文章编号：2095-2163(2020)09-0036-07 中图分类号：TP393.4 文献标志码：A

IPv6 网络拓扑测量目标选择技术

产毛宁，张 宇
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：由于 IPv6 地址空间巨大，使得 IPv6 网络拓扑测量成为一个巨大的挑战。本文基于 IPv6 存活地址列表，提出了一种 IPv6 网络拓扑测量目标选取技术，来提高 IPv6 网络拓扑测量的有效性和完整性。首先，收集并融合了不同来源的 IPv6 存活地址列表；其次，分析了 IPv6 存活地址列表的特征，接着提出了 IPv6 存活地址前缀列表预测算法 PrefixPrediction，对比 Entropy/IP 评估了算法的正确率；最后，给出了 IPv6 网络拓扑测量目标选取的综合方案。

关键词：IPv6 网络拓扑测量；IPv6 存活地址列表；IPv6 存活地址前缀列表预测

Target selection technique for IPv6 network topology measurement

CHAN Maoning, ZHANG Yu
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Due to the large IPv6 address space, IPv6 network topology measurement becomes a great challenge. Based on IPv6 Hitlists, this paper proposes an target selection technology for IPv6 network topology measurement to improve the effectiveness and completeness of IPv6 network topology measurement. First, IPv6 Hitlists is collected and merged from different sources, then the characteristics of IPv6 Hitlists is analyze and PrefixPrediction is proposed: the IPv6 prefix prediction algorithm based on Hitlists. Compared with Entropy/IP , the accuracy of the algorithm is evaluated. Finally a comprehensive scheme for target selection of IPv6 network topology measurement is presented.

【Key words】 IPv6 network topology measurement; IPv6 Hitlists; IPv6 prefix prediction based on Hitlists

0 引 言

互联网向基于 IPv6 的下一代互联网演进已成为全球普遍共识，目前 IPv6 的部署和使用正在飞速增长，但是由于 IPv6 地址空间巨大，地址规划复杂，地址空间划分政策多样，以及地址实际使用率低等这些不同于 IPv4 的问题，使得 IPv6 网络拓扑测量成为一个巨大的挑战。作为 IPv6 网络拓扑测量的输入，IPv6 拓扑测量目标选择技术研究如何通过有效地选择目标来提高拓扑测量的有效性和完整性。

为了克服上述问题，大量的研究工作从 IPv6 存活地址列表收集、IPv6 地址分析、IPv6 地址生成和 IPv6 网络拓扑测量的角度展开。IPv6 存活地址列表(Hitlists)的收集技术包括主动技术如随机探测::1[1]，完整探测每个声明的/32 前缀中的每个/48 中的::1 地址[2]，rDNS zone 游走，被动技术如基于 BGP update 或从流量中获取[3,4]。Gasser 等观察到存活地址列表本身包含聚类，从均衡性和无偏性的角度提出了 TUM Hitlist[5]；Ullrich 等使用递归的算法来发现和提取 IPv6 地址模式[6]；Foremski 等提出了 Entropy/IP 系统，使用了贝叶斯模型和聚类分析的机器学习技术来发现 IPv6 地址结构[7]；Murdock 等设计了 6Gen 目标生成器，利用地址局域性，迭代识别高密集区域[8]。上述工作的主要目标为 IPv6 网络主机发现，而非 IPv6 网络拓扑发现。CAIDA Ark 利用分布式测量点对每个全球声明的 BGP 可路由的/48 或更短前缀中 1 个::1 地址和 1 个随机地址做 ICMP-Paris traceroute，但实践中无法细粒度地抽样，同时存在大量冗余探测。与本文类似，Beverly 等将存活地址列表用于 IPv6 接口级网络拓扑发现[9]，但缺少预测存活地址前缀列表的独立研究。

本文从 IPv6 存活地址列表的角度，首先探究多源的 IPv6 存活地址列表收集技术，用尽可能多的来源保证 IPv6 存活地址的完整性；然后从多角度分析收集到的 IPv6 存活地址列表的特征，观察并总结该列表所体现出特性以及针对 IPv6 网络拓扑测量的使用上的指导，为了补充列表采集中的可能缺失和猜测未来可能增长的地址空间，提出 IPv6 存活地址

