6EDL: 高效的大规模活跃IPv6地址探测系统

宋光磊[1),2)] 张文健[1)] 林金磊[1)] 韩东岐[1)] 王之梁[1),2)]
张 辉[1),3)] 杨家海[1),2)]

[1)](清华大学网络科学与网络空间研究院 北京 100084)
[2)](中关村实验室 北京 100081)
[3)](泉城实验室 济南 250000)

摘 要 互联网规模的急剧增长导致IPv4地址资源逐渐耗尽,IPv6的大规模部署有效地解决了IPv4地址耗尽的问题.然而,IPv6地址空间庞大的特性为活跃地址的探测带来了巨大挑战.当前已有的活跃IPv6地址探测方法存在探测速度较慢、命中率偏低、探测范围受限等问题.针对这些挑战,本文提出了高效、快速、适用范围广的活跃地址探测系统6EDL.6EDL将地址探测分为无种子地址场景和有种子地址场景,并针对每种场景设计高效探测算法.在无种子地址场景下,本文提出了6EDL-N,使用神经网络挖掘BGP前缀信息与地址配置模式之间的潜在关系,实现了有种子区域到任一无种子区域的地址迁移,从而扩展了地址探测的边界.此外,本文引入了预探测机制,有效缓解了大规模地址探测中的资源浪费问题.实验表明,6EDL-N的命中率达到12.69%,覆盖度为21.97%,单位时间发现的活跃地址数(NPT)为233.09个/s.与现有工作相比,6EDL-N的命中率是其的8.13倍,NPT为14.94倍,覆盖度为1.84倍.在有种子地址场景下,本文提出基于生成对抗网络(GAN)的活跃地址探测方法6EDL-S,通过精细的种子地址分布规律学习,并采用环境反馈机制来缓解种子地址采样偏差,有效提升了命中率.实验表明,6EDL-S的命中率达到了25.91%,是已有方法的1.23~10.89倍.同时,NPT为466.72个/s,是已有方法的1.49~6.20倍.最终,经过持续探测,6EDL系统成功发现了29.77亿个活跃地址,包含5.66亿别名地址和24.11亿非别名地址,覆盖了125 101个BGP前缀和40 137个AS.本文构造的活跃IPv6地址集将有效支撑IPv6网络测量和安全分析等多种应用,进一步打开了IPv6网络研究的大门.

关键词 网络测量;IPv6;活跃地址探测;机器学习;IPv6活跃地址集
中图法分类号 TN915 **DOI号** 10.11897/SP.J.1016.2024.01949

6EDL: Efficient Large-Scale Active IPv6 Address Probing System

SONG Guang-Lei[1),2)] ZHANG Wen-Jian[1)] LIN Jin-Lei[1)] HAN Dong-Qi[1)]
WANG Zhi-Liang[1),2)] ZHANG Hui[1),3)] YANG Jia-Hai[1),2)]

[1)](*The Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing* 100084)
[2)](*Zhongguancun Laboratory, Beijing* 100084)
[3)](*Quancheng Laboratory, Jinan* 250000)

Abstract The rapid growth of the size of the Internet has led to the gradual depletion of IPv4 address resources, and the large-scale deployment of IPv6 has effectively solved the problem of IPv4 address exhaustion. However, the vast expanse of the IPv6 address space presents

收稿日期:2023-11-08;在线发布日期:2024-04-30.本课题得到国家重点研发项目互联网IP地址空间与域间路由系统关键信息感知技术(No.2022YFB3105001)、中关村实验室项目、清华大学-中国电信集团有限公司下一代互联网技术联合研究中心项目资助.**宋光磊**,博士,助理研究员,中国计算机学会(CCF)会员,主要研究领域为地址探测、网络测量.E-mail:songgl@mail.zgclab.edu.cn.**张文健**(共同第一作者),硕士研究生,主要研究领域为地址探测、网络测量.E-mail:zhang-wj23@mails.tsinghua.edu.cn.**林金磊**,博士研究生,主要研究领域为网络测量和网络安全.**韩东岐**,博士研究生,主要研究领域为态势感知网络测量.**王之梁**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为网络空间态势感知、网络测量.**张 辉**,硕士,高级工程师,主要研究领域为网络测量、网络管理.**杨家海**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为下一代互联网体系结构网络测量等.Email:yang@cernet.edu.cn.

significant challenges for the detection of active IPv6 addresses. Existing methods for detecting active IPv6 addresses suffer from issues such as slow speed, low hit rates, and limited detection coverage. To address these challenges, we propose the active IPv6 address detection system 6EDL, which is efficient, fast, and broadly applicable. 6EDL divides active IPv6 address detection into scenarios without seed addresses and with seed addresses (e. g., the known active IPv6 addresses), and designs efficient active IPv6 address detection algorithms tailored to each scenario. In the scenario without seed addresses, we propose the 6EDL-N method. This method uncovers the latent relationship between BGP prefix information and address configuration patterns, enabling the migration of addresses from areas with seed addresses to any area without seed addresses, thereby extending the boundary of address detection. Additionally, we establish a pre-scanning mechanism that effectively mitigates resource waste in large-scale address detection. The experimental results show that in the scenario without seed addresses, 6EDL-N achieves a hit rate of 12.69%, a coverage rate of 21.97%, and the number of active IPv6 addresses discovered per unit time (NPT) of 233.09 addresses per second. Compared with existing methodologies, 6EDL-N exhibits a remarkable improvement in hit rate, which is 8.13 times higher than current methods, and an NPT that is 14.94 times higher than the current methods. Additionally, its coverage is 1.84 times greater than that of existing methods. In the scenario with seed addresses, we propose the 6EDL-S method for active IPv6 address detection based on adversarial networks. This method leverages precise learning of seed address distribution patterns and employs an environmental feedback mechanism to mitigate seed address sampling biases, effectively enhancing the hit rate. Experimental results demonstrate that the hit rate of 6EDL-S reaches 25.91%, which is 1.23～10.89 times higher compared with existing methods. Furthermore, NPT is 466.72 addresses per second, which is 1.49～6.20 times higher than existing methods. Ultimately, through continuous active IPv6address probing, the 6EDL system successfully discovered 29.77 billion active IPv6 addresses, comprising 5.66 billion alias addresses and 24.11 billion non-alias addresses, encompassing 125 101 BGP prefixes and 40 137 autonomous systems. The active IPv6 address set (IPv6 hitlist) constructed in our work will effectively support various applications such as IPv6 network measurement and security analysis, further opening the door to IPv6 network research.

Keywords network measurement; IPv6; active address detection; machine learning; IPv6 hitlist

1 引 言

全球互联网范围的活跃地址扫描技术是大规模网络测量和网络调查的基础前提，支撑多种网络上层应用[1-7]. 例如，活跃地址扫描技术支持网络拓扑探测，揭示网络设备的连接关系，深入了解网络内部结构[1-2]；支持探测特定技术或者服务的部署情况，调研其可用性和性能[3-4]；支持发现上层指纹信息，评估网络资产，感知资产态势[5-6]；支持感知网络关键节点，识别设备漏洞，维护网络安全[7-17]. 活跃地址扫描技术对网络空间测绘、网络攻防乃至网络安全具有重要意义.

得益于网络快速扫描能力，在IPv4(Internet Protocol Version 4)空间能快速实现活跃IPv4地址扫描. 例如，Masscan①能够在短短5分钟内完成IPv4网络上特定端口的活跃地址扫描. 然而，随着云计算、5G技术和物联网的迅速崛起，网络中的设备数量急剧增加，这导致了IPv4地址的日益枯竭. 2019年11月26日，欧洲IP网络资源协调中心(RIPE Network Coordination Centre，RIPE NCC)宣布其可用资源池中最后的/22前缀全部分配完毕，这意味着IPv4的整个地址空间已经完全耗尽. 为应对IPv4地址枯竭问题，IPv6(Internet Protocol

① Graham R. MASSCAN. https://github.com/robertdavidgraham/masscan. 2023,10,20.

Version 6)协议在全球加速部署. 截至2023年9月21日，超过43.3%的网络用户通过接入IPv6网络访问Google服务①. IPv6的部署虽然有效解决了IPv4地址耗尽的问题，但其巨大的地址空间给活跃地址扫描带来巨大挑战. 使用当前最快扫描器暴力扫描整个IPv6地址空间，至少要消耗数百万年时间. 因此，暴力扫描整个IPv6地址空间是不可行的.

针对暴力扫描整个IPv6地址空间不可行的问题，研究人员尝试使用开放数据获取[11,18-22]、被动采集[23-25]和主动探测[11,18,26-37]等诸多方法探测活跃IPv6地址，这在一定程度上可发现部分活跃IPv6地址. 但是，这些方法存在诸多限制和挑战. 公开数据获取，例如域名系统(Domain Name System，DNS)解析[19-21]获取的IPv6地址存在分布不均衡、分布范围小等问题；被动采集通过网络关键节点获取网络流量或者日志提取活跃IPv6地址，该方法需要获取网络关键节点权限，且发现的活跃IPv6地址分布范围受限；主动探测方法通过主动对生成的候选地址发送探测包来发现活跃IPv6地址，当前方法不能有效权衡命中率和探测速度，探测范围受限. 例如，已有方法6Graph[34]虽然具有较高的命中率，但由于其过高的时间复杂度，导致每秒只能发现82个活跃IPv6地址. 因此，该方法在大规模地址探测中效率低. 此外，该方法只能应用于有种子的地址场景，在无种子地址探测场景失效. 即使后来提出的AddrMiner-N[18]实现了无种子地址场景探测，但其探测效率低下，扩展地址探测边界有限. 因此，活跃地址探测方法面临速度慢、命中率低、探测范围受限的问题. 当前仍然缺少更加高效、全面的全球活跃IPv6地址探测方法.

针对当前挑战，本文采取主动探测方法探测全球活跃IPv6地址，根据被探测区域是否包含种子地址将地址探测分为两种探测场景：无种子地址场景和有种子地址场景. (1)在无种子地址场景，针对探测效率低、探测范围受限的问题，本文提出基于深度学习的BGP前缀相似度算法6EDL-N. 该无种子地址探测方法基于深度神经网络来挖掘BGP前缀信息与地址配置模式之间的关系，并通过寻找与无种子目标区域相似的有种子区域进行地址迁移，为无种子BGP前缀地址探测开辟了新思路. (2)在有种子地址场景，针对当前方法不能有效权衡资源利用率和探测算法时间效率的问题，本文提出有种子地址探测方法6EDL-S. 通过简洁、高效的深度学习模型学习种子地址的分布特征，生成可能存活的候选地址. 同时，使用反馈机制调整地址生成方向，缓解种子地址抽样偏差对探测效率的影响. 该算法参数少、学习效率高，在提升探测命中率的同时，保持快速的活跃地址发现能力. (3)除此之外，本文还设计了更全面的评价体系，使用命中率、单位时间内发现的活跃地址数量、覆盖度来衡量模型的效率、探测范围. 实验证明，在无种子地址区域，6EDL-N在每个BGP前缀下，选取预算(候选地址/目标地址的数量)为1千、1万、10万、100万进行比较分析. 在前缀预算为100万时，命中率为12.69%，覆盖度为21.97%，单位时间内发现活跃IPv6地址数量(NPT)为233.09个/s，发现了29.77亿种子地址. 与AddrMiner-N[18]相比，6EDL-N的命中率是AddrMiner-N的8.13倍，NPT为其的14.94倍，覆盖度为其的1.84倍. 在有种子地址区域，6EDL-S模型的活跃地址命中率是已有方法的1.23～10.89倍，NPT高达466.72个/s，是当前方法的1.49～6.20倍.

本文在无种子区域、有种子区域均提出了活跃地址探测策略，可以进行全面的IPv6活跃地址探测，本文的主要贡献如下：

(1)设计了高效的无种子区域探测方法6EDL-N，利用神经网络挖掘BGP前缀信息与地址配置模式之间的关联关系，实现在无种子区域快速、高命中率的活跃地址探测，进一步扩大地址探测边界，提升探测覆盖度；

(2)提出高效的有种子区域探测方法6EDL-S，通过深度学习实现对种子地址的精细学习，实现在有种子区域进行大规模活跃地址探测时的高命中率和低时间复杂度；

(3)提出系统化的地址探测评价指标，多角度、更全面地衡量活跃IPv6地址探测性能；

(4)部署6EDL系统，通过持续探测成功收集到了29.77亿个活跃地址，覆盖了125 101个BGP前缀和40 137个AS.

