



下一代互联网活跃地址探测 技术体系研究与实践

清华大学

任 罡

CERNET第三十一届学术年会，哈尔滨



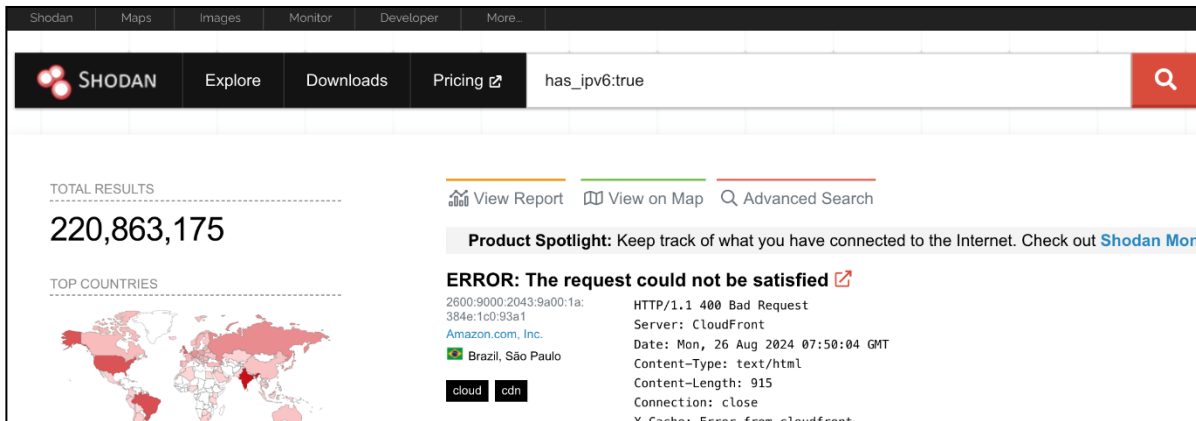
主要内容

- 下一代互联网活跃地址探测的背景和意义
- 国内外研究进展
- 活跃地址探测技术体系设计
- 面向种子密集区域的基于实时模式生成的活跃地址探测
- 面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测
- 基于被动分析和主动检测融合的别名前缀检测
- 相关实践与成果



背景和意义

- IPv6的规模部署对网络空间安全带来新的挑战。传统的IPv4地址扫描方法不再可行。IPv6活跃地址探测，从RFC5157到RFC7707，从公开信息提取，被动监测到主动测量，技术不断演进发展。
- IPv6活跃地址探测是下一代网络空间测绘、网络安全评估、网络空间安全态势感知等技术的基础，有利于：
 - 发现IPv6网络空间中的各类资产，为网络管理和测量等提供数据支持；
 - 定位IPv6网络空间中的薄弱环节，抢占网络空间安全对抗信息高地；
 - 把握IPv6网络空间整体安全态势，为国家战略安全决策提供依据。



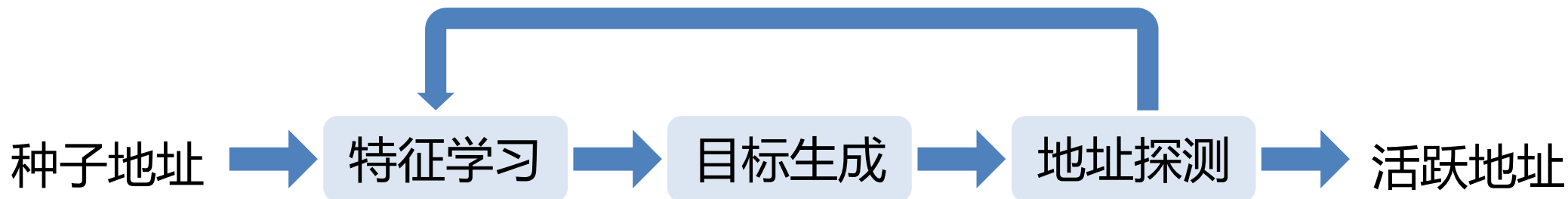


主要内容

- 下一代互联网活跃地址探测的背景和意义
- 国内外研究进展
- 活跃地址探测技术体系设计
- 面向种子密集区域的基于实时模式生成的活跃地址探测
- 面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测
- 基于被动分析和主动检测融合的别名前缀检测
- 相关实践与成果



面向种子地址密集区域的IPv6活跃地址探测研究进展



■ 早期方法：Pattern-based、Entropy/IP

分类	代表方法	特征形式	优点	局限
基于聚类	6Gen	聚类模式	模式挖掘能力较强	算法复杂度高 $O(N^3)$; 不适用大规模地址扫描
	6Tree、6Hit、6Forest、6Graph、6Scan、DET、AddrMiner-S、HMap6、6Subpattern、6Loda等	聚类模式 (密度空间树)	算法效率高, 适用于大规模地址扫描	模式挖掘能力和命中率有待进一步提高。
基于深度学习	6GCVAE、6VecLM、6GAN、6Former、6SENSE、6GAI、6Diffusion等	神经网络	理论上具备更强的特征学习能力	算法开销较大; 大规模地址扫描受限



面向种子稀疏区域的IPv6活跃地址探测研究进展

- 稀疏区域：种子地址数量小于等于某一阈值的BGP前缀空间或AS空间
- 关键问题：如何在缺少有效模式的情况下生成稀疏区域的目标地址

分类	方法	地址生成方法	优点	局限
基于特定模式	Low-Bytes	BGP_prefix::*	小规模预算下命中率较高	大规模预算下命中率低
	Random-Bytes	BGP_prefix::**** BGP_prefix::****:0001	部分缓解Low-Bytes的问题	模式的选择主观性较大
基于模式迁移	GAG、AddrMiner-N	2001::** →BGP_Prefix::**	利用区域相关性提高命中率	可迁移模式挖掘能力相对有限
	6Rover	2001::** →BGP_Prefix::**	利用强化学习提高命中率	可迁移模式挖掘能力相对有限
基于机器学习	6Vision	图像生成技术PixelCNN	特征学习能力更强	算法开销较大、未利用区域相关性
	6Diffusion	扩散模型+Transformer		