基金项目：国家重点研发计划(2018YFB1800702；2016YFB0801303；2016QY01W0103)。

作者简介：产毛宁（1995-），男，硕士研究生，主要研究方向：网络拓扑测量；张 宇（1979-），男，博士，副教授，主要研究方向：网络测量、网络安全、未来网络。

收稿日期：2020-06-08

前缀列表的预测算法并评估；最后给出 IPv6 网络拓扑测量目标生成的技术方案，研究的总体流程如图 1 所示。本文的主要贡献如下：

（1）利用 IPv6 存活地址列表指导 IPv6 网络拓扑测量目标选取。

（2）提出存活地址前缀预测算法。

（3）提出 IPv6 网络拓扑测量目标选取的综合方案。

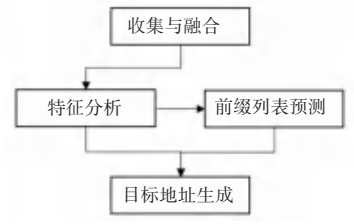

图 1　目标选取技术总体流程设计

Fig. 1　Overall process design of target selection technology

1　IPv6 网络拓扑测量目标选取技术

1.1　IPv6 存活地址列表收集与融合

类似于 IPv4，IPv6 存活地址列表（Hitlist）指一组能够大体上覆盖和代表 IPv6 网络的地址集合，并具有存活性，地址被分配、部署并在使用中；完整性，覆盖所有存活的 IPv6 地址空间；稳定性，地址列表随着时间变化尽可能小。本文尝试从 5 种不同的地址来源获取 IPv6 存活地址列表，包括：

（1）Rapid7 _ FDNS：FDNS 数据是 Rapid7 的 Project Sonar 公开数据一部分，本文采用其中的 fdns _any 数据，包括所有对 ANY 查询的回复，通过提取其中的 AAAA 记录，最终得到 IPv6 地址。

（2）CAIDA_Ark：IPv6 拓扑数据集作为 CAIDA Ark 平台测量数据的一部分，通过使用 Paris Traceroute 技术探测所有声明的/48 或者更短的 IPv6 前缀中的随机地址。本文从 CAIDA 2 月的 IPv6 拓扑数据里提取路径中出现所有 IPv6 地址。

（3）Bitnodes：Bitnodes 通过发现网络中的所有可达节点来评估比特币网络的大小，本文使用 Bitnodes 提供的 API，从节点名称中提取 IPv6 地址。

（4）TUM _ Responsive：德国 TUM 大学通过对 IPv6 存活地址列表的研究，提供 IPv6 Hitlist 服务，本文直接采用服务提供的回复的 IPv6 地址。

（5）TUM_Seeds：德国 TUM 大学的 Hitlist 服务作为一个收集项目，本身也采用了不同的收集源，根据其仓库中的实际数据，包括有 Alexa、CAIDA _ dnsname、CT（Certifcate Transparency）、Zonefiles、Openipmap 和 Traceroute，其中前 4 个根据提取的域名，查询 AAAA 记录，来获取 IPv6 地址；Openipmap 为从 RIPE ipmap 项目中提取的 IPv6 地址；Traceroute 为对所有其它源中的地址 Traceroute 再提取所有的路由器 IPv6 地址。本文将该仓库里不同来源地址综合作为一个收集源。

最后对不同来源存活地址列表清洗再合并，得到融合后的 IPv6 地址存活列表。由于收集到的地址中还包括特殊目的的全球单播地址，如过渡方案中的 6to4 地址等，需根据 IANA 提供的最新全球单播地址分配做筛选。

1.2　IPv6 存活地址特征分析

为了探究 IPv6 存活地址列表在 IPv6 网络拓扑测量上的最佳实践，有必要对收集到的 IPv6 存活地址列表深度分析，观察 IPv6 地址空间的部署和使用，揭示 IPv6 地址结构上的聚类，总结 IPv6 存活地址列表对 IPv6 网络拓扑测量目标选取上的指导作用。IPv6 地址一般可以分为网络前缀和接口标识两部分。

1.2.1　IPv6 网络前缀分析

IPv6 网络前缀一般指 IPv6 地址的前 64 比特。本文为了拓展讨论将前缀长度放宽至 128 比特，选择长度范围 [32,128] 来分析前缀，/32 作为下届，因为/32 前缀通常是区域性互联网注册机构（Regional Internet Registries，RIRs）分配给本地互联网注册机构（Local Internet Registries，LIRs）的最小块，128 是 IPv6 地址的最大长度。输入为不同来源的 IPv6 存活地址列表，并尝试从频率分布，相邻长度前缀关联性，密度 3 个角度分析 IPv6 前缀。首先从 IPv6 存活地址列表中提取不同长度前缀列表，再依次进行以下分析：