2　研究背景

本节介绍全球活跃IPv6地址探测的相关研究背景以及IPv6地址的特性.

① Google. IPv6 Adoption Statistics, https://www.google.com/intl/en/ipv6/statistics.html#tab=ipv6-adoption, 2023, 10, 20.

互联网的迅猛发展使得IPv6部署成为必然趋势，全球范围内各方积极倡导和推动IPv6网络的实施. 一方面，大型企业基于应用驱动快速推进IPv6部署. 2012年全球范围内的网络运营商、网站运营商和设备制造商，如Akamai、Google、Cisco等，联合成立了“World IPv6 Launch”，承诺他们的产品和服务将永久支持IPv6. 自此，IPv6网络在全球范围内取得了飞速的发展，过去十年间，全球IPv6流量增长超过了5000%①. 此外，国际互联网组织积极提倡全球IPv6的推广；互联网架构委员会(Internet Architecture Board，IAB)建议互联网工程任务组(IETF)在新的RFC(Request For Comments)标准中停止要求新设备和扩展协议必须兼容IPv4，鼓励新协议在IPv6的基础上进行优化②；IPv6受到全球各国政府的高度重视，成为各国在网络空间博弈的新战场. 例如，美国联邦管理和预算办公室(Office of Management and Budget，OMB)发布了《IPv6过渡计划新草案指南》③，要求截止到2025财年美国联邦网络上IPv6的网络资产占比应至少达到80%；欧盟委员会鼓励欧盟成员加速部署IPv6④，要求大型企业和运营商推进IPv6部署；我国各政府部门联合出台一系列政策文件以促进IPv6的部署. 特别是在2021年发布《IPv6流量提升三年专项行动计划(2021—2023年)》，要求到2023年底全国基本形成应用驱动、协同创新的良性发展格局.

IPv6地址长度为128位，它由三部分构成：全局路由前缀、本地子网标识符和接口标识符(Interface Identifier，IID). IPv6地址通常使用八组四位十六进制字符表示，每组含有16位，用冒号(“:”)分隔. 例如，IPv6地址2001:0db8:0000:0000:0008:8000:200c:417a，该地址也可以简化表示为2001:db8::8:8000:200c:417a. 相对IPv4地址，IPv6地址具有以下主要特点：

(1)IPv6地址空间巨大：IPv6地址包含了128位可变比特，其空间大小为2^{128}. 与IPv4的32位地址空间(总共2^{32}个地址)相比，IPv6的地址空间是IPv4的2^{96}倍. 巨大的地址空间有效地解决了IPv4地址资源枯竭的问题.

(2)IPv6地址分布稀疏：相比IPv6巨大的地址空间，网络中活跃网络设备较少. 因此，整个IPv6地址空间内活跃IPv6地址的密度非常低. 如果假设IPv6网络设备和IPv4网络设备的规模相等，那么整个IPv6地址空间内活跃IPv6地址的密度大约是IPv4的$\frac{1}{2^{96}}$.

(3)IID分配类型多样：IPv6中的接口标识符可以以多种方式进行配置(无状态自动配置⑤，IPv6动态主机配置协议⑥⑦或静态分配生成). 配置方法的多样性导致IID分配类型多样. 常见的IID类型包括Random⑧、Low-byte⑦、Embedded-IPv4⑨、Teredo隧道⑩、EUI-64⑪、Embedded-word⑫、Embedded-port⑦等.

IPv6巨大的地址空间导致暴力扫描整个IPv6地址空间不可行. 同时，IPv6地址分布稀疏的特性和IID分配方式多样的特性进一步增加了活跃IPv6地址在网络空间中的隐匿性，导致活跃IPv6地址探测更加困难.

3 相关工作

根据地址获取方式的不同，当前IPv6活跃地址探测方法可以分为以下三个主要类别：(1)开放数据获取、(2)被动采集和(3)主动探测.

3.1 开放数据获取

开放数据获取是指依赖于公开可用的网络数据资源，并通过查询或解析方法来提取这些地址. 这种方法充分利用了各种公开数据源，其中一些代表性的资源包括：域名系统(Domain Name System，DNS)[19-21]、众包平台[11,22]和公开数据源[11,18].

① IPv6 Launch. IPv6 is the New Norma. https://www.worldipv6launch.org/. 2023,10,20.

② IAB. IAB Statement on IPv6. https://www.iab.org/2016/11/07/iab-statement-on-ipv6/. 2023,10,20.

③ OMB to Agencies: Time to Finish IPv6 Transition. https://www.nextgov.com/it-modernization/2020/03/omb-ipv6-transition/163459/. 2023,10,20.

④ APNIC. IPv6 Capable Rate by Country. https://stats.labs.apnic.net/ipv6. 2023,10,20.

⑤ Thomson S, Narten T, Jinmei T. IPv6 stateless address autoconfiguration. 2007. https://doi.org/10.17487/RFC4862. 2023,10,20.

⑥ Dynamic Host Configuration Protocol for IPv6 (DHCPv6). RFC, 2018, 8415: 1-154. https://doi.org/10.17487/RFC8415. 2023,10,20.

⑦ Dynamic host configuration protocol for ipv6 (DHCPv6). RFC, 2003, 3315: 1-101. https://doi.org/10.17487/RFC3315. 2023,10,20.

⑧ Network Reconnaissance in IPv6 Networks. RFC, 2016, 7707: 1-38. https://doi.org/10.17487/RFC7707. 2023,10,20.

⑨ IPv6 Addressing of IPv4/IPv6 Translators. RFC, 2010, 6052: 1-18. https://doi.org/10.17487/RFC6052. 2023,10,20.

⑩ Tunneling IPv6 over UDP through Network Address Translations (NATs). RFC, 2006, 4380: 1-53. https://doi.org/10.17487/RFC4380. 2023,10,20.

⑪ IP Version 6 Addressing Architecture. RFC, 2006, 4291:1-25. https://doi.org/10.17487/RFC4291. 2023,10,20.

⑫ A Recommendation for IPv6 Address Text Representation. RFC, 2010, 5952: 1-14. https://doi.org/10.17487/RFC5952. 2023,10,20.

（1）DNS解析：DNS可靠地提供获取活跃IPv6地址的途径[19-21]. 研究者通过多种渠道获取候选域名，例如IPv4地址反查地址指针记录（Pointer Record，PTR）、NXDOMAIN记录提取等，然后对候选域名进行逆向查询，并从逆向查询的AAAA记录中获取活跃IPv6地址. 尽管上述方法可以获得活跃IPv6地址，但存在着探测到的活跃IPv6地址数量较为有限以及探测效率较低的挑战.

（2）众包平台：一些学术研究借助众包平台志愿者的请求数据，致力于提取活跃IPv6地址[11,22]. Huz等研究者率先在亚马逊的MTurk[22]平台上进行了IPv6应用的评估工作，其结果涵盖了38个活跃IPv6地址. Gasser等人从MTurk和ProA①平台的6967名参与者中获取了2032个IPv6地址[11]. 不过，这类方法所需成本较高. 例如，Gasser等人获取活跃IPv6地址的成本高达0.15美金/个. 因此，该方法不适用于全球范围内大规模的IPv6地址活跃性探测.

（3）公开数据源：网络中存在多种公开的数据源，这些数据源可被有效地利用以构建活跃IPv6地址集[11,18]. Gasser等研究人员从多个公开数据源（例如公开域名列表②③④、FDNS⑤、CT证书⑥、AXFR和TLDR⑦、Bitnodes⑧、RIPE Atlas⑨等）中提取IPv6地址，最终成功地获取了超过5千万个活跃IPv6地址. 然而，这种方法存在一些问题，包括活跃IPv6地址数量有限、分布不均匀以及范围受限等挑战. 尽管如此，该方法在一定程度上仍然有助于发现全球范围内的活跃IPv6地址，特别是IPv6服务器地址.

3.2　被动采集

被动采集是指通过获得网络关键节点权限，并从这些节点的日志或流量数据中提取IPv6活跃地址[23-25]. 具体来说，Plona等人的研究利用Akamai公司启用了IPv6的内容分发网络（Content Delivery Network，CDN）服务器作为关键节点[23]，并从这些服务器的请求日志中提取了超过18亿个IPv6活跃地址. 另外，Gasser等人[24]在互联网交换节点（Internet Exchange Point，IXP）采用了1∶10 000的采样比例来获取网络流量，并从采样的流量数据中提取了1.46亿个IPv6活跃地址. 最新的研究中，Rye等人[25]则将NTP服务器作为网络关键节点，从用户的请求记录日志中提取了超过79亿个活跃IPv6地址.

被动采集方法能够全面地获取网络关键节点所在网络的活跃IPv6地址. 例如，Plona等人的工作可以全面解析Akamai组织内启用的CDN服务器地址，而Gasser等人的研究则能够观察到所在自治系统（Autonomous System，AS）和前缀内90%以上的地址. 然而，被动采集方法也存在明显的缺点. 首先，它要求获得网络关键节点的权限，这对于大多数网络研究人员来说是不可行的. 其次，它所收集的活跃IPv6地址仅覆盖有限的空间区域，通常只能观察到少数BGP前缀空间，无法完整反映整个IPv6网络的部署状况.

3.3　主动探测

IPv6活跃地址主动探测根据有无可以学习的种子地址可以分为两个场景，无种子地址区域的主动探测[18]和有种子地址区域的基于种子地址的主动探测[11,26-37].

3.3.1　无种子地址区域的主动探测

无种子地址探测缺乏可以学习的种子地址，基于种子地址的IPv6地址探测方法在无种子区域失效. 且当前公布的活跃IPv6地址集仅覆盖了24.3%的BGP前缀，超过75%的BGP前缀空间无种子地址. 因此，现有方法仅适用于少量的有种子IPv6网络空间，探测范围极其有限. 如何扩大探测边界，实现全球活跃IPv6地址的大规模探测，提升地址集覆盖度是目前面临的巨大挑战.

由于无种子区域探测难度大，没有种子地址进行学习，大部分探测方法失效，因此目前的已有研究进展非常缓慢. Song等人基于图社区聚类的算法提出了AddrMiner-N[18]，AddrMiner-N的核心思想是模式迁移，通过挖掘有种子地址区域的通用地址模式，基于同一组织的地址配置方式具有相似性的假设，将通用地址模式迁移到无种子地址区域生成目标地址，进行无种子区域的探测. AddrMiner-N虽然为无种子地址空间提供可行的探测思路，扩大了活跃IPv6地址探测的范围. 但仍存在：（1）无种子区

① Prolific Academic. https://www.prolific.ac/. 2023,10,20.
② The Spamhaus Project https://www.spamhaus.org/. 2023,10,20.
③ APWG: Cross-industry Global Group Supporting Tackling the Phishing Menace. http://antiphishing.org. 2023,10,20.
④ PhishTank. A Nonprofit Anti-phishing Organization. http://www.phishtank.com. 2023,10,20.
⑤ Rapid7 Project Sonar . Forward dnsdata. https://opendata.rapid7.com/sonar.fdns_v2/. 2023,10,20.
⑥ Media TUM. https://mediatum.ub.tum.de/1452739. 2023,10,20.
⑦ TLD AXFR transfers. https://github.com/mandatoryprogrammer/TLDR. 2023,10,20
⑧ BitnodesAPI. https://bitnodes.earn.com/. 2023,10,20.
⑨ RIPE NCC. IPMap. https://ftp.ripe.net/ripe/ipmap/. 2023,10,20.

域探测范围有限.基于组织关联的通用地址模式迁移,仅能在同一组织内不同子网的模式迁移;(2)无种子地址探测的效率低,图社区聚类的时间复杂度高,无法进行大规模的地址探测;(3)命中率低,存在探测资源的浪费,AddrMiner-N简单的组织匹配策略导致活跃地址命中率低.

