平均值得到图 3 所示的结果。注意,6Gen 处理百万的种子地址的耗时已经需要接近 9 h (31 576 s),远远高于其他算法的耗时。因此为了可视化的需求,笔者将 6Gen 曲线做了一定的调整绘制在图 3 中(将 6Gen 的各个测量点耗时秒数计算其平方根绘制在曲线中),但是各个测量点的真实耗时秒数都标注在折线上。

在种子地址集合较小的时候 4 种探测算法的耗时相差不大,但是随着种子地址数量增加,耗时差异逐渐扩大。其中,Pattern-based 和 6Tree 的耗时较小,改进的 Entropy/IP 耗时较大,但是都在可接受范围。相比之下,6Gen 耗时增加迅速,如果没有对算法本身做一些修改,很难在未来有更长远的应用。

考虑到 6Gen 时间性能问题,本文只选取了 100 000 的种子地址作为输入,通过这 4 种算法分别生成了 100 000 个新地址,使用 ZMAPv6 探测存活情况得到图 7 所示的存活率。

从结果横向比较分析,这些探测技术的整体命中率都比较低,最多的 6Tree 也只发现了 813 个存活的 IPv6 地址。相比之下,Pattern-based 命中率特别低,只发现不到 9 个存活的 IPv6 地址。从结果

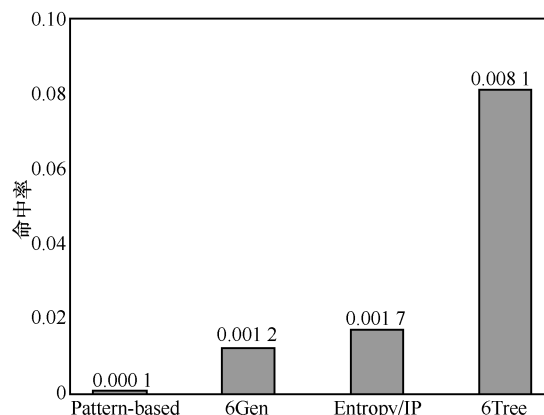


图 7 4 种探测算法生成 100 000 个地址的命中率

的横向对比,可以发现反馈的算法命中率会比其他的方法好一些,这也更加启示应该将带有反馈的模型结合到 IPv6 探测的工作中来。

下面将实验结果和论文结果做一个简单的纵向对比,比较都只是用了参考文献[13]公开的数据集的工作——6Tree 和 2018 年的改进 Entropy/IP。和他们的工作相比,这部分实验的命中率均有所下降,比如 6Tree^[14]扫描了 3 亿地址发现了大约 4×10^6 个的存活地址(命中率为 0.015);2018 年参考文献[13]的 Entropy/IP 扫描 1.1 亿地址发现约 2.7×10^5 多个地址(命中率 0.002);2018 年参考文献[13]复现的 6Gen 扫描 1.2 亿地址发现约 4.8×10^5 多个地址(命中率 0.003)。下降的原因很可能有 2 个:一是公开的 IPv6 地址数据本身有许多地址不再存活,二是本文采集的 CERNET 北京节点数据有相当多的临时分配的客户端地址。

尽管我们的实验和论文的工作都采用了相似的数据集,但是纵向比较仍然只能作为参考。首先,随着时间的推移,收集的 IPv6 地址集合本身就会有大量地址不再存活,并且在不同的时间和地点探测的存活率也是不一样的。其次,参考文献[13]公开的 IPv6 数据目录内容较多,研究者是如何处理和筛选数据的方法也不尽相同,没办法完全比较各自输入的种子地址集合差异。

综上,虽然很难有完全精确地纵向比较这

些基于种子的地址探测技术的实验,但是通过在相同的实验条件下的横向对比,可以发现不同探测技术的优缺点,尤其是在时间性能和命中率方面的好坏,同时横向比较的结果对研究者如何提出基于种子地址的新探测技术会有所启示。

5 结束语

如今,在全球大规模部署 IPv6 下一代互联网协议已经是不可阻挡的趋势。网络扫描在网络研究领域有着不可替代的位置,探测技术是发现安全问题、排查网络故障和了解网络现状的核心手段。本文主要对 IPv6 地址探测技术做了分类,重点讨论了基于种子地址的 IPv6 地址探测技术。将这些最新的技术梳理、总结和分析,比较各自算法的优缺点。在现有研究成果的基础上,进一步总结了基于种子地址的 IPv6 研究领域所面临的挑战。