（1）不同长度前缀频率分布分析。统计不同长度前缀列表中的不同前缀频率。

（2）相邻长度前缀关联性分析。Plonka 等介绍了前缀聚集计数比率（Aggregate Count Ratio，ACR）[10]，对于给定的 N 个地址，将其聚合成不同长度的前缀，接着对于给定前缀长度 p，聚集计数 n_p 作为不同/p 包含的地址数量，并设定 p 为 0 时 n_p 为 1，则 $r_p = n_{p+x}/n_p$，其中 x 代表相邻前缀的长度差，单位为比特。本文依次取 x 为 4、8、12 和 16 计算 ACR。

（3）不同前缀长度密度分析。统计不同长度前缀列表中不同前缀包含的 IPv6 存活地址数 M，再计算各长度前缀列表中 M 的四分位数。

1.2.2　IPv6 接口标识分析

IPv6 接口标识一般指 IPv6 地址的后 64 比特，从生成类别上可以分为人工指定和算法自动生成两种，

不同的生成方案见表 1。张千里等总结了不同的接口标识生成方案,其中人工指定的方案包括[11]:

(1)基于低字节的接口标识(Low-byte)生成方案,一般指的是除最低两个字节外接口标识的大部分设为 0。

(2)基于 IPv4 地址(Embed-IPv4)的生成方案,一般指将 IPv4 地址作为接口标识的最后 4 个字节或者将 IPv4 地址的每个字节对应编码到最后的 4 个 16 比特中。

(3)基于服务、端口号(Embed-port)的生成方案,一般指将服务器的端口号或编号嵌入到接口标识中。

(4)基于单词(wordy-based)的生成方案,一般指使用易于记忆的单词代替 16 进制数字作为接口标识。

算法自动生成的方案包括:

(1)基于 MAC 地址的 IEEE EUI-64(IEEE-based)的生成方案,通常由 48 位的硬件地址,中间嵌入 0xfffe 得到,使得接口标识绑定机器硬件且全局唯一。

(2)保护用户隐私的生成方案,一般随机生成且随时间变化而变化。

(3)稳定、语义不透明的生成方案,目的是为了解决临时地址变化频繁导致的复杂的网络管理和部署问题,接口标识随机生成但对同一子网、同一网络接口产生地址相同。

(4)绑定地址与主机的生成方案,源于局域网内安全风险与多宿主机,生成 IID 随机,一般有加密生成接口标识方案和基于哈希的地址生成方式。本文使用 addr6 工具统计融合后 IPv6 存活地址列表中不同接口标识生成方案的占比,分析不同方案的实际部署和使用情况。

表 1 常见的接口标识生成方案

Tab. 1 Common interface identification generation scheme

方案	生成类别	实例
基于低字节的方案	人工	2001:db8::1
基于 IPv4 地址的方案	人工	2001:db8::192.0.2.1
基于服务、端口号的方案	人工	2001:db8::80:1
基于单词的方案	人工	2001:db8::bad::café
基于 IEEE EUI-64 的方案	自动	2001:1210:100:210:ee08:6bff:fe73:286a
保护用户隐私的方案	自动	/(随机地址)
稳定、语义不透明的方案	自动	/(随机地址)
绑定地址与主机的方案	自动	/(随机地址)

1.3 IPv6 存活地址前缀列表预测

考虑到 IPv6 存活地址列表由于收集来源的缺失如不包括非公开数据源,和限制如缺少长时间的积累,可能带来的不完整性,以及 IPv6 地址部署和使用的快速增长带来的滞后性,有必要设计一种方法能够根据已知收集的 IPv6 存活地址列表,预测未被收集的潜在存活和未来可能部署和使用的地址,作为 IPv6 网络拓扑测量目标的补充来提高拓扑发现的完整性。IPv6 存活地址前缀列表指一组包含 IPv6 存活地址的前缀集合。本文为了更完整地探测特定 IPv6 可路由前缀 R 下的目标网络拓扑,基于收集的 IPv6 存活地址列表,从网络拓扑测量中按前缀均匀抽样的常用方法出发,提出 IPv6 存活地址前缀列表预测算法 PrefixPrediction,即根据 IPv6 存活地址列表提取前缀集合,再结合多层级密度聚类和启发式后缀拓展,来预测未被收集的 IPv6 存活地址前缀列表,最终作为拓扑测量目标生成的一部分。