IPv6别名前缀检测研究进展

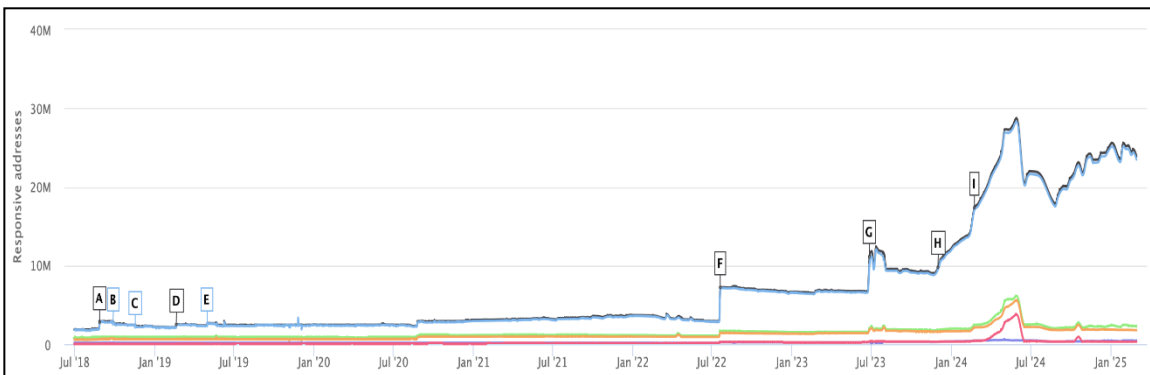
- IPv6别名前缀：整个前缀绑定在同一个接口，通常该前缀下任意地址均是可响应的。
- 对IPv6地址扫描的影响：极大浪费探测资源，大幅降低扫描结果的可信度。

分类	方法	检测方法	优点	局限
基于概率	APD、6SENSE-APD	在/96前缀下随机探测3个地址	简单、高效	长度固定准确性不高
	6Tree-APD、MAPD、HMap6-APD、Luori	探测16个子前缀下的随机地址等	不受别名子前缀干扰	易受网络丢包干扰，开销较大
基于指纹	Speedtrap、FBAR	利用协议栈指纹区分不同主机	可区分活跃前缀和别名前缀	易受网络丢包干扰，协议栈指纹准确性不高
	UAV6、FAPD	利用MTU缓存区分不同主机	可区分活跃前缀和别名前缀，准确性高	易受网络丢包干扰，适用范围受限

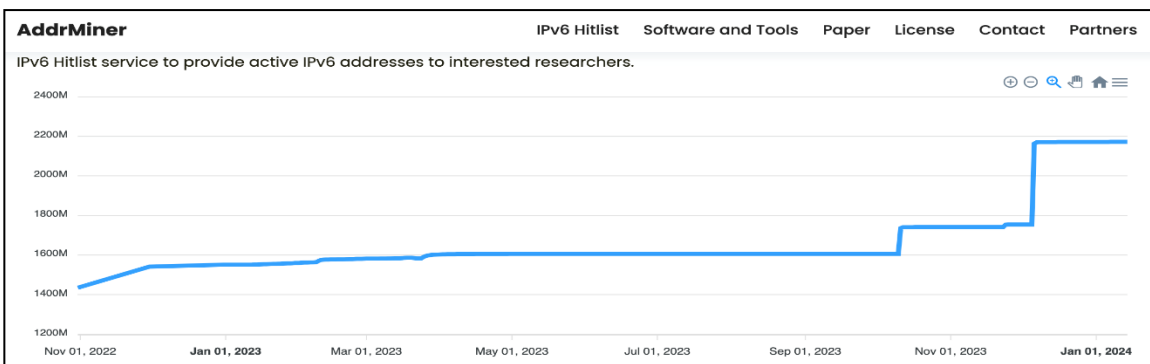


全球IPv6活跃地址探测系统研究进展

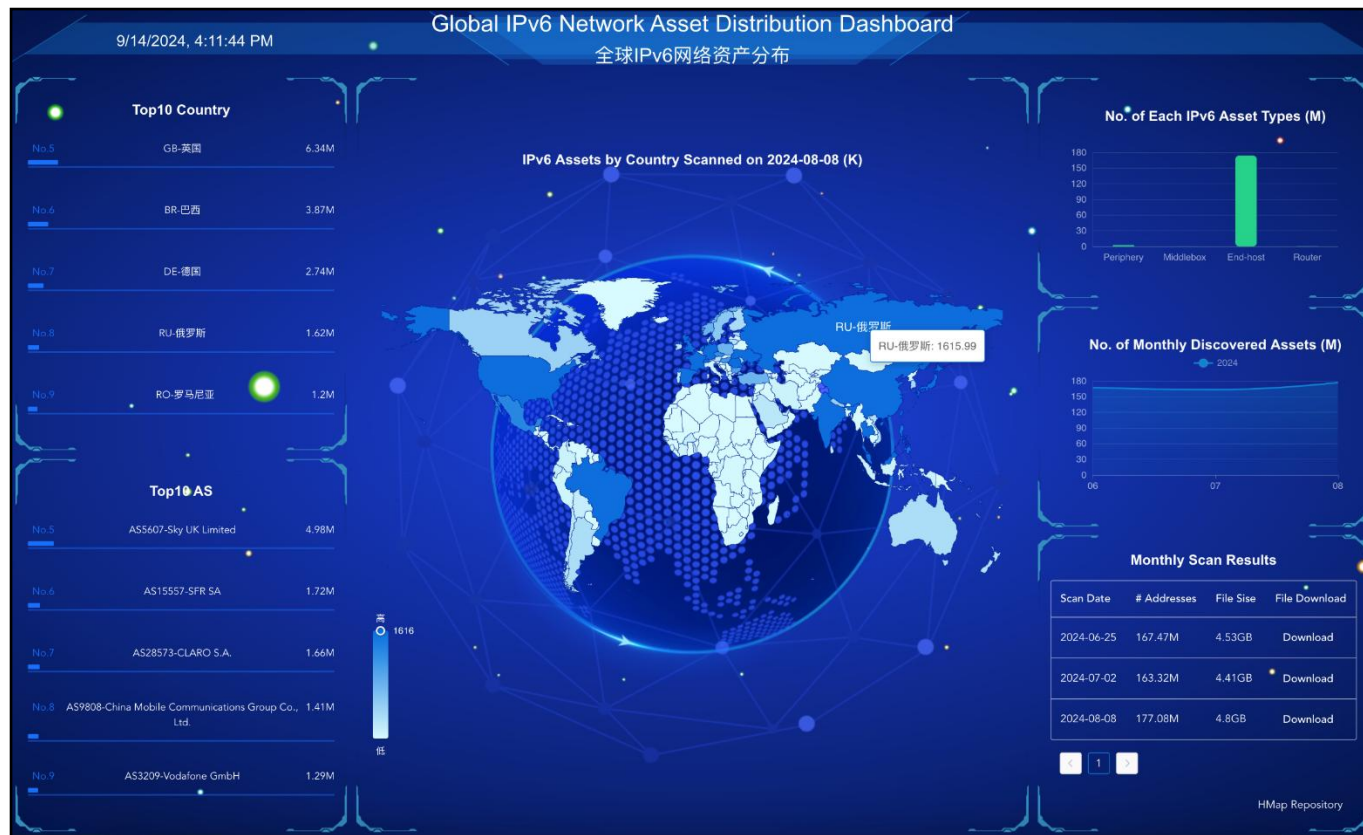
- 除了商业化资产测绘系统外，学术界也提出了一些卓有影响的IPv6活跃地址探测系统，包括IPv6 Hitlist, AddrMiner和HMap等