3.3.2 基于种子地址的主动探测

有种子地址区域基于种子地址驱动的活跃IPv6地址探测技术,其工作流程如图1所示.首先,从多种公开数据源构造的活跃IPv6地址列表(IPv6 hitlist)中获取种子地址.然后,设计目标地址生成方法学习种子地址的特性并生成可能存活的候选地址(目标地址).紧接着,扫描目标地址获取新的活跃IPv6地址.最后,将发现的活跃IPv6地址进行别名前缀检测,消除别名地址,并将非别名地址加入到IPv6 hitlist.根据目标地址生成方法学习的种子地址结构特征或地址特性的不同,基于种子地址的主动探测方法可以分为基于统计学的主动探测方法[11,18,26-34]和基于机器学习的主动探测方法[35-37].

图1 基于种子地址的主动探测流程

基于统计学的主动探测包括基于种子地址统计特征的主动探测方法[11,27]和基于种子地址特性的主动探测方法[18,29-34].其中,基于种子地址统计特征的主动探测方法旨在通过学习种子地址的内部结构规律,并根据这些内部结构的关联关系来生成目标地址.代表性工作之一是Entropy/IP[11,27],该方法由Foremski等提出[27].具体地,Entropy/IP首次引入熵值来度量地址特定维度的取值情况,通过构建贝叶斯网络生成目标地址.尽管该方法为后续地址探测和可视化研究提供很好的研究思路,但存在以下两个问题:首先,它过度依赖种子地址质量,种子地址抽样偏差导致探测效率低下.其次,Entropy/IP的高时间复杂度$O(n^3)$限制了其发现速度,因此不适用于大规模的活跃IPv6地址探测任务.基于种子地址特性的主动IPv6地址探测方法旨在通过学习种子地址的特性来生成目标地址.现有工作学习的种子地址特性具体包括密度特性和层次特性.具体地,DET[29]、6Gen[30]、AddrMiner-S[18]根据种子地址的密度特性,寻找种子地址高密度区域,并在高密度区域生成目标地址.该方法有效证明在种子地址的高密度区域生成的目标地址更可能是活跃的.6Tree[31]、6Forest[32]、6Hit[33]、6Graph[34]根据种子地址的层次特性将地址空间划分为不同的子网,然后在这些子网中生成目标地址.这类方法考虑了地址路由的寻址过程,极大程度缩小地址探测范围.虽然6Hit和AddrMiner-S提出强化学习机制纠正种子地址抽样偏差,但他们不能很好平衡地址发现速率和探测效率.尤其是前者,过低的地址发现速率使其不能进行大规模活跃地址探测.

基于机器学习的主动探测方法将目标地址生成转化为文本生成问题.Cui等人提出6GAN[35]、6GCVAE[36]、6VecLM[37].其中,6GCVAE构建门控卷积变分自编码器(VAE)[38]模型,引入了两种种子分类技术,有效地提高了深度学习模型的地址生成性能.6VecLM将地址映射到向量空间,并基于Transformer网络结构[39]实现了一个可以生成地址的IPv6语言模型.但这两种方法都没有进行别名检测,且模型参数大.6GAN基于生成对抗网络(GAN)[40]、循环神经网络(LSTM)[41]和强化学习(RL)的方法建立模型,并引入了别名检测机制.但6GAN难以在大规模数据集上进行地址探测.基于机器学习的方法普遍存在运算开销大,难以进行大规模的地址探测、探测和探索难以均衡,导致探测的活跃地址局限于少量BGP前缀.

现有的探测方法存在各种问题,公开数据获取和被动探测存在探测效率低、仅覆盖有限的空间区域、一些探测方法需要管理员权限的问题.主动探测存在探测命中率低,探测时间复杂度高,探测范围有限的问题.针对这些问题,本文使用深度学习的方法,实现全面的大规模主动地址探测,在无种子、

有种子区域均可进行高效探测，实现了高命中率、低时间复杂度、覆盖范围广的主动探测.

4　地址探测框架

本文介绍了一个全球活跃IPv6地址探测框架6EDL，旨在系统、全面地进行大规模的IPv6地址活跃性探测. 图2展示了该框架的工作流程. 首先，它从公共数据源中获取活跃IPv6地址，构建了一个活跃地址列表，即IPv6 hitlist. 然后，根据BGP前缀是否包含种子地址，它将地址划分为两个探测场景：无种子场景和有种子场景. 针对不同的探测场景，6EDL采用不同的策略进行高效的主动地址探测. 在无种子场景中，它采用6EDL-N进行探测，该模块利用神经网络挖掘BGP前缀信息与地址配置模式之间的关系，为无种子BGP前缀提供精准的候选地址，实现有种子地址区域到任意无种子地址区域的迁移生成，在保证高效的地址探测的同时扩大了地址探测边界. 在有种子场景中，使用6EDL-S进行探测，该模块基于生成对抗网络(GAN)[40]，通过学习种子地址的分布规律，能够快速且准确地生成候选地址. 接着，探测到的活跃地址经过别名前缀检测，以消除别名地址，并将去除别名后的活跃地址添加到IPv6 hitlist中. 6EDL框架是一个综合的工具，用于全球活跃IPv6地址的探测，为深入测量和安全分析IPv6网络提供了更多的数据支持. 对比AddrMiner地址探测框架，6EDL将种子地址探测简化为两个场景：无种子场景和有种子场景. 这种简化方法避免了AddrMiner框架中少量种子地址场景难以界定的问题，有效降低了部署难度，更适用于全球活跃IPv6地址的探测.

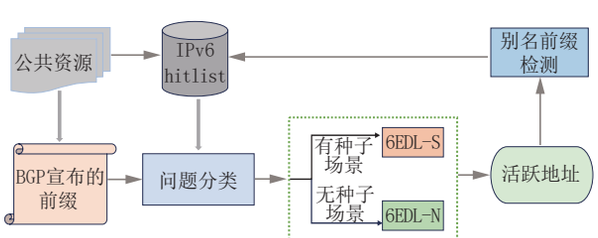

图2　地址探测框架6EDL

5　无种子地址探测模型6EDL-N

由于缺乏可用于学习的种子地址，基于种子地址的活跃地址探测算法不适用于无种子地址的场景. 然而，无种子地址空间在IPv6地址空间中占比超过75%[18]，且目前尚无高效的无种子区域活跃地址探测算法. 现有的无种子地址探测方法存在探测范围受限的问题，导致大量无种子BGP前缀的活跃地址难以被发现. 此外，已有算法的探测效率较低，资源浪费严重，难以实现高效的大规模地址探测. 为了解决上述问题，本文提出了一种基于深度学习的BGP前缀相似度算法6EDL-N. 该算法通过运用深度学习方法挖掘BGP前缀信息与地址配置模式之间的内在关系，从而提高地址探测效率. 此外，6EDL-N能够将有种子区域的地址配置模式更精确地迁移至任意无种子区域，并实现细颗粒度的地址生成，从而扩大地址探测范围，提升活跃地址探测的覆盖度.

5.1　6EDL-N模型

在IPv6地址配置模式中，不同的BGP前缀其地址配置模式之间存在一定的相似性. 这一相似性使得可以将种子地址从有种子区域迁移到无种子区域，从而实现目标地址的生成. 先前的研究，如AddrMiner-N[18]，尝试通过观察组织内部地址配置的相似性来识别不同的地址配置方式，主要依赖于组织名称的匹配. 然而，这种基于组织名称的匹配方法并不能充分揭示BGP前缀地址配置模式之间的内在关系. 此外，它仅能在相同组织内实现模式迁移，从而限制了地址探测的范围.

本文提出了6EDL-N算法，一种基于深度学习的方法，用于挖掘BGP前缀信息与地址配置方式之间的关联关系. 6EDL-N模型能够为无种子BGP准确匹配最相似的有种子BGP前缀，从而实现在无种子BGP前缀区域的活跃地址探测，且不受组织相同的限制，可以适用于任何无种子区域. 6EDL-N的核心思想在于挖掘BGP前缀信息和地址配置模式之间的关系，以定量方式描述BGP前缀之间的相似性. 图3展示了6EDL-N的模型结构，主要包括特征提取、相似度学习、目标地址生成和预探测机制模块，各模块的作用如下：

(1)特征提取：该模块旨在从Whois信息中提取BGP前缀特征，以便后续的相似度学习；

(2)相似度学习：该模块利用深度学习模型来量化前缀之间的相似程度，以确定它们之间的关联性；

(3)目标地址生成：基于相似度学习的结果，生成目标地址，以实现从有种子BGP前缀到无种子BGP前缀的地址迁移和候选地址生成；

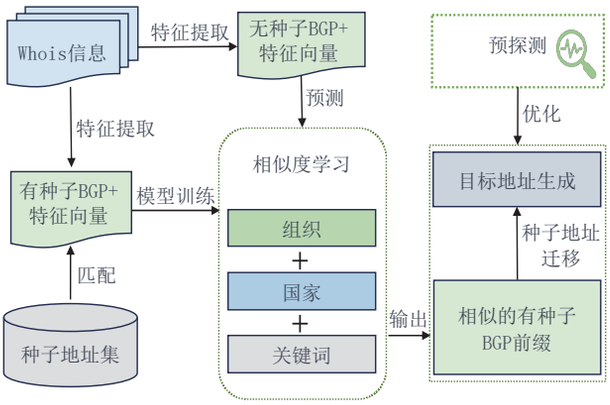

图3 6EDL-N模型结构

(4)预探测机制:对每一个无种子BGP前缀,预先分配少量资源进行探测,避免探测方向出现偏差.

6EDL-N的具体工作流程如下:首先收集了BGP前缀的Whois信息,并划分有种子地址的BGP前缀和无种子地址的BGP前缀.接着,从Whois信息中提取了组织、国家和描述信息,利用TF-IDF(TermFrequency-Inverse Document Frequency,词频-逆向文件频率)关键词提取方法[42]得到关键词.这些信息用于构建相似度向量,从而为相似度计算提供训练数据.随后,通过深度学习方法挖掘有种子BGP前缀信息和种子地址配置模式之间的关系.最后,对无种子BGP前缀使用训练好的模型进行相似度预测,选择相似度高的有种子BGP前缀,进行目标地址生成.下面将对四个功能模块进行详细介绍.

5.1.1 特征提取

IP的Whois信息包括了关于IP地址的所有者、注册商、注册时间、到期时间等关键信息,这些信息由互联网注册管理机构(Regional Internet Registry,RIR)提供,并且可以通过免费查询获得.通过获取BGP前缀的Whois信息,可以从中提取BGP前缀的特征.

首先进行了BGP前缀的Whois信息收集,这些信息包括了地址所属的组织(org字段)、国家(country字段),以及详细的描述信息(ipdescr字段)等.然而,目前尚未存在一个完整的、开源的Whois数据库可供使用,通常需要通过查询不同的RIR来获取这些信息.因此,本文基于全球五大RIR机构进行了信息查询与收集,并构建了离线的BGP前缀Whois信息库.可以通过查询信息库,来获取BGP前缀的相关信息.

在特征提取阶段,本文根据Whois信息所包含的字段信息以及这些信息反映的BGP前缀特性,对每个BGP前缀进行了三类特征的提取,分别为BGP前缀所属组织、国家,以及描述BGP前缀详细信息的关键词.对于BGP前缀的组织特征和所属国家特征提取,可以直接从Whois信息的org字段和country字段中提取这些信息.在关键词提取过程中,需要从Whois信息的描述字段中提取关键词.由于描述信息通常较为简短,而且其中的词语往往是专有名词,前后关联性较弱,本文选择了TF-IDF方法来进行关键词的提取.TF-IDF方法通常用于评估一个词语在文档中的重要性,其重要性与词语在文档中的频率成正比,但与其在整个语料库中的频率成反比.TF-IDF算法能够较好地衡量关键词在文档中的独特性,既考虑了词频因素,同时也考虑了关键词在整个语料库中的独特性,因此适用于处理简短的描述信息.关键词权重计算方法:

$$IDF_x=\log\left(\frac{N}{df_x}\right) \tag{1}$$

$$w_{x,y}=TF_{x,y}\times IDF_x \tag{2}$$

其中$w_{x,y}$为单词x在文档y中占有的权重,$TF_{x,y}$为单词x在文档y中出现的频率,IDF_x为单词x的逆文档频率.df_x为包含单词x的文档数量.N为文档总数.