根据实践中地址分配体现出的层级性,以半字节即 4 比特为单位划分;和局部聚集性,分配的地址聚集在相邻或相近的前缀下,本文提出的算法核心思想为针对特定 IPv6 可路由前缀 R 下的目标网络,根据 R 包含的 IPv6 存活地址列表 HL 中提取的 IPv6 存活地址前缀列表,预测固定长度为 L 的 IPv6 存活地址前缀列表。R 的长度 $L_R \in [32,64]$,$L \in \{40,48,\cdots,64\}$。对于用 32 位 16 进制数 x_j 表示的 IPv6 地址 X_i,则包含 n 个地址的 HL 形式化表示(1):

$$HL = \{X_1, X_2, \cdots, X_i, \cdots, X_n\},$$
$$X_i = x_1^i x_2^i \cdots x_j^i \cdots x_{32}^i,$$
$$x_j^i \in \{'0', '1', \cdots, 'f'\}. \quad (1)$$

(1)多层级密度聚类。按不同前缀长度 32~64,以 8 位间隔来划分层级,采用 32 作为最小长度,而 64 比特的位置一般作为网络标识和接口标识的分界。对于层级 $Level_p$ 形式化表示(2):

$$Level_p = \{x_9^1 x_{10}^1 \cdots x_{p/4}^1, x_9^2 x_{10}^2 \cdots x_{p/4}^2, \cdots, x_9^i x_{10}^i \cdots x_{p/4}^i,$$

$$\cdots, x_9^{n1} x_{10}^{n1} \cdots x_{p/4}^{n1}\} \qquad p \in \{40, 48, \cdots, L\} \tag{2}$$

再以各 $Level_p$ 的前缀值 $x_8^i x_9^i \cdots x_{p/4}^i$ 依次作为密度聚类的输入，从 X_i 第 8 位开始是由于/32 前缀固定。采用 DBSCAN 密度聚类算法，设聚类成 K 类，对 $Level_p$ 的第 k 类的 l 个值升序排序得集合 C_k^p，用数学公式可表示为式(3)：

$$C_k^p = \{C_{k,1}^p, C_{k,2}^p, \cdots, C_{k,j}^p, \cdots, C_{k,l}^p\},$$
$$C_{k,j}^p = x_9^{k,j} x_{10}^{k,j} \cdots x_{p/4}^{k,j}. \tag{3}$$

将 C_k^p 相邻值间的缺失值作为预测 CP_k^p，p 越大对应 CP_k^p 预测优先级越高，表示为公式(4)：

$$CP_k^p = \{C_{k,1}^p + 1, C_{k,1}^p + 2, \cdots, C_{k,2}^p - 1, C_{k,2}^p + 1, C_{k,2}^p + 2, \cdots, C_{k,l}^p - 1\}. \tag{4}$$

(2)启发式后缀拓展。对于预测的长度比 L 短的前缀，采用启发式拓展后缀的方法，具体的对某个预测的前缀，采用步骤(1)中对应的聚类类别后缀集合 CS_k^p 来拓展，其中 $C_{k,j}^p$ 为包含 m 个元素的后缀集合，并设定这些后缀集合的交集 $CSIN_k^p$ 为预测高优先级，相应的形式化表示为式(5)：

$$CS_k^p = \bigcup_{j=1}^{l} CS_{k,j}^p,$$
$$CS_{k,j}^p = \{CSE_{k,j,1}^p, CSE_{k,j,2}^p, \cdots, CSE_{k,j,t}^p, \cdots, CSE_{k,j,m}^p\},$$
$$CSE_{k,j,t}^p = x_{p/4+1}^{k,j,t} x_{p/4+2}^{k,j,t} \cdots x_{L/4}^{k,j,t},$$
$$CSIN_k^p = \bigcup_{j=1}^{l} CS_{k,j}^p. \tag{5}$$

结合(1)(2)步，得到 $Level_p$ 的预测的 IPv6 存活前缀列表 HPL_p，形式化表示为式(6)：

$$HPL_p = \bigcup_{k=1}^{K} HP_k^p,$$
$$HP_k^p = \{x_1 x_2 \cdots x_8\} \times CP_k^p \times CS_k^p. \tag{6}$$

最终合并不同 HPL_p 得到输出 HPL。

1.4　IPv6 网络拓扑测量目标选取

在 IPv6 网络拓扑测量目标选取一般采用对 BGP 路由前缀列表均匀随机抽样的策略。基于这种均匀随机抽样的方法，本文从 IPv6 存活地址列表的角度出发，分为两步选取 IPv6 网络拓扑测量的目标：