IPv6 Hitlist: 累计地址最多3.2B，活跃地址23M



AddrMiner Hitlist: 累计地址2.2B，活跃地址74M



HMap Hitlist: 活跃地址最多177M



主要内容

- 下一代互联网活跃地址探测的背景和意义
- 国内外研究进展
- 活跃地址探测技术体系设计
- 面向种子密集区域的基于实时模式生成的活跃地址探测
- 面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测
- 基于被动分析和主动检测融合的别名前缀检测
- 相关实践与成果



IPv6活跃地址探测数据集多维指标分析

现有IPv6地址扫描方法评价指标:

- 命中率
- 新发现网络数
- 网络覆盖数
- 累积分布曲线

未能全面刻画扫描结果的分布特性

多维度IPv6活跃地址探测数据集分布偏差指标:

- 地址分类维度: 国家、AS、BGP前缀
- 量化地址在头部的聚集和尾部的稀疏程度:
 - 稀疏指标: 稀疏区域占比 $s_i = \frac{1}{k} \sum_j 1_{\|X_j\| \leq i}$
 - 稠密指标: 稠密地址占比 $c_i = \frac{1}{\|X\|} \sum_{j=1}^i \|Y_j\|$

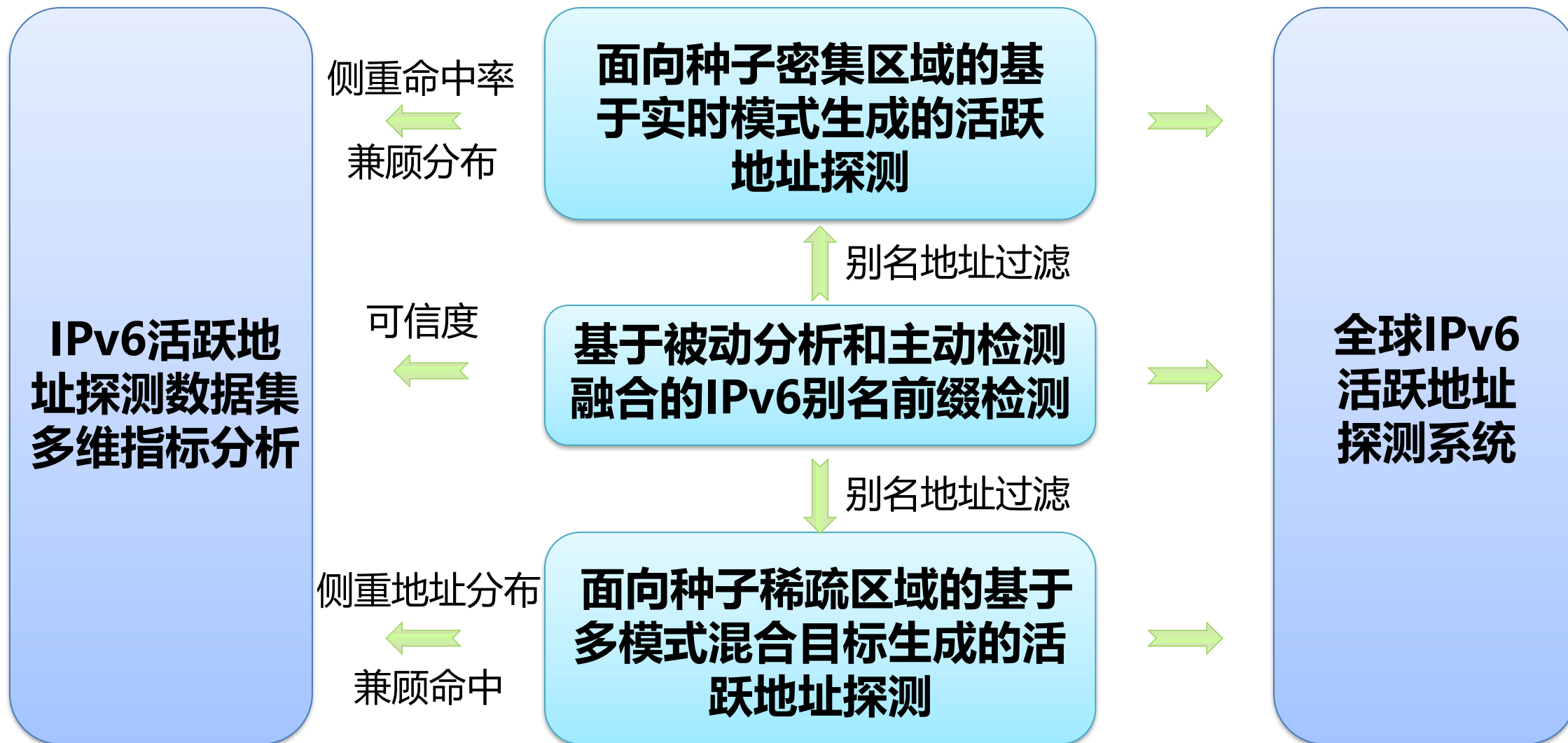
代表性的活跃地址探测数据集分析:

- 头部集中性与稠密区域探测价值
 - 绝大部份地址来自少数高密度AS
 - 深耕稠密区域依然是实现大规模获取活跃地址的高效路径
- 尾部稀疏性与稀疏区域探测挑战
 - 绝大部份AS已知地址数量十分稀少
 - 现有方法未实现对稀疏区域有效覆盖

数据集	维度	稀疏性 (Sparsity)					稠密性 (Concentration)				
		s	s_0	s_1	s_{10}	s_{100}	c	c_1	c_{10}	c_{100}	c_{1000}
D_{Gasser} (21M)	国家/地区	7.44%	4.55%	5.45%	15.45%	34.55%	95%	31.58%	82.35%	99.92%	-
	AS	48.51%	38.54%	47.03%	70.42%	90.86%	98.95%	20.17%	48.92%	83.44%	96.43%
	BGP 前缀	71.44%	63.48%	72.70%	86.29%	95.91%	99.12%	7.84%	33.57%	69.35%	87.81%
D_{Addr} (74M)	国家/地区	9.19%	4.55%	8.64%	17.27%	40.00%	99.31%	89.38%	97.92%	99.98%	-
	AS	53.78%	41.24%	55.45%	77.55%	93.86%	99.81%	85.52%	92.27%	97.12%	99.32%
	BGP 前缀	75.54%	68.55%	76.92%	88.15%	95.96%	99.55%	2.49%	12.77%	58.26%	93.10%
D_{HMap6} (177M)	国家/地区	8.29%	5.91%	6.82%	14.55%	33.64%	97.68%	43.36%	93.66%	99.99%	-
	AS	53.50%	44.95%	51.96%	72.62%	90.74%	99.83%	40.17%	84.32%	95.88%	99.51%
	BGP 前缀	76.92%	71.24%	77.76%	87.50%	95.23%	99.86%	12.88%	65.96%	90.54%	97.32%