(1) 别名地址问题

一方面从探测 IPv6 地址数量来讲,别名问题并不是大问题,能够被 ICMPv6 甚至 TCP80 扫描响应的地址都可以算是存活地址。但是,另一方面,别名问题和别名前缀会影响基于种子地址的探测算法的普适性,会使得算法探测的存活地址很可能都是网络上同一个接口的地址或者是同一个中间设备响应的地址。而基于种子地址的探测算法对种子地址敏感度较高,如果不能处理好别名问题,将大大地降低发现的真实 IPv6 设备数量。

(2) 处理大量种子地址输入的能力

现有的地址探测算法时间复杂度普遍太高,并且随着种子地址数量增加耗时也会不断增加,比如 6Gen 的时耗提升得非常快。即使对于建模时间复杂度已经很低的 6Tree,由于其地址生成过程是与扫描动态结合的,会给扫描和地址生成过程都带来巨大的性能瓶颈。而 IPv6 地址空间未来是

必然会不断增大的,合理地优化算法的复杂度变得非常重要。

(3) 应对随机化的能力

除了 IPv6 的后 64 位分配策略本身有很多随机的情况,随着 IPv6 种子地址集合的不断增大,其整体的表现特点必然呈现随机性(设想全部地址空间作为种子地址集合,则 IPv6 的每个 bit 都是 0 和 1 等概率,香农熵全部为 1)。这会使得这些挖掘种子地址特征的方法不断失效,甚至完全失效。因此合理的过滤和划分种子地址、正确的处理地址较为随机的部分是未来需要解决的问题。

(4) 种子地址有效性

作为给互联网设备分配地址的 IPv6 协议,这些 IPv6 地址很可能在一定时间过后就不再存活,比如临时分配的地址。在实际网络测量中也经常发现种子地址集合中大量地址无法被探测,可能是不存活也可能是被网络访问策略阻塞。但是无论是哪种情况,这些不存活的地址如果被用于基于种子的地址探测技术很可能会降低生成地址的命中率。

尽管 IPv6 地址探测面临着非常多的挑战,笔者仍然认为未来有如下可以继续研究的方向。

(1) 有反馈的启发式方法

多数工作的算法都是基于种子地址建模,并使用模型生成地址,再将生成地址作为扫描器的输入。这三部分往往都是独立的,但是 6Tree 给人们一个非常好的启示就是扫描和地址生成是可以相互反馈的。由此可以进一步思考是否能将反馈作用于建模过程,是否能够提出更加优化的有反馈的启发式方法。

(2) 基于 IPv6 情报

本文探讨的 IPv6 地址探测技术都是利用 IPv6 地址本身作为种子地址。这样做的缺点是输入的数据除了 IPv6 地址本身的信息外,就没有更多的信息能够被模型处理。相反,如果采



用带有额外信息的 IPv6 情报作为输入, 利用这些额外的信息来优化地址生成算法也许能提高探测命中率。

(3) 机器学习方法的应用

贝叶斯模型、聚类算法都是在 IPv6 地址探测中应用比较成功的方法, 除此之外一些近年主流的机器学习模型能否进一步用于 IPv6 探测领域仍然是一个值得探索的问题。

参考文献:

- [1] DURUMERIC Z, BAILEY M, HALDERMAN J. An internet-wide view of internet-wide scanning[C]//USENIX Conference on Security, Aug 20-22, 2014, San Diego, USA. [S.l.:s.n.], 2014: 65-78.
- [2] DURUMERIC Z, WUSTROW E, HALDERMAN J A. ZMap: fast internet-wide scanning and its security applications [C]//USENIX Conference on Security, Aug 14-16, 2013, Washington D.C., USA. [S.l.:s.n.], 2013: 605-620.
- [3] SAILAN M, HASSAN R, PATEL A. A comparative review of IPv4 and IPv6 for research test bed [C]//International Conference on Electrical Engineering and Informatics, Aug 5-7, 2009, Selangor, Malaysia. Piscataway: IEEE Press, 2009: 427-433.
- [4] IETF. Network reconnaissance in IPv6 networks: RFC7707[S]. 2016.
- [5] OTTOW C, VLIET F, BOER P. The impact of IPv6 on penetration testing[C]//Information and Communication Technologies, EUNICE 2012. Heidelberg: Springer, 2012: 88-89.
- [6] STROWES S D. Bootstrapping active IPv6 measurement with IPv4 and public DNS[J]. Computing Research Repository, 2017.
- [7] FIEBIG T, BORGOLTE K, HAO S, et al. Something from nothing (there): collecting global IPv6 datasets from DNS[C]//International Conference on Passive and Active Network Measurement. Heidelberg: Springer, 2017.
- [8] FIEBIG T, BORGOLTE K, HAO S, et al. In rDNS we trust: revisiting a common data-source's reliability [C]//Passive and Active Measurement Conference, March 26, 2018, Berlin, Germany. Heidelberg: Springer, 2018.
- [9] BORGOLTE K, HAO S, FIEBIG T, et al. Enumerating active IPv6 hosts for large-scale security scans via DNSSEC-signed reverse zones[C]//2018 IEEE Symposium on Security and Privacy (SP), May 20-24, 2018, Francisco, CA, USA. Washington: IEEE Computer Society, 2018: 770-784.
- [10] ULLRICH J, KIESEBERG P, KROMBHOLZ K, et al. On reconnaissance with IPv6: a pattern-based scanning approach[C]//2015 10th International Conference on Availability, Reliability and Security (ARES), 2015.
- [11] FOREMSKI P, PLONKA D, BERGER A. Entropy/IP: uncovering structure in IPv6 addresses[C]//ACM Internet Measurement Conference, Nov 14-16, 2016, Santa Monica, USA. New York: ACM Press, 2016.
- [12] MURDOCK A, LI F, BRAMSEN P, et al. Target generation for internet-wide IPv6 scanning[C]//Internet Measurement Conference, Nov 1-3, 2017, London, UK. New York: ACM Press, 2017.
- [13] GASSER O, SCHEITL Q, FOREMSKI P, et al. Clusters in the expanse: understanding and unbiasing IPv6 hitlists[C]//Internet Measurement Conference (IMC), Oct 31-Nov 2, 2018, Boston, USA. [S.l.:s.n.], 2018: 364-378.
- [14] LIU Z Z, XIONG Y Q, XIN L, et al. 6Tree: efficient dynamic discovery of active addresses in the IPv6 address space[J]. Computer Networks, 2019.
- [15] IETF. Basic requirements for IPv6 customer edge routers: RFC 7084 [S]. 2013.
- [16] HUZ G, BAUER S, CLAFFY KC, et al. Experience in using MTurk for network measurement[C]//SIGCOMM C2B(I)D Workshop, Aug 17, 2015, London, UK. New York: ACM Press, 2015.
- [17] QASIM L, MATTHEW L, MACIEJ K, et al. Using crowd sourcing market places for network measurements: the case of spoofer[C]//Network Traffic Measurement and Analysis Conference, June 26-29, 2018, Vienna, Austria. Piscataway: IEEE Press, 2018.
- [18] VARVELLO M, BLACKBURN J, NAYLOR D, et al. EYEORG: a platform for crowdsourcing Web quality of experience measurements[C]//ACM Conference on Emerging Networking Experiments and Technologies, Nov 30-Dec 3, 2016, Philadelphia, USA. New York: ACM Press, 2016.
- [19] CHOWN T. IPv6 implications for network scanning internet requests for comments: RFC5157[S]. 2008.
- [20] BARRERA D, OORSCHOT P. Security visualization tools and IPv6 addresses[C]//International Workshop on Visualization for Cyber Security, Oct 11, 2009, Atlantic City, USA. Piscataway: IEEE Press, 2009: 21-26.
- [21] GASSER O. Evaluating network security using internet-wide measurements[D]. München: Technische Universität München, 2019.
- [22] IETF. IPv6 Addressing of IPv4/IPv6 translators: RFC6052[S]. 2010.
- [23] IETF. Teredo: tunneling IPv6 over UDP through network ad-

dress translations (NAT): RFC4380[S]. 2006.

[24] IETF. IP version 6 addressing architecture: RFC4291[S]. 2006.

[25] IETF. A recommendation for IPv6 address text representation: RFC5952[S]. 2010.

[26] 刘觅, 杨海川, 李振华. 运营商构建电信大数据生态之道[J]. 电信科学, 2018, 34(1): 109-119.

LIU M, YANG H C, LI Z H. Way for operators to build the telecommunication big data ecology[J]. Telecommunications Science, 2018, 34(1): 109-119.

[作者简介]



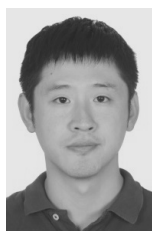
李果 (1996-), 男, 清华大学硕士生, 主要研究方向为网络空间探测、网络空间安全和数据挖掘。



何林 (1991-), 男, 博士, 清华大学博士后, 主要研究方向为互联网可审计性与隐私性、网络体系结构与协议设计。



宋光磊 (1992-), 男, 清华大学博士生, 主要研究方向为 IPv6 网络资源探测。



王之梁 (1978-), 男, 博士, 清华大学副教授、博士生导师, 主要研究方向为网络测试测量、互联网体系结构与协议、软件定义网络等。



杨家海 (1966-), 男, 清华大学教授、博士生导师, 主要研究方向为网络管理、网络测量、云计算以及网络安全。

李子木 (1978-), 男, 博士, 清华大学高级工程师, 主要研究方向为互联网体系结构与协议、网络安全与管理等。