(1)对存活地址列表按/64 前缀均匀随机抽取；

(2)根据 IPv6 存活地址前缀列表预测算法，得到预测的/64 前缀列表，并对每个前缀拼接随机接口标识，最终合并这两步的输出。基于 IPv6 存活地址列表的 IPv6 网络拓扑测量目标选取的综合方案如图 2 所示。

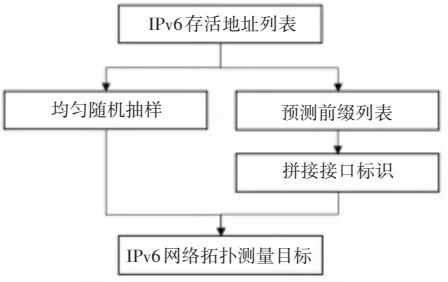

图 2　目标选取的综合方案

Fig. 2　Comprehensive scheme of target selection

2　实验结果与分析

2.1　IPv6 存活地址列表收集与融合

合并不同来源的存活地址列表，得到融合的 IPv6 存活地址列表约 17M。不同来源的 IPv6 存活地址列表统计结果见表 2，可以观察到不同来源在融合的 IPv6 存活地址列表中贡献的差异。其中，独占：融合的 IPv6 存活地址列表中仅由该源提供的地址数量；占比：采用该源的地址与融合的 IPv6 存活地址列表交集数占融合的 IPv6 存活地址列表的比例。可以发现不同来源占比以及独占比差距不大，CAIDA_Ark 甚至基本一致，说明这些不同来源背后采集技术的不同，导致不同来源存活地址列表独立性强，都应该被采用。

表 2　IPv6 存活地址列表统计

Tab. 2　IPv6 Hitlists statistics

收集源	方法	属性	时间	清洗前地址数	清洗后地址数及占比	独占地址数及独占比
Rapid7_FDNS	Fwd. DNS	Servers	2020/1/25	4 610 956	4 517 992 (27.00%)	2 728 797 (16.31%)
CAIDA_Ark	Traceroute	Routers	2020/02	5 382 052	5 312 785 (31.75%)	5 139 168 (30.71%)
Bitnodes	Active	Mixed	2020/2/25	1 208	1 160 (0.01%)	628 (<0.01%)
TUM_Responsive	Active	Mixed	2020/2/22	6 390 778	6 390 778 (38.19%)	5 684 819 (33.97%)
TUM_Seeds	Collection	Mixed	2020/1/4	3 891 940	3 876 883 (23.17%)	2 393 567 (14.30%)

2.2 IPv6 存活地址特征分析

IPv6 存活地址特征主要针对最终融合的 IPv6 存活地址列表。

2.2.1 IPv6 网络前缀分析

不同长度前缀频率分布结果如图 3 所示。数据包括最终融合的 IPv6 存活地址列表 Total 及其最主要的 3 个收集来源，Rapid7_FDNS、CAIDA_Ark 和 TUM_Responsive，可以观察到不同来源 IPv6 存活地址列表的频率分布存在共性，随着前缀长度的增加，在 [32,64] 的区间内呈现明显的上升趋势，而在 [64,128] 的区间内趋于平缓，表明存在大量的 IPv6 存活地址聚集在 64 长度以下的 IPv6 前缀中。

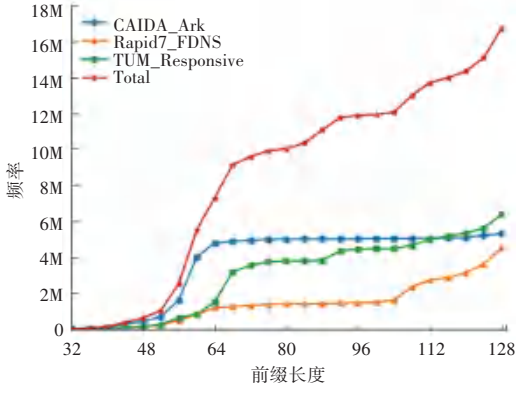

图 3 IPv6 存活地址列表不同长度前缀频率分布

Fig. 3 Different length prefix frequency distribution of IPv6 Hitlists

融合 IPv6 存活地址列表的不同长度前缀 ACR 值分布如图 4 所示。可以观察到不同相邻前缀的长度差下 ACR 值随着前缀长度的增长，总体在 [32,64] 区间变化大且多呈现下降趋势，在 [64,128] 区间变化较小且增减交替，相邻前缀的长度差越小，ACR 值变化越多，表明相邻前缀的长度差越小，越能反应相邻前缀的差异。