活跃地址探测技术体系设计





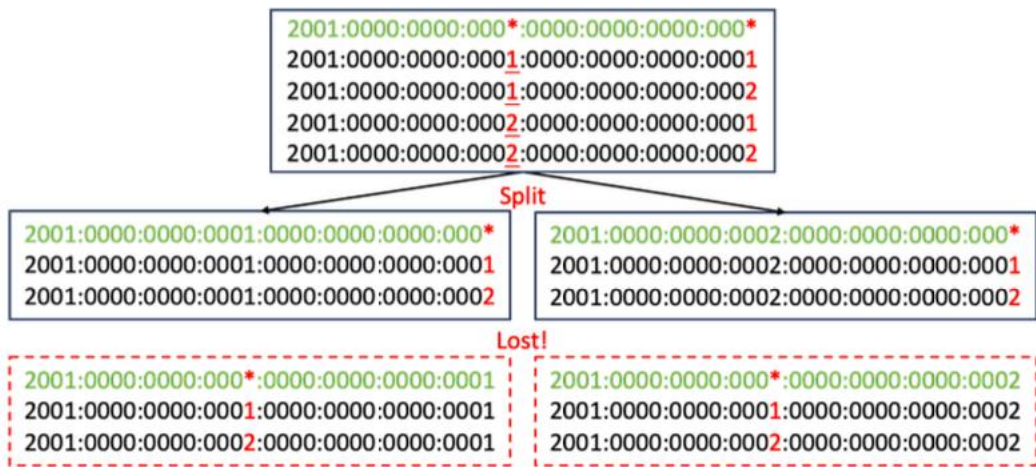
主要内容

- 下一代互联网活跃地址探测的背景和意义
- 国内外研究进展
- 活跃地址探测技术体系设计
- 面向种子密集区域的基于实时模式生成的活跃地址探测
- 面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测
- 基于被动分析和主动检测融合的别名前缀检测
- 相关实践与成果



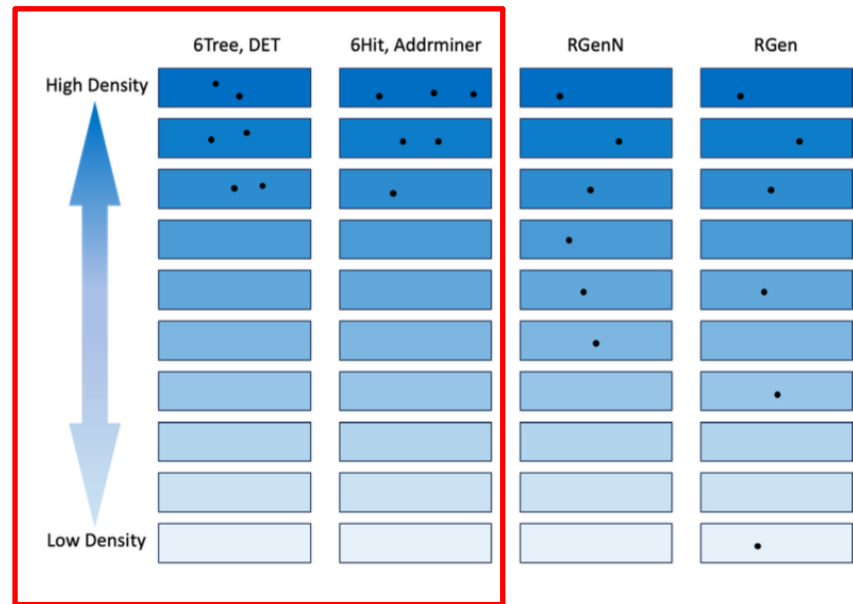
面向种子密集区域的基于实时模式生成的活跃地址探测 设计思想

现有基于分裂层次聚类算法的模式挖掘



存在问题：部份高密度模式丢失；
 实时发现新的地址模式开销较大

现有基于地址密度反馈的目标地址生成



存在问题：目标地址多样性不足

解决思路

- 去聚类化：放弃传统的聚类机制，转而从单个种子地址直接生成模式，避免高密度模式丢失；
- 实时处理：设计流式架构，便于利用新探测到的活跃地址进行实时反馈以发现新的模式；
- 随机策略：引入基于密度的随机选择机制，平衡高密度模式的深挖与低密度模式的探索。

