IPv6地址生成探测调研报告

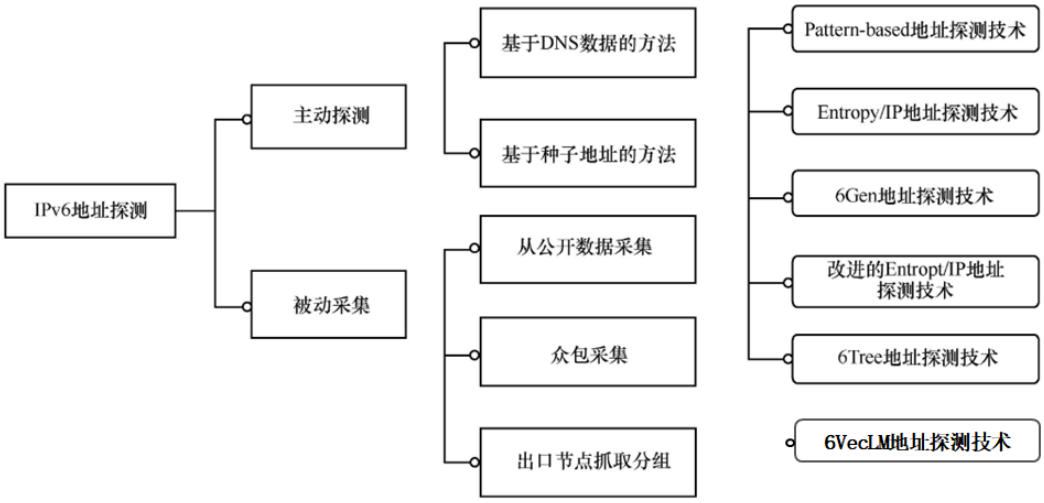
面临的问题

一方面由于IPv6的地址空间(2^{128})相较于IPv4的地址空间(2^{32})有了指数级的变化，使得以目前的技术对全网所有地址进行**遍历扫描变得不可能**，且由于IPv6的特殊格式，其后半部分的接口标识主要采取**随机化生成的方式**，这对资产发现以及网络测量等相关项目造成了很大的干扰。因此，需要尽可能压缩需要探测的地址数量，即***尽可能只探测活跃有效的IPv6地址***

另一方面，由于IPv6的特性，存在某一子网前缀下存在多个地址指向一个主机的情况，这使得针对IPv6的生成测量工作存在了*冗余*，降低了IPv6探测的实际效率，因此，需要使用相应方案对该问题进行解决。

解决方案调研

目前，IPv6地址探测方案可汇总[1]分类为下图：



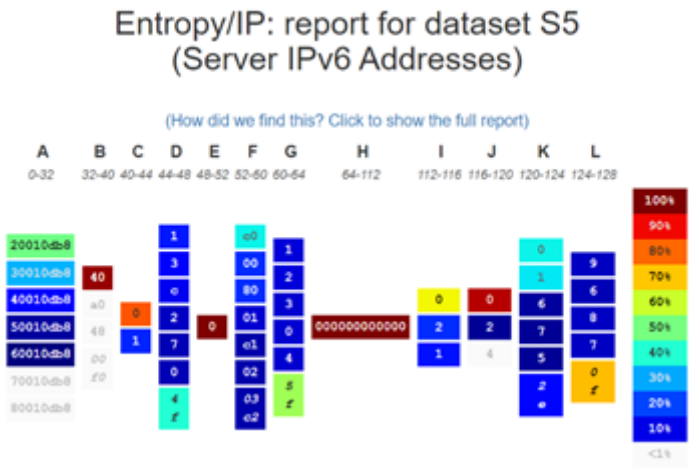
目前，由于其更好的预测与发现性能，业界将发展重点放在了基于种子地址的一类IPv6地址探测技术上。下面对主流的一系列方案进行分析

Pattern-base

自动分析地址的规律，形成pattern，并根据pattern生成新地址，并进行探测。作为该领域的开山之作，该方案**简单，无需人为分析，效果优于暴力搜索**，但是对于集合要求十分严格，易受后期随机化生成技术的影响，效果不佳