关键词的提取流程如下:首先,需要对所有Whois信息中的描述信息进行规范化处理,包括去除标点符号并将所有单词转化为小写.然后,去除停用词,采用了一个包含常见的153个停用词的停用词表.接下来,使用了sklearn内置的TF-IDF库函数,将单词的最低出现频率设置为3.在对BGP前缀的Whois信息库初步筛选后,构建了一个4311×1073的词条-文档矩阵.由于每个句子的长度通常较短,因此这个矩阵非常稀疏.为了提取关键词,需要设定了一个阈值t,只有当$w_{x,y}$大于阈值t时,才会被初步选择.这一策略不仅有助于筛选出具有较高重要性的词语作为关键词,同时还能够降低后续模型训练过程中的时间消耗.为了在降低算法复杂度的同时尽可能多地选取关键词,以确保每个描述信息中的重要词语尽可能包含在最终建立的关键词表中(保证后续模型训练的覆盖前缀范围),本文定义了句子覆盖率($seq_{coverrate}$)、词语覆盖率($voc_{coverrate}$)来进行阈值的选择.

$$\mathrm{seq}_{\mathrm{coverrate}}=\frac{hit_{seq}}{M} \tag{3}$$

$$\mathrm{voc}_{\mathrm{coverrate}}=\frac{hit_{voc}}{N} \tag{4}$$

其中hit_{seq}为选定词表后，包含关键词的句子数量，M为句子总数．hit_{voc}为选定词表后，整个文档中包含关键词的个数，N为初步删选后所包含的单词数目．本文通过离线BGP前缀Whois信息库获取的句子总数和单词数量分别为$M=4311$和$N=1073$.

为了在关键词覆盖的前缀范围和算法时间复杂度之间找到平衡，图4中绘制了阈值t与句子覆盖率和词语覆盖率的关系．当阈值t为0.5时，句子覆盖率达到了88.0%，词语覆盖率为83.4%．在这种情况下，关键词覆盖了绝大多数句子，确保了后续模型的覆盖范围，同时关键词的数量尽可能减少，从而降低了算法的时间复杂度．因此，选择了阈值$t=0.5$，并最终从1073个初步筛选的候选词中选出了895个关键词来构建关键词表．

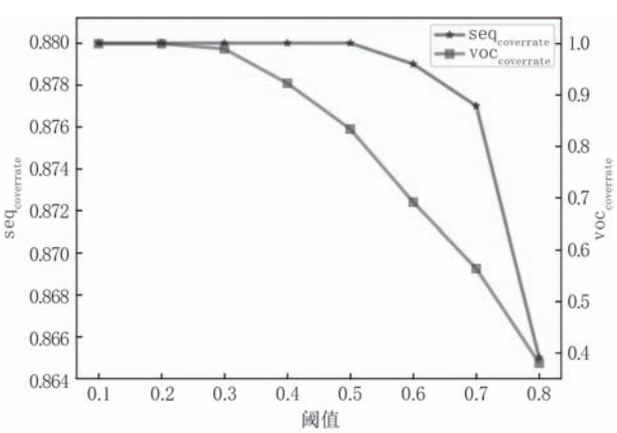

图4　句子覆盖率、词语覆盖率与阈值关系

5.1.2　相似度学习

在提取了Whois信息中的组织、国家和关键词等特征之后，需要深入挖掘BGP前缀信息与地址配置模式之间的内在联系．利用深度学习方法可以有效地揭示这两个因素之间潜在的关联．因此，需要构建适当的特征向量，并使用神经网络来计算两个BGP前缀的相似度，对于输入的特征向量，需要衡量组织、国家和关键词这三个特征之间的相似性．本文首先对特征进行独热编码，并利用同或运算构建了相似度向量．相似度向量由组织相似向量、国家相似向量和关键词相似向量组成，以衡量它们之间的相似性．本文定义了相似度分数作为输出向量来度量两个BGP地址配置模式之间的相似程度．特征向量的构建如下：

(1)组织相似向量

$$org_{or}(i,j)=org_i\odot org_j(i,j\in N) \tag{5}$$

其中org_i为组织向量，使用独热编码，维度为$k+1$，k为训练集中组织的总个数，第$k+1$位为保留位，当需要计算无种子BGP前缀和有种子BGP前缀的相似度时，如果输入的无种子BGP的组织没有在前k个组织中，则将第$k+1$位的值设置为1.

(2)国家相似向量

$$country_{or}(i,j)=country_i\odot country_j(i,j\in N) \tag{6}$$

其中$country_i$为国家向量，使用独热编码，与组织向量一样，维度为$m+1$，m为训练集中国家的总个数，第$m+1$位为保留位．如果输入的无种子BGP的国家没有在前m个，则将第$m+1$位的值设置为1.

(3)关键词相似向量

$$word_{or}(i,j)=word_i\odot word_j(i,j\in N) \tag{7}$$

其中$word_i$为关键词向量，使用独热编码，维度为n维，n为关键词表的长度．

根据上述三个相似向量的定义，输入的相似度向量的特征向量可以表示为：

$$X_{or}=[org_{or},\ country_{or},\ word_{or}] \tag{8}$$

(4)BGP配置规律的相似程度：

首先，定义地址表示方法如下：

$$a=a_1a_2a_3\cdots a_{30}a_{31}a_{32}a_i\in[0x0,0xf] \tag{9}$$

区域A内两个地址a_j,a_k间的距离计算公式：

$$S(a_j,a_k)=\sum_{i=0}^{32}\overline{(a_{ji}==a_{ki})}\{a_j,a_k\in A\} \tag{10}$$

那么，区域A,B内地址的平均距离：

$$S(A,B)=\frac{1}{|A||B|}\sum_{j=1}^{|A|}\sum_{k=1}^{|B|}S(a_j,b_k)\{a_j\in A,b_k\in B\} \tag{11}$$

最终得到区域A,B的相似度分数：

$$score=1-\frac{S(A,B)}{32} \tag{12}$$

相似度向量在表示两个BGP前缀的地址配置模式之间的相似程度时是可解释的．如果相似度向量全为0，这表示两个BGP前缀的输入信息之间没有相似性．由于一个BGP前缀只有一个组织信息和一个国家信息，因此组织相似向量和国家相似向量最多只包含一个1，其他位都是0．而关键词相似向量可能包含多个1．对于输出的相似度分数，如果两个BGP前缀之间的平均距离越大，表示它们的地址配置模式差异越大，因此计算出的相似度分数会越小，表示它们之间的相似度越低．为了便于相似

度的衡量，上述计算方法中相似度分数的值已经被归一化至0到1之间.

在构建完输入和输出的特征向量后，需要对神经网络进行训练. 本文选取了有种子BGP前缀来构建训练集，并从这些BGP前缀的Whois信息中提取了组织、国家和关键词的特征，并构建相似度向量来表示BGP前缀信息的相似性，并使用有种子BGP下的种子地址计算了两两BGP之间的相似度分数，以表示两个BGP地址配置模式之间的相似性. 神经网络的结构如下：输入层为相似度向量，即三个特征向量的拼接，接着连接了两个隐藏层，最后输出相似度分数. 这些隐藏层的设计旨在挖掘BGP前缀信息与地址配置模式之间的相似性. 图5为相似度学习的示意图.

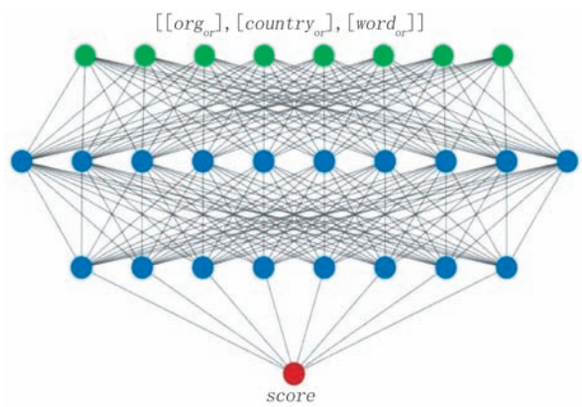

图5 基于神经网络的相似度学习

5.1.3 目标地址生成

在进行相似度学习后，神经网络挖掘出BGP前缀信息与地址配置模式之间的关联关系. 对于无种子BGP前缀，可以匹配出与之相似度分数高的有种子BGP前缀，由于这些有种子BGP前缀与该无种子BGP前缀的地址配置模式存在相似性. 可以进行种子地址的迁移，实现细颗粒度的候选地址生成. 迁移的方法如下：将相似度分数高的有种子BGP前缀下的种子地址的前缀替换为无种子BGP前缀. 首先确定无种子BGP前缀的长度（假设为n），然后将种子地址的前n位替换为该无种子BGP前缀. 如图6所示，使用BGP前缀2001:1410::/32来对地址2001:1400:ffee:1::3进行前缀替换，对于BGP前缀2001:1410::/32，其前32个半字节是2001:1410，因此，需要将地址2001:1400:ffee:1::3的前32个半字节替换为2001:1410. 替换后的结果是2001:1410:ffee:1::3，获得了具有相似配置模式的地址，为无种子BGP前缀生成更精确的候选地址.

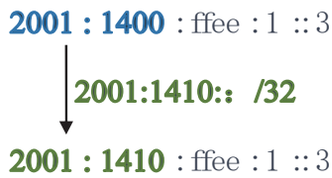

图6 替换举例示意图

5.1.4 预探测机制

本文引入了预探测机制，旨在避免有种子地址区域到无种子地址区域地址配置模式迁移失效的问题，以及减少探测资源的浪费，以提高探测效率.

6EDL-N的设计基于地址配置具有相似性这一核心思想，旨在实现从有种子地址区域到无种子地址区域的地址配置模式迁移和地址生成. 然而，种子地址抽样偏差可能导致BGP前缀和地址配置模式之间的关联性出现偏差. 因为有种子地址BGP前缀下的种子地址可能存在抽样偏差，所以该前缀下的种子地址的配置模式不一定反映真实的地址配置方式. 因此，在后续的迁移过程中，可能会导致迁移的地址模式与被探测前缀的模式不一致，从而降低探测效率.

为了规避可能存在的风险，6EDL-N引入了预探测机制. 对于每个无种子BGP前缀，系统首先生成满足预算的目标地址，然后使用部分目标地址，即占预算比例为p进行预探测. 如果预探测的命中率小于阈值ht，这意味着选取的地址模型不适合被探测的无种子BGP前缀，不再对这个BGP前缀下的其他目标地址进行探测. 这种预探测方法通过避免无效的迁移更合理地利用探测资源，提高了活跃IPv6地址探测的命中率. 且如果该无种子BGP探测到的所有活跃地址与候选地址的比值大于ct，则认为该无种子BGP前缀为别名前缀.

6 有种子地址探测模型6EDL-S

当前基于种子地址的IPv6活跃地址探测方法主要追求提高命中率，而忽视了活跃地址的发现速度，因此这些方法虽然保持较高活跃地址命中率，但这往往以牺牲地址发现速度为代价. 此外，由于种子地址抽样存在偏差问题，导致了地址探测命中率的下降. 为了解决上述问题，本研究提出了一种基

于深度学习的动态学习方法6EDL-S. 这种方法具有参数数量少、学习速度快的优势，通过引入了动态反馈机制以减轻种子地址抽样偏差问题，从而在提高活跃地址检测的命中率的同时，兼顾了高效的地址发现速度. 本节将详细介绍6EDL-S的设计.

6.1　6EDL-S模型

先前基于种子地址的探测方法证明将地址进行细颗粒度的分类能够提高地址的探测效率[35]. 本文提出基于深度学习的动态学习方法6EDL-S细粒度学习种子地址分布特征，并根据学习的地址分布特性生成进行活跃IPv6地址探测. 6EDL-S方法主要包含四个模块：种子地址聚类、地址特征学习、动态反馈机制、别名前缀检测. 各模块的作用如下：

(1)种子地址聚类：将粗粒度的种子地址集合划分更细粒度的地址簇，为后续地址学习提供数据输入；

(2)地址特征学习：学习每一个种子地址簇中种子地址的分布规律，生成目标地址；

(3)动态反馈学习：优化目标地址生成器的参数，动态引导目标地址生成方向；

(4)别名前缀检测：检测别名前缀，去除别名地址，消除别名前缀对活跃地址探测的影响.