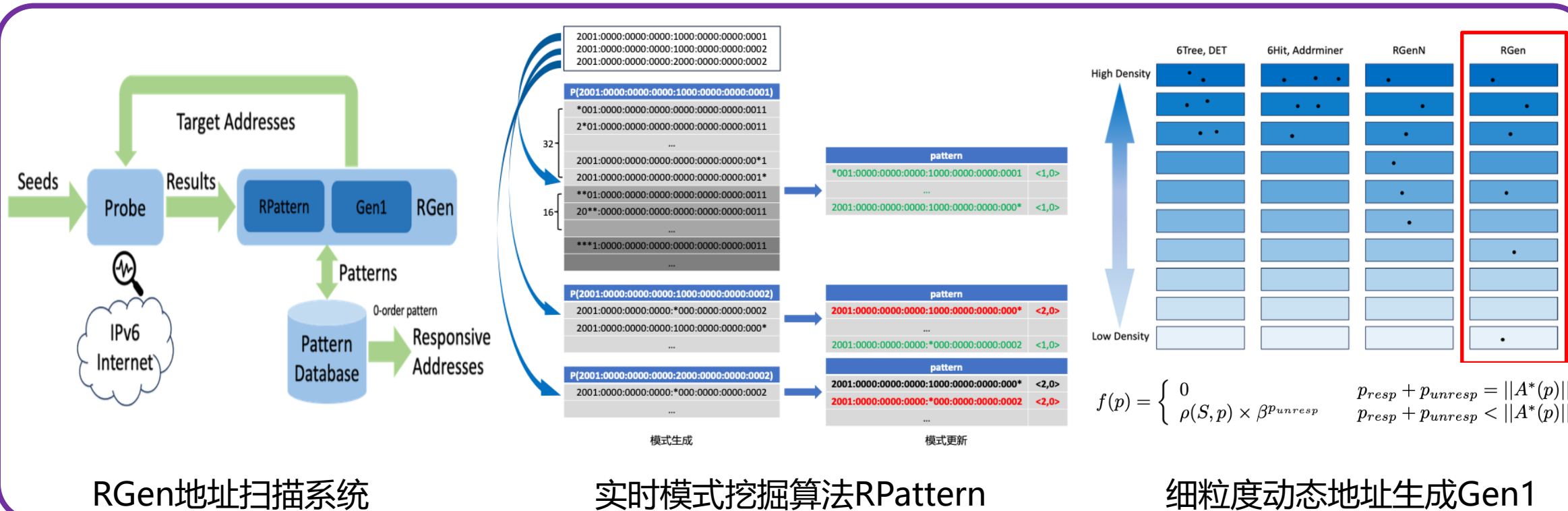


面向种子密集区域的基于实时模式生成的活跃地址探测 技术方案

核心组件

基于实时模式生成的活跃地址探测系统RGen

- 实时模式挖掘算法RPattern：实时利用种子地址和新发现的地址挖掘地址模式，针对每个地址生成一组优选模式，避免模式空间爆炸；通过维护模式数据库来避免耗时的聚类。
- 细粒度动态地址生成算法Gen1：采用轮盘赌算法选择模式，而非仅选择最高密度模式，每个选中模式每轮最多生成1个地址，兼顾目标地址高命中率和多样性。





面向种子密集区域的基于实时模式生成的活跃地址探测 评估分析

实验设置：种子地址21M，预算100M

TGA	De-aliased Hit	new AS	new BGP Prefix	new /64 Prefix
6Tree	29M	3	165	6M
DET	31M	231	1,190	8M
AddrMiner-S	36M	0	221	9M
6Subpattern	47M	473	2,108	11M
6Sense	7M	46	1,318	4M
RGenN	59M	22	699	11M
RGen	58M	129	1,339	12M

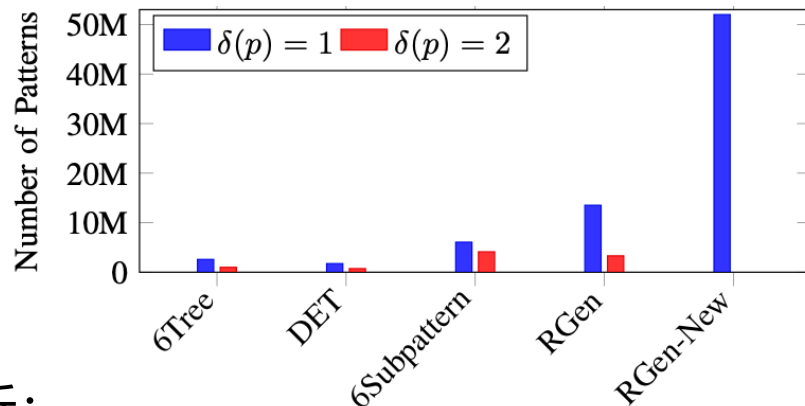
实验结果：

- **命中率58%**，比现有最佳方法6Subpattern高出 23%。
- **网络发现能力**：发现1339个新的BGP前缀和1200万个新的/64子网，远超不使用随机策略的对照组方法。

RGen在取得高命中率的同时，保持了较高的对未知网络区域的探索能力

模式挖掘能力分析：

- **初始挖掘1阶模式1400万**，是6Subpattern的2倍
- **实时挖掘1阶模式5200万**，极大扩展高密度区域



性能分析：

- 种子地址处理速率 44k/s
- 地址生成速率 456k/s
- 满足大规模扫描

Method	Time (min)		Memory (GB)	
	Init	Generation	Init	Generation
6Tree	73	137	17	49
DET	117	102	18	50
AddrMiner-S	109	4,574	18	311
6Subpattern	311	54	19	54
6Sense	480*	3,100	92*	31
RGen	8	219	70	262



主要内容

- 下一代互联网活跃地址探测的背景和意义
- 国内外研究进展
- 活跃地址探测技术体系设计
- 面向种子密集区域的基于实时模式生成的活跃地址探测
- 面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测
- 基于被动分析和主动检测融合的别名前缀检测
- 相关实践与成果



面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测 设计思想

现状：种子稀疏区域广泛存在

Region	ASes	%	Addresses	%
r_0	13,338	38.55	0	0
r_1	3,708	10.72	3,708	0.02
r_{10}	8,387	24.24	38,388	0.17
r_{100}	6,396	18.49	216,418	0.96
r	31,829	92.00	258,514	1.14

基于IPv6 Hitlist 活跃地址集：

- 92%的AS，发现活跃地址不超过100个
- 39%的AS，尚未发现活跃地址

现有聚类方法在稀疏区域局限

- 依赖种子地址临近聚集，在种子稀疏区域失效。
- 混淆“BGP前缀相似性”与“管理员配置习惯相似性”，未充分学习跨网络配置规律。
- 为处理海量数据采用近似算法，但在数据稀疏时反而造成信息丢失。

现有针对稀疏区域的地址扫描方法在覆盖率和命中率方面尚需提高。

解决思路

- 可迁移模式学习：利用IPv6地址语义（前后缀解耦）学习真正的跨网络配置规则。
- 孤立种子利用：最大化利用稀疏区域中少量宝贵的种子地址。
- 聚类模式最大化：针对种子数量稀少的特点，持续进行实时、完备的聚类分析。