Entropy/IP

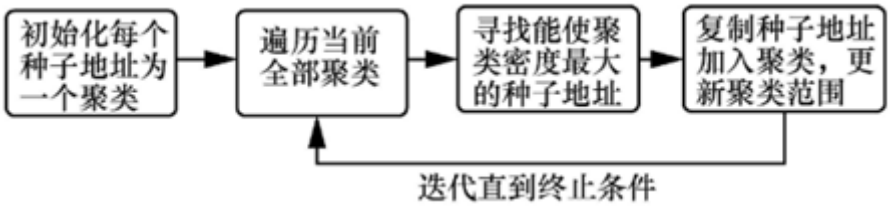
将IPv6地址进行拆分，计算每半字节的熵值，将值相近的合并成为段，并利用贝叶斯网络建模，随机遍历地址空间，从而得到最贴近的预测地址集合。该算法**首次利用了机器学习的方法，并且设计了一套良好的可视化方案**[2]



在6Gen算法提出之后，Entropy/IP又对其工作进行了改进，改进后的方案采用全局遍历，并改用新的别名检测算法和利用ICMPv6协议进行探测，这些对实际的探测效果有一定的帮助。但与pattern-base方案类似，该方案易受随机化影响，随机化会使各段熵值接近为1，造成命中率不高的情况

6Gen

该方案基于一条假设：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。通过识别地址取值空间中的密集区域，在这些区域中取新地址，则新地址存活的可能性最大，该方案流程[1]如下：

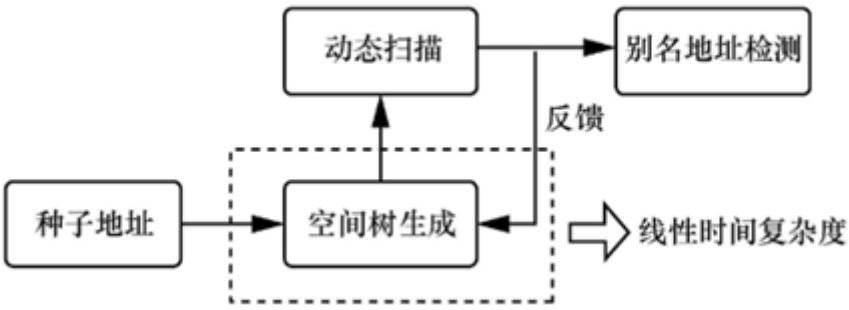


另一方面，作者在其论文里首次提出了别名地址的问题，即在某些地址前缀下，存在大量IPv6地址指向同一个网络接口。针对于此，作者提出了自己的解决方案，在前缀下随机生成地址，若都存活，则认为是别名地址。该思想基于IPv6空间里不同主机的活跃地址是处于稀疏的状态的现实，因此不同主机在同一前缀下的地址均为存活的可能性不大。

该算法的**命中率有所提高**，并首次提出了*别名问题*，并进行初步解决，但由于算法的设计缺陷，多个密集区域经常出现重复的情况，这使得在时间、性能开销过大

6Tree

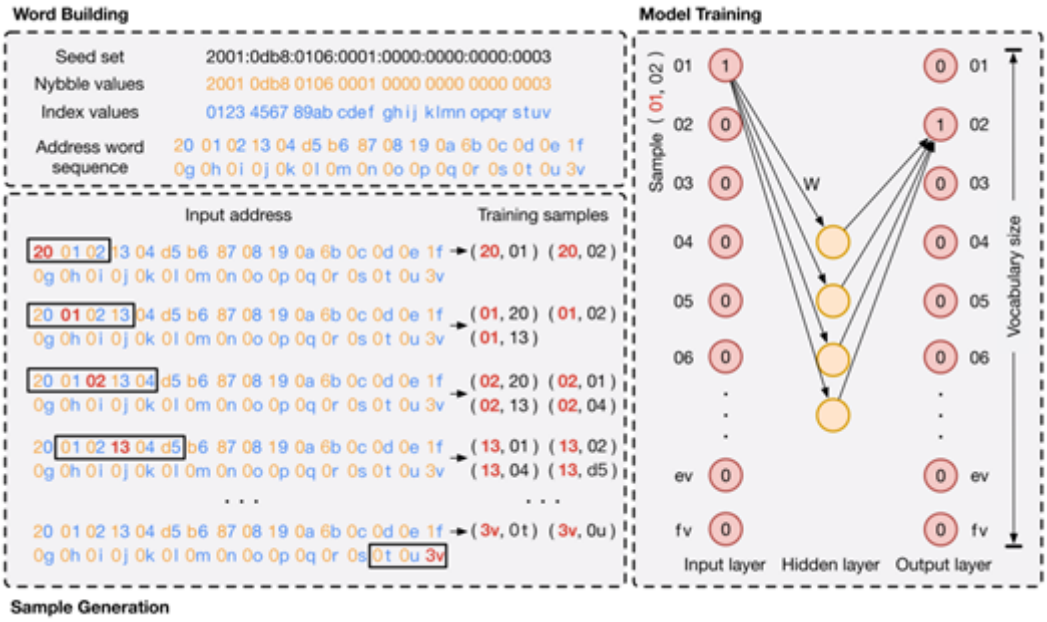
该方案基于分裂聚类树的思想，通过将v6地址转化为高维向量，在每个向量上执行分裂聚类，形成空间树，同时进行扫描与生成[3]，流程如下：