6EDL-S整体探测流程如图7所示，伪代码见算法1和算法2. 首先，对输入的种子地址进行聚类，获取细粒度的种子地址簇. 紧接着，使用生成器和判别器对每个种子地址簇进行对抗学习，学习其地址分布规律，生成目标地址. 紧接着在环境中验证目标地址是否活跃，并进行别名前缀检测，消除别名地址，获取去除别名地址的活跃IPv6地址. 根据环境中反馈探测结果优化生成器参数，引导地址特征学习模块学习真实网络中活跃地址分布规律，进一步生成更可能存活的目标地址. 6EDL-S核心包括两个关键设计：一是简化深度学习模型. 6EDL-S的生成器和判别器使用简单的全连接层训练生成器和判别器，在保证训练效果的同时，具备可训练的参数少、训练速度快的特点，学习过程的时间复杂度低. 同时，生成器快速的目标地址生成能力可以在短时间内完成目标地址的生成. 二是动态反馈机制. 6EDL-S根据反馈机制优化模型参数，随着迭代次数的增加引导生成更可能存活的目标地址. 6EDL-S的两个关键设计在提升了活跃IPv6地址探测效率的同时，提高了活跃地址的发现速度，使其适合大规模活跃IPv6地址探测. 下面对6EDL-S方法中的四个核心模块展开介绍.

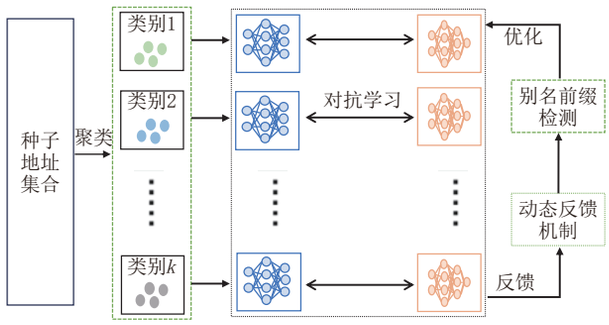

图7　6EDL-S模型结构图

算法1.　6EDL-S.

输入：Seeds(S)

输出：Candidate addresses(CA)

```
FUNCTION 6EDL-S(S)：
  CA=set()
loss_fn=BCEloss()#种子地址划分为k类
  FOR i in range(k)：#初始化生成器和判别器
      FOR train_epoches DO：
multiGANs (Si)
      END FOR
      FOR env_epoches DO：
          C=Gi(random_noise)
 Active_C=ZMapv6(C)
 multiGANs(Active_C)
      END FOR
 CA. add(Gi(random_noise))
 END FOR
 END FUNCTION
```

算法2.　MultiGANs.

```
FUNCTIONMultiGANs(S)：
real_output = Di (S)
Di_real_loss=loss_fn(real_output，torch. ones_like
(real_output))#判别器在真实数据上的损失
gen_seeds = Gi(random_noise)
fake_output = Di (gen_seeds)
Di_fake_loss=loss_fn(fake_output，torch. zeros_like
(fake_output))
#判别器在生成数据上的损失
Di_loss = Di_real_loss + Di_fake_loss
  Di _loss反向传播
Di参数更新
 fake_output = Di(gen_seeds)
 Gi_loss = loss_fn(fake_output，torch. ones_like
(fake_output))
  Gi _loss反向传播
 Gi参数更新
ENDFUNCTION
```

6.1.1 种子地址聚类

对于输入的种子地址集合，由于其广泛的地址分布范围和复杂多变的分布规律，而且在不同地区存在显著的差异．为了更好地发现地址分布规律，首要任务是对这些种子地址进行聚类，以使生成器能够更有效地学习它们的分布规律．在地址聚类过程中，本文使用Gasser等人提出的熵聚类[11]方法．熵聚类方法采用了信息论中的熵值概念，用以表示地址特定维度的变化情况．具体来说，对于地址中的每个半字节(4个比特位)，第j位的熵值$H(X_j)$的计算方法如下[11]：

$$H(X_j)=-\frac{1}{4}\sum_{\omega\in\Omega}P(X_j=\omega)\cdot logP(X_j=\omega)$$
$$\Omega=\{0,1,2,\cdots,f\} \tag{13}$$

其中，X_j表示第j位半字节的取值，熵值表示地址对应半字节的取值变化．$H(X_j)=0$表示对应第j个半字节的取值为一个固定值，相反的，$H(X_j)=1$表示在第j个半字节的取值在$[0-f]$上是机会是均等的．

已知特定地址半字节的熵值的计算方法，那么一个种子集合的香农熵特性向量的定义如下[11]：

$$F_b^a=\left(H(X_a),\cdots,H(X_j),\cdots,H(X_b)\right) \tag{14}$$

其中，a,b分别表示香农熵特性向量的初始和终止计算位置．

在聚类过程中，本文使用k-means聚类[43]方法对种子地址集合的香农熵特性向量进行处理，以获得更精细的种子地址簇．图8示例为聚类后的香农熵特性向量热力图．在这个热力图中，每一行代表了经过聚类后的一类地址集合的香农熵特性向量，而每个小方块的颜色则表示了该位置上香农熵的数值大小．

本文采用了熵聚类方法对种子地址集合进行分组，同时划分的类别数可以根据需要进行选择．具体地，实验中设置参数a为0，参数b为32，并参考以往的研究经验，选择了$k=6$来表示聚类的数量．

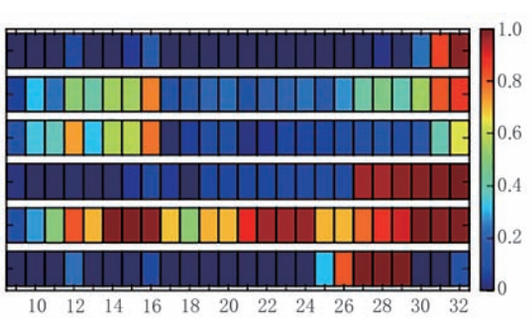

图8 香农熵特征向量热力图[11]($a=9,b=32$)

6.1.2 地址特征学习

为了学习种子地址的配置规律，本文采用了生成对抗网络(GAN)进行地址特征学习．在对种子地址进行聚类后，将它们细分为地址簇，然后基于GAN对每个类别中地址的分布特征进行详细学习．

在这个过程中，生成器和判别器作为博弈的双方．每个种子地址可以被看作一个32维的输入向量，其中每个维度相对独立．生成器的任务是学习这些种子地址，以拟合它们的分布，使生成的伪造地址尽可能地欺骗判别器．判别器的目标是区分真实的种子地址和来自生成器的伪造地址．这两个网络不断对抗，最终达到一种平衡状态，其中生成器充分拟合了训练种子地址的分布，判别器无法准确识别伪造地址．此时，生成器能够在短时间内生成大量候选地址．生成器和判别器使用全连接网络，以便更好地探索每个半字节的可能性．这一模型具有强大的探索能力．下面是对抗结构的示意图，如图9所示．以下是两个核心组件生成器和判别器的具体构建信息．

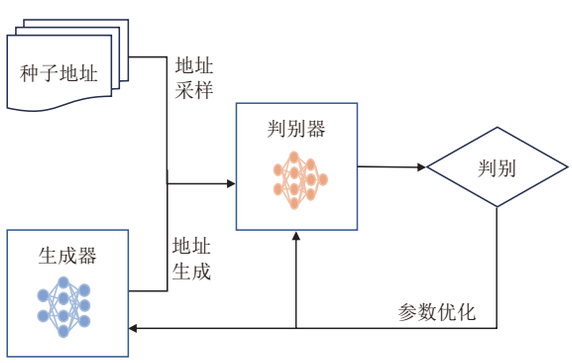

图9 模型对抗学习过程

生成器：输入为噪声数据z，经过全连接层，输出为生成数据$G(z)$，为32个半字节的地址．

生成器和判别器的优化目标函数如下：

$$\min_G\max_D V(D,G)=E_{x\sim p_{data}(x)}\left[logD(x)\right]+$$
$$E_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right)\right] \tag{15}$$

其中x表示种子地址，z表示输入的噪声，$G(z)$表示生成器生成的假地址，$D(*)$表示判别器判断是否为真实地址的概率．

6.1.3 动态反馈机制

尽管模型学习了种子地址特征的分布规律，但由于IPv6地址空间巨大的特性，无法全面采集种子

地址，因此实际采样是有偏的．这导致了采样的种子地址和真实网络中存在差异．采样数据的分布不能够完全反映真实活跃地址的分布，这可能会误导生成器，使其无法准确学习真实活跃地址的分布情况．为了生成更可能存活的候选地址，提高地址探测资源的有效利用率，本文引入了动态反馈机制．动态反馈机制的过程如图10所示．首先，进行生成器和判别器的对抗学习训练，直到其训练稳定．在这个阶段，生成器学习到了种子地址特征的分布情况．然后，对生成器的参数进行修正，以使其尽可能地学习真实活跃地址的情况．修正方法如下：在每轮训练中，使用生成器生成候选地址，通过ZMap进行活跃性探测，然后将探测到的活跃地址反馈给生成器．同时，进行别名地址的去除，以进一步提高生成器的学习效果．生成器再对这些活跃地址进行对抗学习，不断地优化其参数．多次的反馈过程有助于降低偏差，引导生成器朝向更可能存活的地址方向生成候选地址．动态反馈机制的引入能够更好地逼近真实活跃地址的分布，从而提高生成的候选地址的准确性和存活性．

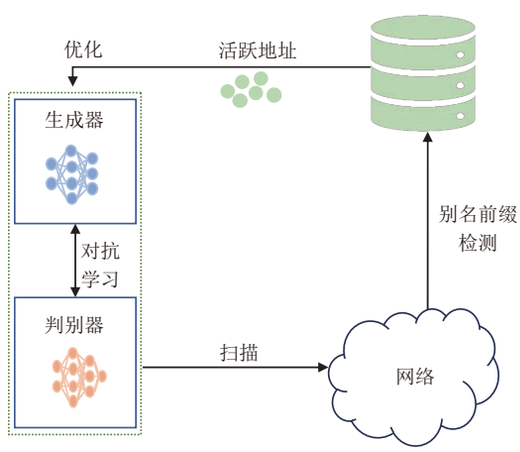

图10　动态反馈机制示意图

6.1.4　别名前缀检测

在IPv6的部署和推广过程中，出现了一种新现象，即别名前缀．这是由于IPv6地址空间极端庞大，网络管理员在配置地址时，通常会将整个地址前缀分配给一个网络主机，这个前缀被称为别名前缀，而其中的地址则被称为别名地址．在进行活跃IPv6地址探测时，经常会遇到一个配置了别名前缀的网络主机，它会消耗大量的探测资源．因此，在全球活跃IPv6地址探测中，别名前缀的检测和移除成为不可或缺的步骤．

在对抗学习的过程中，生成器可以根据输入的活跃地址学习到其分布情况．通过对生成器的输入地址进行别名前缀检测，去除包含别名前缀的活跃地址，从而使生成器仅学习不包含别名地址的种子地址．生成器学习种子地址的来源包括初始的种子地址和从环境中新发现的活跃地址．在别名前缀检测模块分别对两个不同来源的地址进行别名前缀检测和别名地址去除．无论是初始的种子地址还是反馈机制中新发现的活跃地址，本文采用Gasser等人[11]提出的MAPD别名前缀检测方法①．

具体地，如图11所示，对于被探测前缀，遍历其子前缀伪随机生成的16个候选地址．例如，为了检测2001:de8:abc::/48是否为别名前缀，需要对每个4位子前缀空间内生成一个伪随机地址，即2001:de8:abc:[0-f]000::/52．然后，针对每个地址发送一个ICMPv6探测包，如果收到响应的地址数量大于等于15，则认为被探测前缀是别名前缀，该前缀下的地址是别名地址．

```
2001:0de8:0abc:0a32:4fcb:8ca8:7c64:ba26
2001:0de8:0abc:1b48:5f81:3453:8268:6bd2
2001:0de8:0abc:2e91:2900:77e9:03a8:875
...
2001:0de8:0abc:fd37:2443:915e:1d2e:53b2
```

图11　2001:de8:abc::/48伪随机生成地址

7　实验评估

7.1　实验配置

实验在Ubuntu 18.04系统、Intel Core Processor (2 GHz)、125 GB内存的机器上运行．Python版本为3.8.8，Pytorch使用cpu 1.11.0版本．

7.1.1　实验指标

为了衡量种子地址探测模型的性能，本文构建了一套评价体系全面地衡量算法进行大规模种子地址探测的性能，该评价体系由三个指标构成．对于一个地址活跃性的判断，对其发送ICMPv6数据包，

① 基于分片指纹的别名前缀检测方法FAPD[5]需要重复探测数据，探测效率较慢，加之不能适用所有的前缀检测，所以本文使用MAPD进行大规模别名前缀检测．

若收到该地址回复的响应报文，则认为该地址是活跃的.