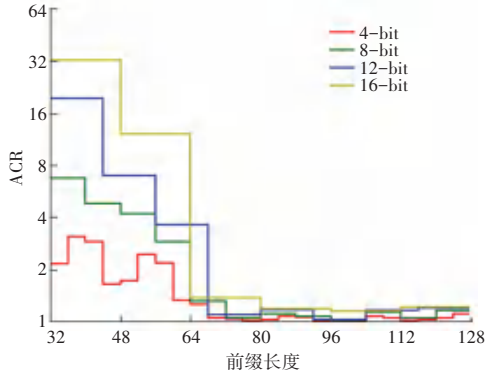

图 4 IPv6 存活地址列表不同长度前缀 ACR 值分布

Fig. 4 Different length prefix ACR's value distribution of IPv6 Hitlists

融合 IPv6 存活地址列表的不同长度前缀密度分布如图 5 所示。可以观察到 [32,64] 区间内前缀长度越短，相对地址密度越大，在前缀长度 [40,128] 区间包含存活地址数的中位数都为 1，甚至在 [60,128] 上四分位数和下四分位数也为 1，表明该区间的大多数前缀仅包含 1 个存活地址，同时 [32,40] 区间的中位数都不超过 10，表明 IPv6 存活地址列表中不同长度的特定前缀包含存活地址都偏少，数值在 1~10 之间，呈现出低密度性。

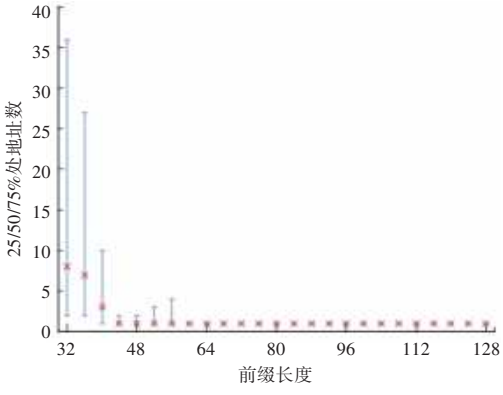

图 5 IPv6 存活地址列表不同长度前缀密度分布

Fig. 5 Different length prefix dense distribution of IPv6 Hitlists

2.2.2 IPv6 接口标识分析

IPv6 接口标识(IID)分析的结果见表 3。可以观察到，随机化的地址在不同来源中占据大部分，为 50%~80%，表明大多数接口标识是随机的，而基于 IPv4 地址和端口号的生成方案地址在不同来源中占比较小，基于低字节和 IEEE EUI-64 方案的地址占比相对较多。此外，不同收集来源下接口标识生成方案占比分布不同，除随机化地址外，Rapid7_FDNS 中基于低字节方案的地址最多，而 CAIDA_Ark 和 TUM_Responsive 中 EUI-64 地址最多。

表 3 不同 IID 模式在存活地址列表中占比统计

Tab. 3 Statistics of the proportion of different IID patterns in the Hitlists

IID 模式	Rapid7_FDNS	CAIDA_Ark	TUM_Responsive	Total
Low-byte	28.54	6.33	17.70	13.28
Embed-IPv4	11.16	1.10	5.74	5.52
Embed-port	0.07	0.01	0.05	0.03
IEEE-based	10.26	8.91	20.63	13.94
Randomized	49.07	83.51	55.22	66.80

综上，融合的 IPv6 存活地址列表呈现高聚集，即大量地址聚集在较短的前缀中；多层次，即长度差越小，越能反应相邻前缀的差异性；低密度，即不同长度前缀包含存活地址数的中位数为 1~10；接口标识不可预测性，即大部分接口标识是随机的。

2.3　IPv6 存活地址前缀列表预测

为了评估预测算法的效果，针对特定/32 前缀网络，对 IPv6 存活地址列表过滤得到/64 存活前缀列表，对该列表的随机 50%作为训练集，剩余 50%作为测试集，计算本文算法 PrefixPrediction（PP）的正确率，并与 Entropy/IP（EIP）结果作比较。预测集：输出的与测试集相同数量的预测前缀列表；正确率：预测集与测试集的交集占测试集的比例。本文在 3 个数据集 H1、H2 和 H3 做实验，实验结果见表 4。