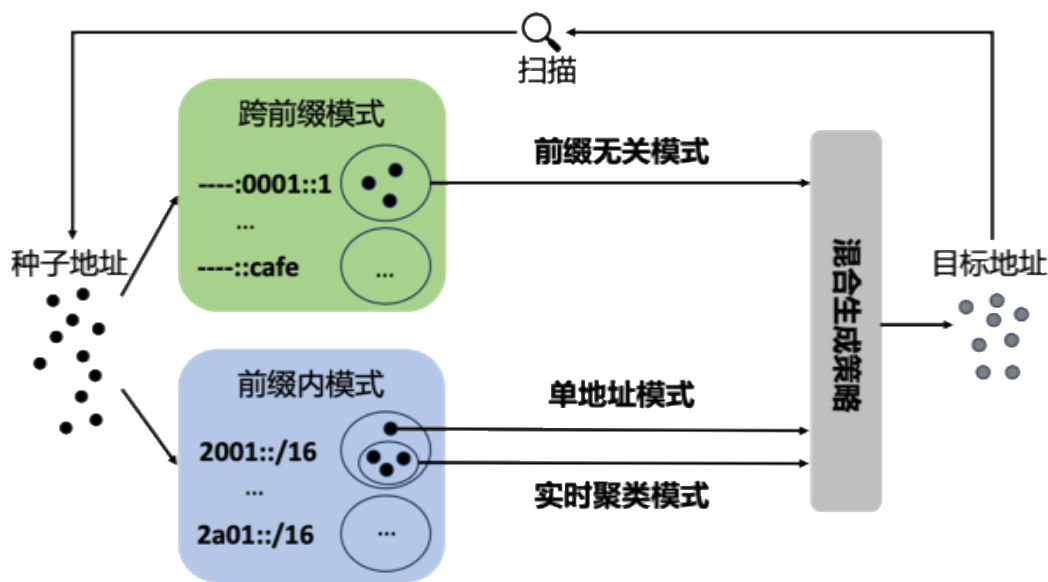


面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测 技术方案

核心组件

混合模式生成的活跃地址探测系统HGen，有三种互补的地址模式：

- 前缀无关模式：用于学习跨网络地址配置模式，聚类时完全忽略BGP前缀，只关注地址的后缀部分，主要负责广度探索。
- 单地址模式：用低成本泛化机制去激活孤立地址，从而提高稀疏区域种子的利用效率。
- 实时聚类模式：针对稀疏区域地址数量少的特点，任何新发现地址，都被立即用于进行一次完备的聚类分析，将聚类模式的高命中率优势发挥到极致，主要负责深度挖掘。



HGen地址扫描系统

```

2001::

```

前缀无关模式

```

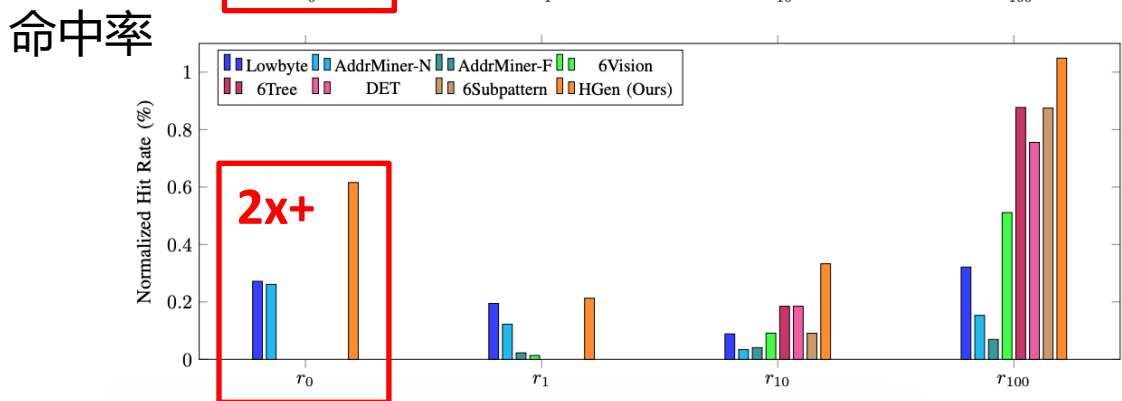
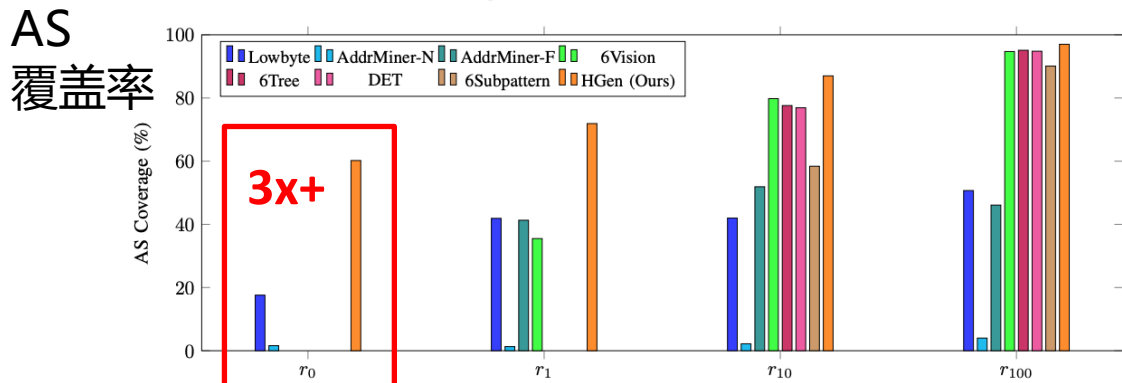
seed  2001:0000:0000:0001:0192:0168:0001:0252
      2001:0000:0000:0001:0192:0168:0001:025*
      2001:0000:0000:0001:0192:0168:0001:02*2
      ...
      2001:0000:0000:000*:0192:0168:0001:0252
  
```

单地址模式



面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测 评估分析

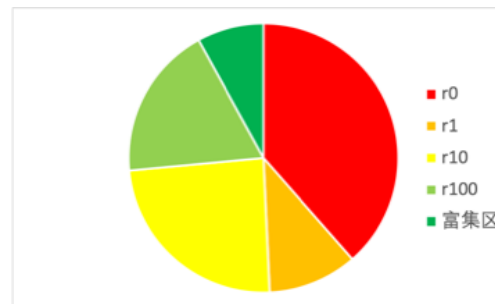
实验设置：4类不同稀疏程度的场景，各随机抽取1000个AS，每个AS预算为4k



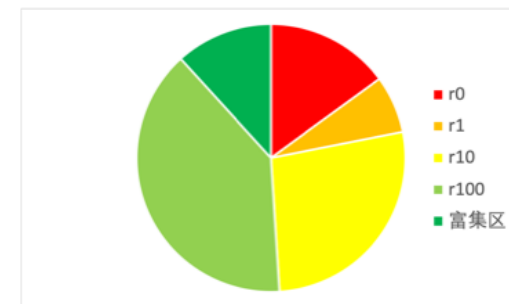
实验结果：HGGen在所有稀疏场景下均取得最佳的AS覆盖率和命中率。特别在最难的无种子场景，覆盖率60%，是最佳基准方法3倍以上。命中率为最佳基准方法2倍以上。

大规模稀疏区域扫描：

- 扫描全部31829个稀疏AS。
- 发现超过100万个活跃地址，覆盖AS数量超2.4万个。
- 无种子AS由13338减少到5170个，大幅拓展了IPv6网络的已知边界。