(1)命中率(*HitRate*)：假设候选地址的集合为C，其使用的种子地址集合为S，命中的活跃地址集合为A，则命中率计算方法如下：

$$HitRate=\frac{|A-A\cap S|}{|C|} \tag{16}$$

(2)单位时间发现的活跃地址数(*NPT*)：假设候选地址的集合为C，其使用的种子地址集合为S，命中地址的集合为A，生成和探测候选地址所需的时间为t，则NPT的计算方法如下：

$$NPT=\frac{|A-A\cap S|}{t} \tag{17}$$

(3)覆盖度(*Coverage*)：假设无种子BGP前缀数量为M，在探测后有N个无种子BGP前缀搜集到种子地址，则覆盖度的计算方法为：

$$Coverage=\frac{N}{M} \tag{18}$$

其中命中率(*HitRate*)、单位时间发现的活跃地址数(*NPT*)分别用来衡量种子地址探测的准确性与效率，覆盖度(*Coverage*)用于衡量无种子地址探测模型的探索范围. 需要说明的是，由于基于深度学习的模型在模型训练完成后不需要再次进行训练即可进行大规模的地址生成，6EDL的每次进行大规模探测时发现的活跃地址具有多样性，而现有算法对于相同的种子地址输入，输出的候选地址均相同. 因此，在计算NPT时，其时间t只需考虑候选地址的生成和探测时间.

7.1.2 数据集

Pyasn是由代尔夫特理工大学网络安全经济学研究小组[①]的研究人员开发和维护的开源库. 作为Python的扩展模块，Pyasn能够实现IP地址到自治系统(AS)号的快速查询. 本文使用了这一扩展模块来获取了截至2022年11月13日的Pyasn库，其中包含了142 649条IPv6 BGP前缀列表. 此外，本文获取了Gasser公开的不包含别名地址的IPv6 hitlist数据集[11]，其中包含653.9万个活跃IPv6地址，并使用Pyasn库进行查询，其所覆盖的BGP前缀情况如表1所示.

本文从全球五大RIR机构获取了上述BGP前缀的Whois信息. 这五大RIR机构分别为美洲互联网号码注册管理机构(American Registry for Internet Numbers，ARIN)[②]、欧洲IP网络资源协调中心(RIPE Network Coordination Centre，RIPE NCC)[③]、亚太网络信息中心(Asia-Pacific Network Information Centre，APNIC)[④]、拉丁美洲及加勒比地区互联网地址注册管理机构(Latin American and Caribbean Internet Address Registry，LACNIC)[⑤]以及非洲网络信息中心(African Network Information Centre，AfriNIC)[⑥]，统计结果见表2.

表1 IPv6 hitlist覆盖BGP前缀情况统计

数据来源	活跃地址数	覆盖前缀	未覆盖前缀
IPv6 hitlist	6539 000	47 761	94 888

表2 BGP前缀的Whois信息获取来源统计

来源	ARIN	RIPE NCC	APNIC	LACNIC	AfriNIC
数量	26 102	45 017	62 412	7671	1447

7.1.3 参数设置

(1)6EDL-N的参数设置

在所有的BGP前缀中，统计出Whois信息中包含8631个组织，225个国家和895个关键词. 在相似度学习阶段，本文构建了一个五层神经网络，其中输入层维度为9753，三个隐藏层的隐藏单元个数分别为256、512和256，最终输出维度为1，用于表示相似度分数. 前四层均采用了relu激活函数，而最后一层则采用了sigmoid激活函数. 优化器选择Adam，学习速率为0.0001，batchsize选取了5000，并且设定了50轮的训练次数.

(2)6EDL-S的参数设置

6EDL-S的生成器采用四层全连接层，根据超参数选择的经验，设置输入维度为100的随机噪声，连接隐藏单元数为128和64的隐藏层，由于种子地址使用熵值聚类方法分类后每一个前导的长度均为8，故为了防止存在生成不合法的前导，将其前8位固定，因此设置输出维度为24. 生成器的每一层均使用Relu激活，并将输出控制在0-15的浮点数，判别器前两层使用LeakyReLU激活，最后一层使用Sigmoid激活函数进行二分类. 损失函数使用交叉

① Economics of Cybersecurity. https://ocw.tudelft.nl/courses/economics-of-cybersecurity/. 2023,10,20.

② ARIN. WHOIS-RWS. https://whois.arin.net/ui/. 2023,10,20.

③ RIPE NCC. RIPE Network Coordination Centr. https://www.ripe.net/. 2023,10,20.

④ APNIC. APNIC Whois Search. https://wq.apnic.net/static/search.html. 2023,10,20.

⑤ LACNIC. MiLACNICQuery. https://query.milacnic.lacnic.net/home. 2023,10,20.

⑥ AfriNIC. African Network Information Centre. https://www.afrinic.net/whois. 2023,10,20.

熵损失，生成器和判别器的优化器均选取Adam，学习率为0.0001，batchsize选取64. 由于训练过程中生成器输出的是0-15之间的浮点数，而在IPv6地址的每一位半字节均为整数，这里采用向下取整. 初始的生成器和判别器对抗轮数为1200轮，在每次从环境中学习的过程中，每轮生成器生成2万个候选地址，使用ZMap[3]探测其活跃性，将活跃地址反馈回生成器，并进行100轮的训练.

7.2　6EDL-N评估

在进行相似度学习时，如果将每两个有种子BGP都作为输入，会面临两个问题：一是训练时间和存储成本高，二是可能存在相同的输入却得到不同输出的情况. 因此，将组织和国家构造成元组，共构建了9059对不同的组合. 对于每一对组合，仅选取一个相似度分数进行训练，从而极大地减少了计算量和存储量. 完成BGP前缀相似度计算模型的训练后，对于输入的无种子BGP前缀，与每一个有种子BGP前缀构建相似度向量，然后将其输入神经网络进行预测，通过对相似度分数进行排序，可以得到与输入BGP前缀相似度最高的有种子BGP前缀. 然后选取相似度最高的m个BGP前缀（对于每个无种子BGP前缀，m的值可能不同），以确保其包含num个种子地址生成的目标地址（num为每个无种子BGP前缀空间地址生成的预算）. 在实验中，选取num为1千、1万、10万和100万进行探测. 由于无种子BGP的地址配置模式未知，候选地址的增多可能导致地址生成方向的偏差加大，存在探测资源浪费的问题. 因此，本文采取预探测的方法来提高其准确性. 在预探测的参数选取中，选取p为10%，ht选取5%，ct选取90%. 具体而言，这表示将预算的10%用于预探测，若预探测的命中率低于5%，将不对这个BGP前缀再进行探测. 另外，如果一个BGP前缀探测到的活跃地址数量超过预算的90%，则将认定这个BGP前缀为别名前缀. AddrMiner-N在无种子地址探测方面取得了最优的结果，因此，本文使用6EDL-N与其进行比较，通过实验验证6EDL-N在无种子地址探测的优势.

7.2.1　6EDL-N与AddrMiner-N比较

6EDL-N的核心目标是实现广泛的无种子地址区域探测，同时保持高准确度和高效率. 因此，本文使用命中率（*HitRate*）、地址发现速度（*NPT*）和地址覆盖度（*Coverage*）指标来评估. 覆盖度越高表示构建的活跃IPv6地址越全面，能更好地反映活跃IPv6地址的全球部署情况. 本文从上述三个指标出发，将6EDL-N与当前无种子地址生成算法AddrMiner-N进行比较，具体评估如下：

（1）命中率（*HitRate*）评估

图12展示了6EDL-N与AddrMiner-N的命中率比较情况. 可以观察到，在命中率方面，随着预算的增加，AddrMiner-N的命中率逐渐降低. 当预算为1000时，AddrMiner-N的命中率最高为1.76%；而在大规模探测时，其命中率降低至1.56%，说明其探测准确度有明显下降. 这可能是由于随着预算的增加，探测模式的空间也随之增大，从而增加了目标地址生成的随机性. 相比之下，6EDL-N采取了预探测机制，在预算为1万、10万和100万时，优先在活跃地址密度较大的区域进行探测，从而缓解了资源浪费的情况. 在预算为1千时，由于预算较小，为保证探测的覆盖度不受预探测影响，未采用预探测机制，导致命中率较低. 而当预算增加到100万时，尽管采取了预探测机制，但由于预算过大，活跃地址数量有限，探测资源浪费的问题难以缓解，导致命中率略有下降. 具体来说，当预算为1千时，6EDL-N的命中率为2.69%，是AddrMiner-N的1.53倍. 在进行大规模探测时，当预算为100万时，6EDL-N的命中率为12.69%，而AddrMiner-N的命中率为1.56%，6EDL-N的命中率是AddrMiner-N的8.13倍.

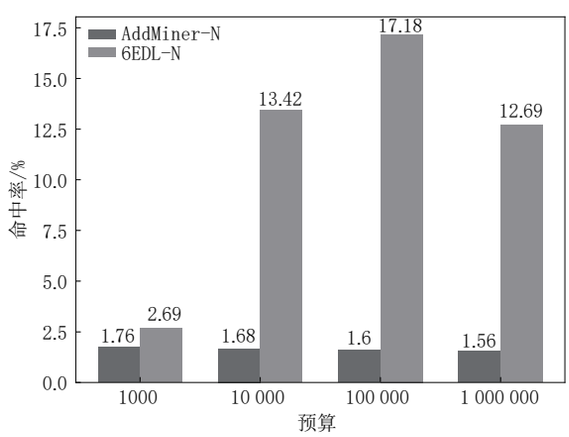

图12　命中率比较结果

（2）单位时间发现的活跃地址数（*NPT*）评估

*NPT*用来评估活跃IPv6地址探测方法发现活跃IPv6地址的效率. 算法的命中率越高，且活跃地址生成的时间越短，活跃地址探测的效率越高.

图13展示了6EDL-N与AddrMiner-N单位时间发现活跃地址数的比较情况. 随着预算的增加，两个模型的*NPT*均呈上升趋势. 这是由于算法的

运行和候选地址生成需要时间，而候选地址生成的数量和算法运行的时间并不是线性关系．当候选地址大规模增加时，地址生成的时间增加不明显，从而导致*NPT*随着预算的增加而上升．6EDL-N显示出快速发现大量活跃地址的能力，*NPT*随着预算的增加而明显提升．而且随着预算的增大，6EDL-N在*NPT*方面的优势也增加．在大规模探测时，当预算为100万时，6EDL-N的*NPT*为233.09个/s，而AddrMiner-N仅为15.6个/s，是其的14.94倍．因此，6EDL-N能够在无种子区域进行大规模高效的活跃IPv6地址探测．

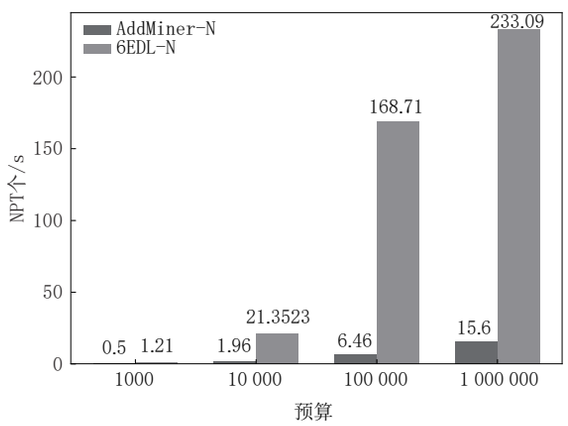

图13 NPT比较结果

(3)覆盖度(*Coverage*)评估

覆盖度用来表示发现的活跃IPv6地址覆盖BGP前缀数量占被探测BGP前缀数量的比值．覆盖度越大表示活跃IPv6地址探测方法适用区域越大，扩展性越强，也表示发现的活跃IPv6地址越全面．如表3所示，在覆盖度方面，随着预算的增加，6EDL-N在每个无种子BGP前缀下生成了更多的候选地址，更可能在无种子BGP下发现活跃的IPv6地址．因此，覆盖度随着预算的增加而增大．当预算为1千时，由于预算较小，覆盖度最低为17.00%．随着预算的增大，覆盖度逐步提升．当预算为100万时，覆盖度最高达到21.97%．此时，6EDL-N的覆盖度是AddrMiner-N的1.84倍．实验结果表明，6EDL-N在无种子区域IPv6活跃地址探测方面表现出覆盖范围广、高命中率和快速的特点．相对于现有的无种子区域地址探测方法AddrMiner-N，6EDL-N获得了显著的提升．