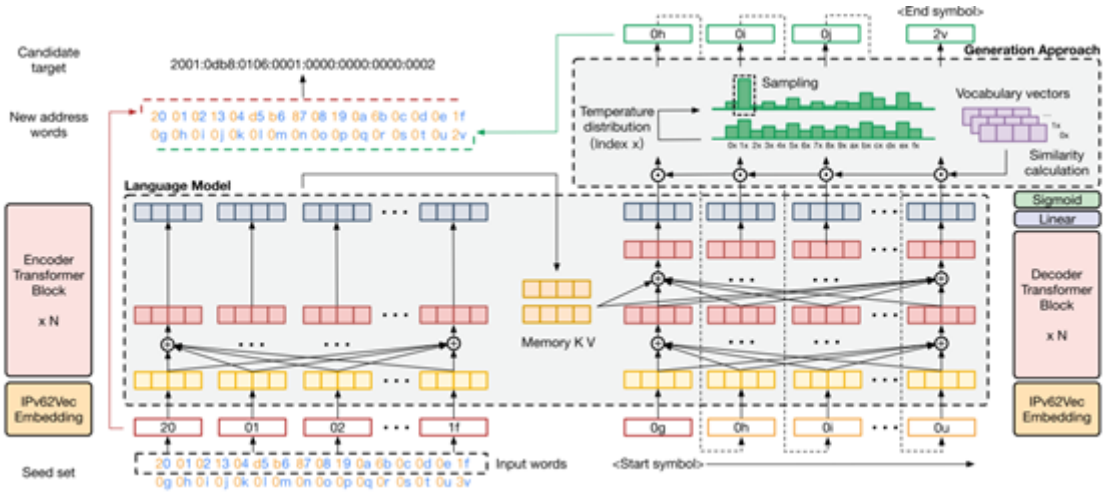
相较于先前的方案，**命中率有了极大提升，能处理大量的种子地址，动态调整生成地址的空间**，并进行别名检测，及时反馈。但由于扫描和地址生成是反馈进行的，随着地址数的增加，耗时提高，同时，交互式的工作流程使得整体效率不高

6VecLM

首次利用自然语言处理的手段，借助IPv62Vec将IPv6转化为向量空间，学习地址之间的语义联系，利用语义进行IPv6地址建模预测，生成最高相似度的地址。



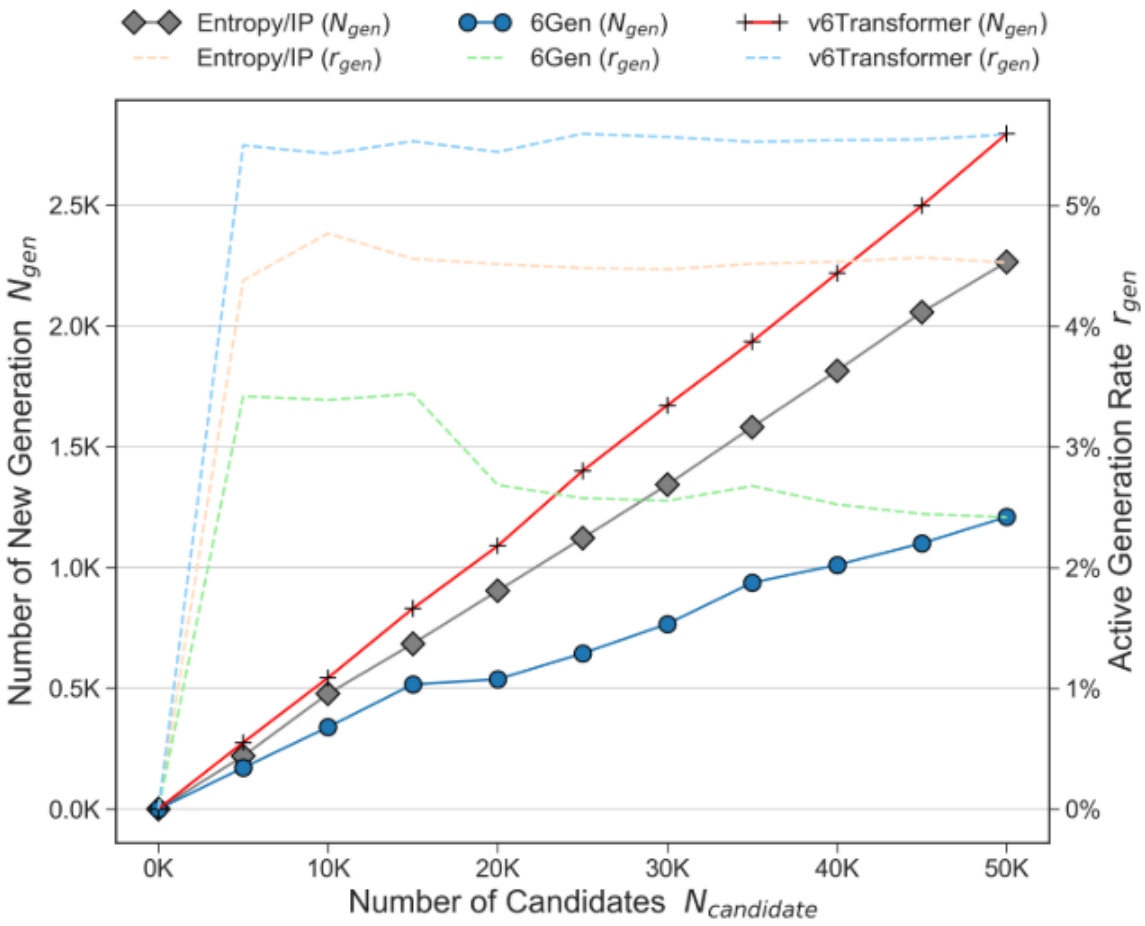
IPv62Vec结构[4]