表 4　PrefixPrediction 与 Entropy/IP 实验结果对比

Tab. 4　Comparison of experimental results between PrefixPrediction and Entropy/IP

数据集	/64 存活前缀数	PP 预测正确数量及占比	EIP 预测正确数量及占比	PP 和 EIP 预测正确交集数量和占比
H1（2001：16b8：：/32）	169 790	1 705（2.01%）	1 217（1.43%）	33（0.04%）
H2（2a02：06b8：：/32）	7 016	2 259（64.40%）	592（16.88%）	493（14.05%）
H3（240e：00e0：：/32）	318 747	38 960（24.49%）	7 159（4.49%）	2 672（1.68%）

实验结果表明，本文的前缀预测算法 PrefixPrediction 正确率相比 Entropy/IP 高，在 1.4～5.5 倍之间，同时二者预测前缀的交集小，有强互补性。预测集增长下的正确率变化趋势如图 6 所示，可以发现随着预测集增长，PrefixPrediction 比 Entropy/IP 正确率都要高，二者的正确率都呈现线性增趋势，表明两种算法的预测能力都具有稳定性。

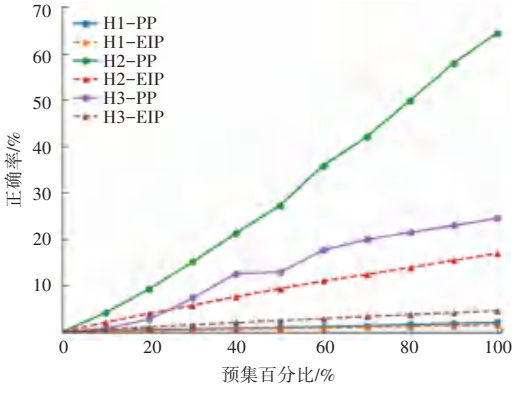

图 6　预测集增长下正确率变化趋势

Fig. 6　The trend of accuracy under the growth of prediction set

2.4　IPv6 拓扑测量目标选取

为了验证本文基于 IPv6 存活地址列表的 IPv6 网络拓扑测量目标选取方法的效果，针对特定/32 可路由前缀的目标网络，分别采用本文方法 HS 和按/64 均匀随机抽样方法 URS，对一定数量目标采用 ICMP-paris 的探测方法，对比结果的节点数和链接数。目标网络包括 T1（2001：16b8：：/32）、T2（2a02：06b8：：/32）和 T3（240e：00e0：：/32）。本文方法中对存活地址列表按/64 均匀随机抽样数的 2 倍，相应的 URS 为从均匀随机抽样的地址中再随机抽取该数量；traceroute 路径中发现的不同 IPv6 接口地址数；traceroute 路径中发现的不同 IPv6 接口地址连接数。本文对 3 个数据集 T1、T2 和 T3 分别采用不同测量点做实验，实验结果见表 5。

实验结果表明，本文基于 IPv6 存活地址列表的 IPv6 网络拓扑测量目标选取方法 HS 比均匀随机抽样方法 URS 明显提高了拓扑发现的完整性，拓扑新发现率在 94%以上，拓扑发现结果随探测目标增长变化如图 7 和图 8 所示，HS 相比 URS 结果表现一直要好。拓扑的完整性即发现的节点数和链接数，拓扑新发现率即 HS 相比 URS 新发现的节点和链接占 HS 的比率。

表 5　HS 和 URS 实验结果对比

Tab. 5　Comparison of experimental results between HS and URS

目标网络	目标数	测量点	HS		URS		HS-URS	
			节点数	链接数	节点数	链接数	节点数	链接数
T1	339 580	VP#NL	99 559	99 647	32 782	32 875	98 318	98 140
T2	14 032	VP#UK	5 414	5 500	11	11	5404	5490
T3	637 494	VP#US	125445	158 629	27 511	64 346	122 516	148 412