表3 覆盖度统计

预算	1千	1万	10万	100万
6EDL-N	17.00%	19.36%	21.58%	21.97%
AddrMiner-N	9.70%	10.22%	10.62%	11.91%

7.3 6EDL-S评估

本文采用6.1.1节中的种子地址聚类方法将IPv6 hitlist划分为6类，得到43个前导(前导是指：分类后地址的前8个半字节)，为了保持每类数据的均衡性，选取种子地址数量小于20万的前导，聚类后类别的种子地址数量统计如表4．现有的有种子地址探测算法较多，其中在大规模探测性能优异的算法有六个：Entropy/IP、6Tree、DET、6Graph、6Forest和AddrMiner-S，由于6Hit算法生成的地址空间过大，导致其运行时间复杂度过高，无法进行大规模实验探测，当前最高效的基于深度学习的地址探测方法6GAN的时间复杂度高，而IPv6 hitlist的种子地址规模大，在合理的时间内，6GAN仅能进行少量轮数的训练，导致学习效果不理想，因此本文未采用这两种算法进行比较．

表4 IPv6 hitlist分类结果表

类别	C0	C1	C2	C3	C4	C5
数量	72 005	171 795	58 303	128 267	10 990	170 593

7.3.1 6EDL-S与现有模型比较

本文分别计算了六种算法以及6EDL-S的命中率(*HitRate*)和单位时间内发现的活跃地址数(*NPT*)．每个算法生成的候选地址数均为611 953，生成的候选地址数量和种子地址数目一致．具体的统计结果如表5：

表5 6EDL-S与上述模型比较

算法	命中地址数	HitRate(%)	地址生成时间(s)	NPT
Entropy/IP	14 593	2.38	193.87	75.27
6Tree	112 120	18.32	565.86	198.14
DET	118 719	19.40	589.73	201.32
6Graph	55 162	9.01	673.19	81.94
6Forest	50 830	8.31	359.73	141.30
AddrMiner-S	128 495	21.00	409.25	313.98
6EDL-S	**158 527**	**25.91**	**339.66**	**466.72**

(1)命中率(*HitRate*)评估

根据表5的结果，可以明显看出6EDL-S在命中率和单位时间内发现活跃地址数两个指标上都取得了显著的提升．在命中率方面，各个算法的命中率从高到低的排列顺序分别为6EDL-S(25.9%)、AddrMiner-S(21.00%)、DET(19.40%)、6Tree

（18.32%）、6Graph（9.01%）、6Forest（8.31%）、Entropy/IP（2.38%）. 在上述的所有算法中，Entropy/IP的命中率最低，仅为2.38%，而AddrMiner-S的命中率最高，为21.00%. 然而，6EDL-S的命中率达到了25.91%，是现有方法的1.23～10.89倍.

（2）单位时间发现的活跃地址数（*NPT*）评估

单位时间发现的地址数（*NPT*）很好地反映了模型探测的效率，因为算法的优劣不仅应该考虑探测资源的利用效率，还应考虑时间成本. 由于各个算法对活跃地址的发现速度不同，因此*NPT*和命中率的排序并不相同. 在单位时间内发现活跃地址方面，从高到低依次是6EDL-S、AddrMiner-S、DET、6Tree、6Forest、6Graph、Entropy/IP. 其中，6Graph的地址生成时间最长，达到了673.19秒，而6EDL-S最快，仅为339.66秒. 虽然现有的活跃地址探测方法6Graph具有很高的命中率，但其内存消耗过大，地址生成时间最长，整体上探测效率较低. 相对的，虽然Entropy/IP的运行时间最快（算法本身并不快，解析源码发现是由于其限制了种子地址数量不能大于10万造成的），但其命中率仅为2.38%，低的命中率导致活跃地址发现速度慢. 相比现有算法，6EDL-S很好地权衡了命中率和探测速率的关系，其命中率为25.91%，而生成候选地址仅花费了339.66秒. 6EDL-S在单位时间发现的活跃地址数（*NPT*）为466.72个/s，是现有方法活跃地址发现速度的1.49～6.20倍.

实验结果表明，6EDL-S能够在有种子区域有效地进行大规模活跃地址探测. 相较于现有的有种子地址探测方法，它在命中率和单位时间发现活跃地址数方面都有显著的提升，很好地权衡了命中率和地址生成时间之间的关系. 此外，6EDL-S还引入了一种从环境中反馈学习的机制来学习真实的活跃地址分布（参考7.3.2节反馈机制验证）. 且在这个反馈过程中，生成器通过学习不包含别名前缀的数据，有效地解决了别名前缀问题.

7.3.2　反馈机制有效性验证

由于抽样偏差会影响活跃地址探测的命中率，因此在反馈过程中观察了命中率的变化情况，以评估反馈机制对抽样偏差的纠正效果. 根据实验设计，每隔20轮统计一次6EDL-S的命中率，并绘制了图14来展示结果. 从图中可知随着反馈轮数的增加，命中率不断提高. 具体地，在没有反馈的情况下（即反馈轮数为0时），命中率为14.95%；而当反馈轮数增加到100时，命中率达到了25.91%，是不加入反馈的1.73倍. 当不加入反馈机制时，即生成器不再次学习反馈得到的活跃地址，命中率在14.95%上下波动，当轮数为20轮时，命中率最低为14.15%，当轮数为80轮时，命中率最高为15.21%，这是由于不加入反馈学习的机制，生成器的参数没有优化，仍然学习的是初始有偏采样的种子地址，导致命中率在初始值附近波动. 而从环境中动态反馈的机制使得生成器在每轮反馈中学习到了更多真实活跃地址的分布情况，而不仅是初始的种子地址. 这使得生成器生成的候选地址逐渐符合真实活跃地址的分布，从而缓解了种子地址抽样偏差的问题. 理论上来说，随着反馈轮数的增加，当探测完整个IPv6地址空间时，生成器将能够学习到所有活跃地址的分布情况，从而完全解决了抽样偏差的问题. 因此，本文提出的反馈机制在一定程度上解决了种子地址抽样偏差的问题.

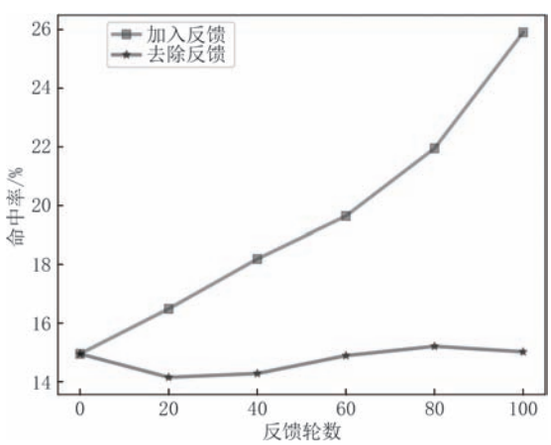

图14　反馈结果有效性验证

8　IPv6　hitlist分析

在第7节中，本文对6EDL的性能进行了评估. 本章利用第4节提出的探测架构，在真实生产网络环境下部署了6EDL系统对全球IPv6互联网进行了全面探测. 6EDL系统会将探测到的活跃IPv6地址添加到活跃IPv6地址列表中，形成IPv6 hitlist. 在本节中，将进一步对IPv6 hitlist进行分析并和Gasser[11]和Song①公开的IPv6 hitlist进行比较.

对构建的IPv6 hitlist，使用Pyasn库来查询活跃

① AddrMiner：A Comprehensive Global Active IPv6 Address Discovery System. . https://addrminer. github. io/IPv6_hitlist. github. io/. 2023，10，20

IPv6地址所属的BGP前缀和AS号.最终得到了IPv6 hitlist所覆盖的BGP前缀数量和AS数量.这些数据反映了地址探测系统的探测范围,即构建的IPv6 hitlist的覆盖广度.同时,分别统计了搜集到的所有活跃地址集合和去除别名后的活跃IPv6地址集合中的活跃地址数量,并列出了不同类型集合中排名前五位AS包含的活跃IPv6地址所占比例.具体的统计结果如表6所示.总共发现了29.77亿个活跃IPv6地址,其中包括了5.66亿个别名地址和24.11亿个非别名地址.此外,从统计活跃地址的分布情况来看,现有的活跃IPv6地址主要集中在大型互联网服务商(Akamai、Amazon、Cloudfare等),这也表明现有的IPv6部署核心以大型互联网厂商主导.由于活跃IPv6地址探测的目的在于发现网络空间中的网络设备.因此,即使一个目标前缀被判定为别名前缀(即该前缀被配置到同一台网络设备,该前缀下的所有地址都是活跃的),在去除别名地址时,需要在该别名前缀下至少保留一个活跃IPv6地址,以确保目标网络中的设备能够被探测到.因此,在表6中,无论是否去除别名地址,构建的IPv6 hitlist所覆盖的BGP前缀和AS数量是相同的.本文构建的IPv6 hitlist涵盖了125 101个BGP前缀和40 137个AS,占据了整体宣告前缀数量的87.67%(截至2022年11月).

使用本文构建的IPv6 hitlist与公开的两个IPv6 hitlist进行了比较,如表7所示,本文的IPv6 hitlist包含了更多的活跃地址数量,覆盖了更多的BGP数量、AS数量.这表明本文的IPv6 hitlist覆盖了大部分IPv6网络空间,更好地反映了活跃IPv6地址的部署情况.

表6 IPv6 hitlist数据统计

类别	活跃地址数量	BGP数量	AS数量	TOP1 AS	TOP2 AS	TOP3 AS	TOP4 AS	TOP5 AS
IPv6 hitlist	29.77亿	125 101	40 137	28.10%★	23.03%●	5.36%◇	4.41%◆	3.48%■
IPv6 *hitlist**	24.11亿	125 101	40 137	34.69%★	28.44%●	6.61%◇	5.44%◆	4.29%☆

注:IPv6 hitlist: 包含别名地址的活跃地址集合; IPv6 *hitlist**:去除别名地址的活跃地址集合★Akamai ● Amazon ◇ Bharti Airtel Ltd ◆ Cloudflare ■Rohmad Kumiadin ☆Google

表7 IPv6 hitlist数据比较

类别	活跃地址数量	BGP数量	AS数量
IPv6 *hitlist**	24.11亿	125 101	40 137
Gasser	0.079亿	44 286	16 347
Song	0.957亿	80 720	22 771

9 总 结

随着云物联网、5G技术的发展,互联网用户暴增,IPv4地址资源耗尽,IPv6在全球开始加速部署.IPv6部署虽然解决了IPv4地址耗尽的问题,但由于其巨大的地址空间,给IPv6活跃地址扫描带来了巨大的挑战.本文提出了高效、快速、适用范围广的活跃地址探测系统6EDL,有效解决了活跃IPv6地址探测速度慢、探测效率低、探测范围受限的问题.6EDL将地址探测分为无种子地址场景和有种子地址场景,并针对每种场景设计高效探测算法.在无种子地址场景下,本文提出了6EDL-N方法,通过使用神经网络挖掘BGP前缀信息与地址配置模式之间的潜在关系,实现了有种子区域到任意无种子区域的地址迁移.在有种子地址场景下,本文提出基于对抗网络的活跃地址探测方法6EDL-S.除此之外,本文还提出了一套更为完整的地址探测评价指标,并使用这些指标多角度衡量活跃地址探测的性能.最终,经过持续探测,累计发现了29.77亿个活跃地址,包含5.66亿别名地址和24.11亿非别名地址,覆盖了125 101个BGP前缀和40 137个AS.