V6Transformer结构[4]

与之前的方案相比，该算法的**生成率与命中率都达到了一个新的高度**

Category	Model	$N_{candidate}$	N_{hit}	N_{gen}	r_{hit}	r_{gen}
Conventional language model	RNN [17]	34,604	995	851	2.88%	2.46%
	LSTM [14]	34,636	727	564	2.10%	1.63%
	GCNN [6]	34,817	787	649	2.26%	1.86%
Target generation algorithm	Entropy/IP [11]	69,167	8,321	2,540	12.03%	3.67%
	6Gen [19]	67,712	4,612	1,638	6.81%	2.42%
Adding IPv62Vec and generation approach	RNN [17]	44,242	12,133	2,409	27.42%	5.44%
	LSTM [14]	61,950	10,640	2,019	17.18%	3.26%
	GCNN [6]	52,046	11,360	2,146	21.83%	4.12%
Our approach	6VecLM	46,461	**15,406**	**2,883**	**33.16%**	**6.21%**



但由于算法的原理不同，在生成地址数方面低于6Tree算法,且由于需要进行模型的训练以及调优，整体的工作难度以及耗时也比较大。

汇总

对比项目	Pattern-based	Entropy/IP&改进型	6Gen	6Tree	6Ve
年份	2015	2016/2018	2017	2019	202
处理较大地址耗时	最快	中等	最慢	快	较慢
关键技术	递归	贝叶斯网络	聚类	分裂树	自然言处

对比项目	Pattern-based	Entropy/IP&改进型	6Gen	6Tree	6Ve
是否考虑别名	否	否/是	是	是	否
可视化能力	较弱	较强	较弱	较强	较弱
扫描是否有反馈	无	无	无	有	有

对比项目	Pattern-based	Entropy/IP&改进型	6Gen	6Tree	6Ve
是否公开数据集	否	否/是	否	否	否
是否开源代码	否	否/是	否	是	是
论文中发现地址量级	10^3	10^6/10^8	10^6	10^8	10^8

参考资料

[1] GUO L, LIN H, GUANGLEI S等. 基于种子地址的IPv6地址测量.(2019). DOI:10.11959/j.issn.1000-0801.2019296.

[2] FOREMSKI P, PLONKA D, BERGER A. Entropy/IP: Uncovering structure in IPv6 addresses[J/OL]. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, 2016, 14-16-Nove: 167–181. http://arxiv.org/abs/1606.04327. DOI:10.1145/2987443.2987445.

[3] LIU Z, XIONG Y, LIU X等. 6Tree: Efficient dynamic discovery of active addresses in the IPv6 address space[J]. Computer Networks, 2019, 155. DOI:10.1016/j.comnet.2019.03.010.

[4] CUI T, XIONG G, GOU G等. 6VecLM: Language Modeling in Vector Space for IPv6 Target Generation[J]. arXiv, 2020.