扫描前



扫描后



主要内容

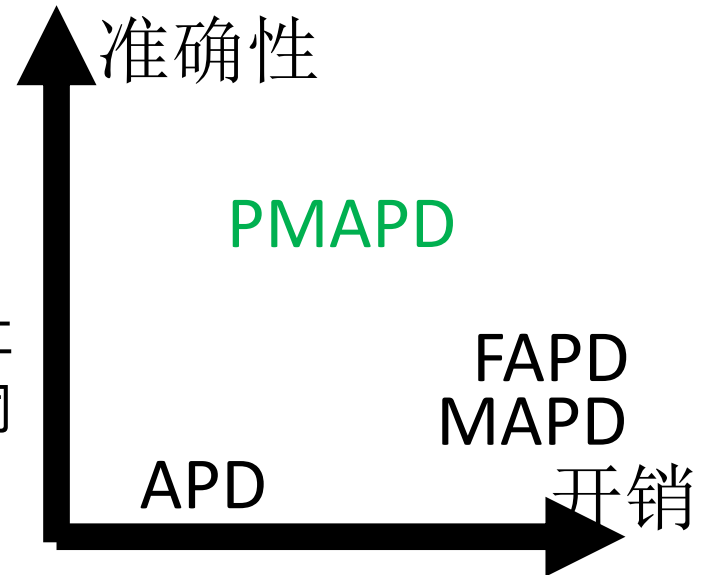
- 下一代互联网活跃地址探测的背景和意义
- 国内外研究进展
- 活跃地址探测技术体系设计
- 面向种子密集区域的基于实时模式生成的活跃地址探测
- 面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测
- 基于被动分析和主动检测融合的别名前缀检测
- 相关实践与成果



基于被动分析和主动检测融合的IPv6别名前缀检测 设计思想

现有IPv6别名前缀检测方法局限：准确性，开销，适用范围等

- 基于概率：利用非别名前缀内地址响应率极低的特性
 - 固定前缀长度，如APD，探测/96下3个随机地址
 - 层次化长度，如MAPD，探测/64至/124下16个随机地址
- 基于指纹：利用协议栈指纹或MTU缓存等机制，验证不同IPv6地址是否指向同一主机，依赖准确的主机指纹获取



解决思路：

- 被动分析优先：复用扫描过程中的响应性、端口、指纹等先识别非别名前缀，只对可疑前缀进行主动检测。
- 层次化检测：支持从BGP前缀长度到 /112 的灵活检测粒度，参考HMap6-MAPD。
- 优化主动探测：采用随机子前缀应对别名子前缀干扰；根据前缀长度调整阈值，平衡误报率与漏报率。

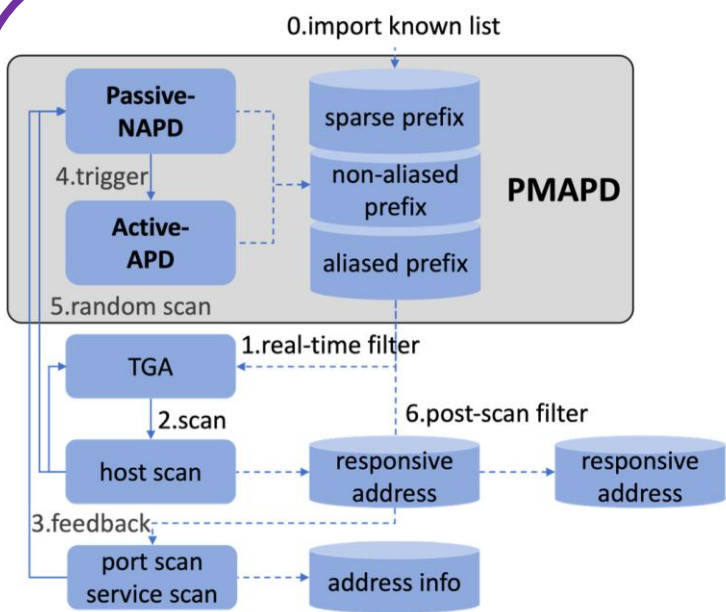


基于被动分析和主动检测融合的IPv6别名前缀检测 技术方案

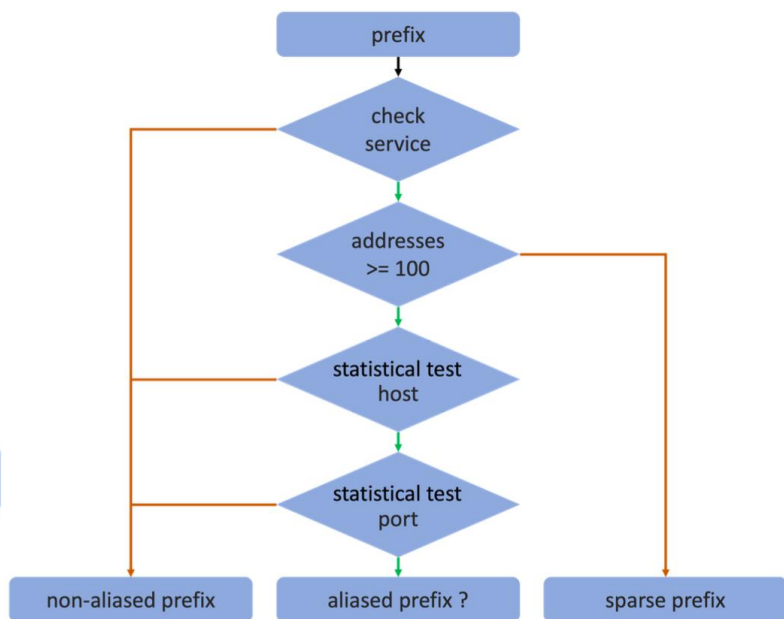
核心组件

被动分析和主动检测融合的别名前缀检测系统PMAPD:

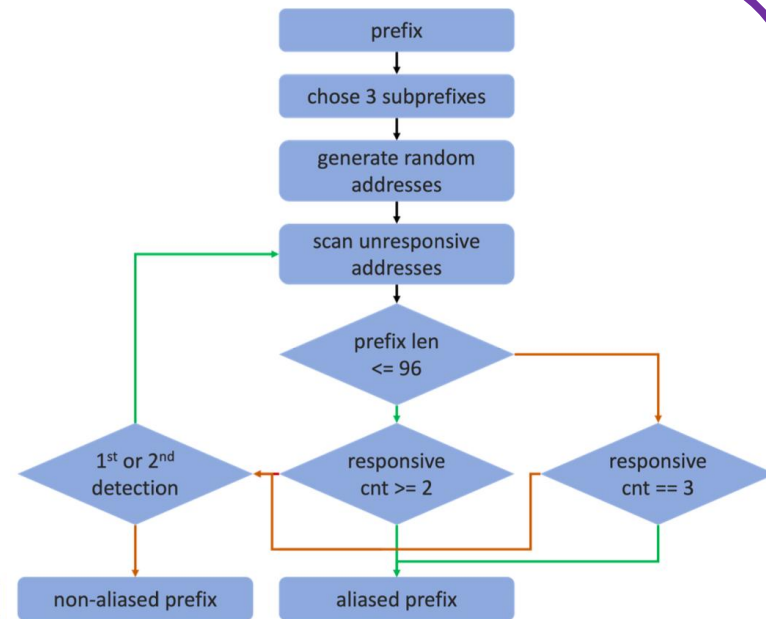
- 被动非别名前缀检测 (Passive-NAPD) : 利用已有的扫描信息识别出非别名前缀, 大幅降低主动检测的开销, 提高检测的准确性。
- 优化的主动别名前缀检测 (Active-APD) : 不仅融合已有多种方法的优点, 而且引入了随机子前缀、动态别名前缀阈值等机制, 在检测的准确性和开销方面取得较好的平衡。