参 考 文 献

[1] Beverly R, Durairajan R, Plonka D, et al. In the IP of the beholder: Strategies for active IPv6 topology discovery// Proceedings of the Internet Measurement Conference 2018. Boston, USA. 2018: 308-321

[2] Vermeulen K, Rohrer J P, Beverly R, et al. Diamond-miner: Comprehensive discovery of the Internet's topology diamonds// Proceedings of the 17th USENIX Symposium on Networked Systems Design and Implementation (Clara Santa, USA. 2020: 479-493

[3] Durumeric Z, Wustrow E, Halderman J A. ZMap: Fast internet-wide scanning and its security applications//Proceedings of the 22nd USENIX Security Symposium (USENIX Security 13. Washington, USA. 2013: 605-620

[4] Izhikevich L, Teixeira R, Durumeric Z. LZR: Identifying unexpected internet services//Proceedings of the 30th USENIX

Security Symposium (USENIX Security 21. Virtual, Online. 2021: 3111-3128

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

[6] Kensuke Fukuda and John Heidemann. Who knocks at the ipv6 door? detecting ipv6 scanning//Proceedings of the Internet Measurement Conference 2018, Boston, USA ,2018.231 - 237

[7] Moon Soo-Jin, Yin Yucheng, Rahul Anand Sharma, et al. Accurately measuring global risk of amplification attacks using ampmap//Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), Virtual, 2021,3881-3898

[8] Jing X, Yan Z, Jiang X, et al. Network traffic fusion and analysis against DDoS flooding attacks with a novel reversible sketch. Information Fusion, 2019, 51: 100-113

[9] Jing X, Yan Z, Pedrycz W. Security data collection and data analytics in the internet: A survey. IEEE Communications Surveys & Tutorials, 2018, 21(1): 586-618

[10] Song G, Yang J, He L, et al. AddrMiner: A comprehensive global active IPv6 address Discovery System//Proceedings of the 2022 USENIX Annual Technical Conference (USENIX ATC 22. Carlsbad, USA. 2022: 309-326

[11] Gasser O, Scheitle Q, Foremski P, et al. Clusters in the expanse: Understanding and unbiasing IPv6 hitlists// Proceedings of the Internet Measurement Conference 2018. Boston, USA. 2018: 364-378

[12] Jing X, Yan Z, Han H, et al. Extendedsketch: Fusing network traffic for super host identification with a memory efficient sketch. IEEE Transactions on Dependable and Secure Computing, 2021, 19(6): 3913-3924

[13] Amann J, Gasser O, Scheitle Q, et al. Mission accomplished?: HTTPS security after diginotar//Proceedings of the 2017 Internet Measurement Conference, London, UK, 2017: 325-340

[14] Gasser O, Hof B, Helm M, et al. In log we trust: Revealing poor security practices with certificate transparency logs and Internet measurements//Proceedings of the PAM 2018, Berlin, Germany, 2018: 173-185

[15] Liu R, Weng Z, Hao S, et al. Addressless: Enhancing IoT server security using IPv6. IEEE Access, 2020, 8: 90294-90315

[16] Modares H, Moravejosharieh A, Lloret J, et al. A survey of secure protocols in mobile IPv6. Journal of Network and Computer Applications, 2014, 39: 351-368

[17] Wang X, Mu Y. A secure IPv6 address configuration scheme for a MANET. Security and Communication Networks, 2013, 6(6): 777-789

[18] Song G, Yang J, He L, et al. AddrMiner: A comprehensive global active IPv6 address discovery system//Proceedings of the 2022 USENIX Annual Technical Conference (USENIX ATC 22. Carlsbad, USA, 2022: 309-326

[19] Fiebig T, Borgolte K, Hao S, et al. Something from nothing (there): collecting global IPv6 datasets from DNS// Proceedings of the Passive and active measurement: 18th International Conference, PAM 2017, Sydney, Australia, 2017, 2017: 30-43

[20] Fiebig T, Borgolte K, Hao S, et al. In rDNSwe trust: Revisiting a common data-source's reliability.//Proceedings of the Passive and active measurement: 19th International Conference, PAM 2018. Berlin, Germany, 2018: 131-145

[21] Borgolte K, Hao S, Fiebig T, et al. Enumerating active IPv6 hosts for large-scale security scans via DNSSEC-signed reverse zones//Proceedings of the 2018 IEEE Symposium on Security and Privacy, San Francisco, California, USA, 2018: 770-784

[22] Huz G, Bauer S, Claffy K C, et al. Experience in using mturk for network measurement//Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data. London, UK, 2015: 27-32

[23] lonka D, Berger A. Temporal and spatial classification of active IPv6 addresses//Proceedings of the 2015 Internet Measurement Conference. Tokyo, Japan. 2015: 509-522

[24] Gasser O, Scheitle Q, Gebhard S, et al. Scanning the IPv6 internet: towards a comprehensive hitlist. arXiv preprint arXiv: 1607.05179, 2016. https://arxiv. org/abs/1607.05179.2023, 10,20

[25] Rye E, Levin D. IPv6 hitlists at scale: Be careful what you wish for//Proceedings of the ACM SIGCOMM 2023 Conference. New York, USA. 2023: 904-916

[26] Li Guo, He Lin, Song Guang-Lei, Wang Zhi-Liang, et al. IPv6 active address discovery algorithm based on multi-level classification and space modeling. Journal of Tsinghua University (Science and Technology), 2021, 61(10): 1177-1185. (in Chinese)
(李果, 何林, 宋光磊, 王之梁, 杨家海, 林金磊, 高浩. 基于多层级分类和空间建模的IPv6活跃地址发现算法. 清华大学学报(自然科学版), 2021, 61(10): 1177-1185)

[27] Foremski P, Plonka D, Berger A W. Entropy/IP: Uncovering Structure in IPv6 Addresses//Proceedings of the 2016 ACM on Internet Measurement Conference, Santa Monica, USA, 2016: 167-181

[28] Ullrich J, Kieseberg P, Krombholz K, et al. On reconnaissance with IPv6: A pattern-based scanning approach//Proceedings of the 10th International Conference on Availability, Reliability and Security. Toulouse, France, 2015: 186-192

[29] Song G, Yang J, Wang Z, et al. Det: Enabling efficient probing of IPv6 active addresses. IEEE/ACM Transactions on Networking, 2022, 30(4): 1629-1643

[30] Murdock A, Li F, Bramsen P, et al. Target generation for internet-wide IPv6 scanning//Proceedings of the 2017 Internet Measurement Conference, London,UK. 2017: 242-253

[31] Liu Z, Xiong Y, Liu X, et al. 6Tree: Efficient dynamic discovery of active addresses in the IPv6 address space. Computer Networks, 2019, 155: 31-46

[32] Yang T, Cai Z, Hou B, et al. 6Forest: An ensemble learning-based approach to target generation for Internet-wide IPv6 scanning//Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications. 2022: 1679-1688

[33] Hou B, Cai Z, Wu K, et al. 6Hit: A reinforcement learning-based approach to target generation for Internet-wide IPv6 scanning//

Proceedings of the IEEE INFOCOM 2021-IEEE Conference on Computer Communications. Vancouver, Canada, 2021: 1-10

[34] Yang T, Hou B, Cai Z, et al. 6Graph: A graph-theoretic approach to address pattern mining for Internet-wide IPv6 scanning. Computer Networks, 2022, 203: 108666

[35] Cui T, Gou G, Xiong G, et al. 6GAN: IPv6 multi-pattern target generation via generative adversarial nets with reinforcement learning//Proceedings of the IEEE INFOCOM 2021-IEEE Conference on Computer Communications, Vancouver, Canada, 2021: 1-10

[36] Cui T, Gou G, Xiong G. 6gcvae: Gated convolutional variational autoencoder for ipv6 target generation//Proceedings of the Advances in Knowledge Discovery and Data Mining, Singapore, Singapore, 2020: 609-622

[37] Cui T, Xiong G, Gou G, et al. 6veclm: Language modeling in vector space for ipv6 target generation//Proceeding of the ECML PKDD 2020, Ghent, Belgium, 2020: 192-207

[38] Kingma D P, Welling M. Auto-encoding variational bayes. arXiv preprint , 2013: arXiv: 1312.6114

[39] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA. 2017: 6000-6010

[40] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 2018, 35(1): 53-65

[41] Graves A. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, Springer, 2012

[42] Aizawa A. An information-theoretic perspective of tf-idf measures. Information Processing & Management, 2003, 39(1): 45-65

[43] MacQueen J. Some methods for classification and analysis of multivariate observations//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967, 1(14): 281-297

SONG Guang-Lei, Ph.D., assistant research professor. His research interests include network measurement and address detection.

ZHANG Wen-Jian, M. S. candidate. His research interests include address detection and machine learning.

LIN Jin-Lei, Ph. D. candidate. His research interests include network measurement and network security.

HAN Dong-Qi, Ph. D. candidate. His research interests include cyberspace situational awareness and network measurement.

WANG Zhi-Liang, Ph. D., associate professor. His research interests include cyberspace situational awareness and network measurement.

ZHANG Hui, M. S., senior engineer. His research interests include network measurement and network management.

YANG Jia-Hai, Ph.D., professor. His research interests include the next-generation Internet architecture and network measurement.

Background

With the development of cloud computing, the Internet of Things (IoT), and 5G technology, the number of Internet users has surged, leading to the depletion of IPv4 address resources. As a result, IPv6 deployment has accelerated globally. For instance, in 2012, less than 1% of Google users accessed services via IPv6. However, by October 2023, this proportion had risen to 44.7%. Similarly, since the World IPv6 Launch event on June 6, 2012, global IPv6 traffic has grown by more than 5000%.

IPv6 active address scanning serves as a crucial foundational technology supporting various IPv6 network applications, such as asset assessment, vulnerability identification, and situational awareness. While IPv6 deployment has addressed the issue of IPv4 address exhaustion, its vast address space presents significant challenges for IPv6 active address scanning. For instance, using the fast scanner ZMap to scan the entire IPv6 address space would take at least millions of years. Therefore, traditional brute-force scanning of the entire address space becomes infeasible.

To address the aforementioned challenges, researchers have employed various technical approaches, including public data acquisition, passive detection, and active probing. These techniques have partially addressed the difficulties in active address discovery but still lack a fast, efficient, and comprehensive solution for active address detection. Particularly, in seedless address regions, there is still a lack of effective detection technology to overcome the low coverage of active addresses.

In response to the current issues, this paper proposes a systematic address discovery approach, categorizing active address detection into two scenarios: seed-based address detection and seedless address detection. In the seed-based address detection scenario, we introduce address generation

technology 6EDL-S based on deep learning and dynamic feedback mechanisms, achieving high hit rates, low time complexity, and large-scale detection in seed-based regions. In the seedless address scenario, we propose the 6EDL-N address detection method based on seed address migration, achieving efficient and high-coverage address detection. Additionally, we design more comprehensive address evaluation metrics to assess the effectiveness of active address detection from multiple perspectives. The experiments verify that the proposed 6EDL algorithm outperforms the current state-of-the-art algorithms in terms of hit rate and detection speed, both in seeded and seedless scenarios. Additionally, the IPv6 hitlist constructed by 6EDL surpasses the currently publicly available IPv6 hitlist in both the quantity and quality of active IPv6 addresses. This research is supported by the National Key Research and Development Project for Internet IP Address Space and Interdomain Routing System Key Information Perception Technology (Project Number: 2022YFB3105001), aimed at perceiving critical information in the Internet and conducting analysis and verification thereof. Additionally, this research also receives funding from Zhongguancun Laboratory Project. This involves addressing the challenges of deep integration and intelligent analysis of multi-source data to achieve multidimensional deep situation awareness capabilities. The authors of this paper have many years of experience in IPv6 active address detection and network measurement, and their representative work has been published in top conferences and journals, such as USENIX ATC'22, NDSS (Distinguished Paper Award), and IEEE/ACM Transactions on Networking (ToN). The authors' team (Network Management and Measurement Laboratory) has accumulated extensive detection data over the years, leading to the construction of a higher-quality active IPv6 address collection known as the IPv6 hitlist. This dataset further reduces the barriers for researchers in the community to study IPv6 network-related applications and security. For more details, please refer to:

https://addrminer.github.io/IPv6_hitlist.github.io/