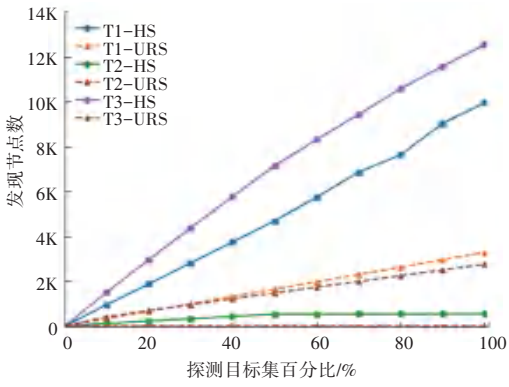

图 7 发现节点数随探测测量目标数增长变化

Fig. 7 The number of nodes found varies with the number of measured targets

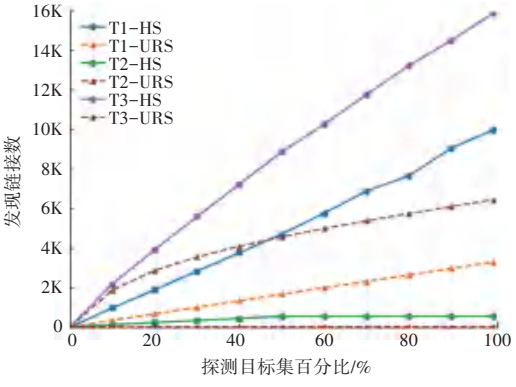

图 8 发现链接数随探测目标数增长变化

Fig. 8 The number of links found varies with the number of measured targets

3 结束语

本文提出一种 IPv6 网络拓扑测量目标选取技术,用于提高 IPv6 网络拓扑测量的有效性和完整性。首先,收集并融合不同来源的 IPv6 存活地址列表得到约 17M 的地址;其次,分析了 IPv6 存活地址列表的特征,发现融合存活地址列表呈现出高聚集、多层次、低密度和接口标识不可预测性,提出了 IPv6 存活地址前缀列表的预测算法 PrefixPrediction,对比 Entropy/IP 预测的正确率更高,发现二者结果具有强互补性;最后,给出了 IPv6 网络拓扑测量目标选取的综合方案HS,对比URS 明显提高拓扑发现的完整性,拓扑新发现率超过 94%。未来将进一步丰富收集技术,提高 IPv6 存活地址列表的完整性,结合多种 IPv6 存活地址前缀预测算法来提高预测正确率,从而更深入地进行 IPv6 网络拓扑测量的研究。

参考文献

[1] CAIDA. The IPv6 Topology Dataset [EB/OL]. [2020]. http://www.caida.org/data/active/ipv6_allpref_ topology_dataset.xml.

[2] ROHRER J P,LAFEVER B, BEVERLY R, et al. Empirical Study of Router IPv6 Interface Address Distributions[J]. IEEE Internet Computing, 2016, 20(4): 36-45.

[3] BORGOLTE K, HAO S, FIEBIG T, et al. Enumerating Active IPv6 Hosts for Large-Scale Security Scans via DNSSEC-Signed Reverse Zones[C]// IEEE Symposium on Security and Privacy. San Francisco, California, USA: IEEE, 2018: 770-784.

[4] FIEBIG T, BORGOLTE K, HAO S, et al. Something from Nothing (There): Collecting Global IPv6 Datasets from DNS [C]// Passive and Active Network Measurement. Sydney, NSW, Australia: Springer, 2017: 30-43.

[5] GASSER O,SCHEITLE Q, FOREMSKI P, et al. Clusters in the Expanse: Understanding and Unbiasing IPv6 Hitlists [C]// Internet Measurement Conference. Boston, MA, USA: ACM, 2018: 364-378.

[6] ULLRICH J, KIESEBERG P, KROMBHOLZ K, et al. On Reconnaissance with IPv6: A Pattern-Based Scanning Approach [C]// Availability, Reliability and Security. Toulouse, France: IEEE, 2015: 186-192.

[7] FOREMSKI P, PLONKA D, BERGER A W, et al. Entropy/IP: Uncovering Structure in IPv6 Addresses [C]// Internet Measurement Conference. Santa Monica, CA, USA: ACM, 2016: 167-181.

[8] MURDOCK A, LI F, BRAMSEN P, et al. Target generation for internet - wide IPv6 scanning [C]// Internet Measurement Conference. London, United Kingdom.: ACM, 2017: 242-253.

[9] BEVERLY R,DURAIRAJAN R, PLONKA D, et al. In the IP of the Beholder: Strategies for Active IPv6 Topology Discovery [C]// Internet Measurement Conference. Boston, MA, USA: ACM, 2018: 308-321.

[10] PLONKA D, BERGER A W. Temporal and Spatial Classification of Active IPv6 Addresses[C]// Internet Measurement Conference. Tokyo, Japan: ACM, 2015: 509-522.

[11] 张千里,姜彩萍,等. IPv6 地址结构标准化研究综述[J]. 计算机学报,2019,42(6):1384-1405.