PMAPD与TGA集成示意图



Passive-NAPD流程图



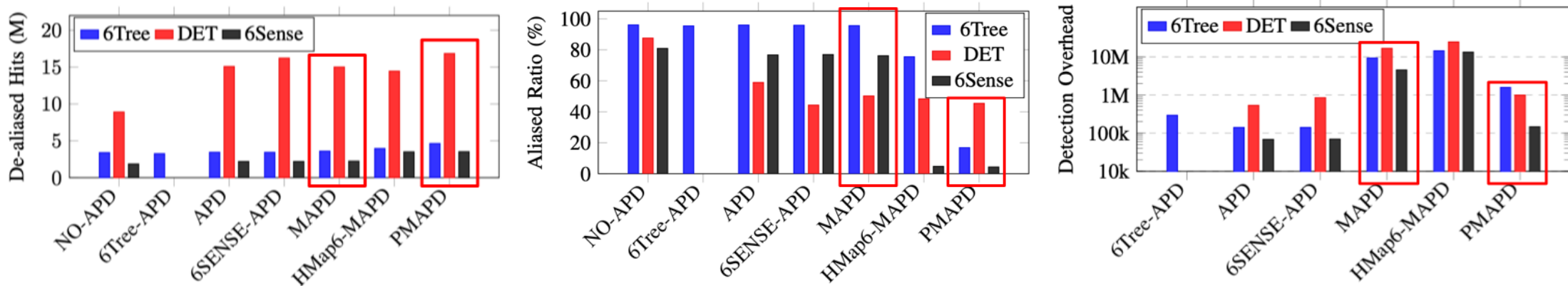
Active-APD流程图



基于被动分析和主动检测融合的IPv6别名前缀检测 评估分析

实验设定:

- 6Tree, DET, 6Sense三种主流TGA算法
- 6Tree-APD, APD, 6Sense-APD, MAPD, HMap6-MAPD等对比方法



实验结果 (与MAPD等相比) :

- 更高的有效命中: 目标生成算法的非别名地址数量提升了 12% 至 57%
- 更干净的扫描结果: 扫描结果中的别名地址率降低了 9% 至 94%
- 更低的探测开销: 探测开销仅为 3% 至 17%

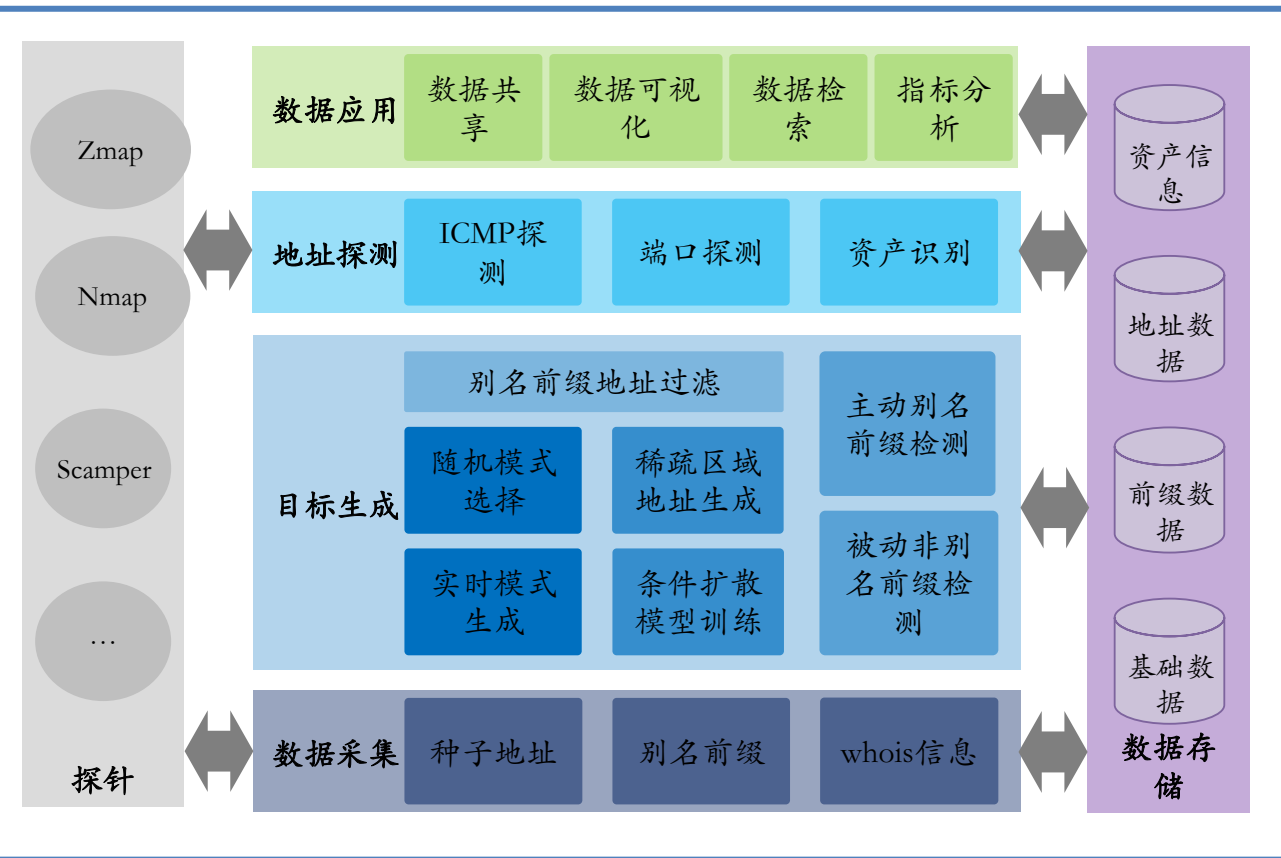


主要内容

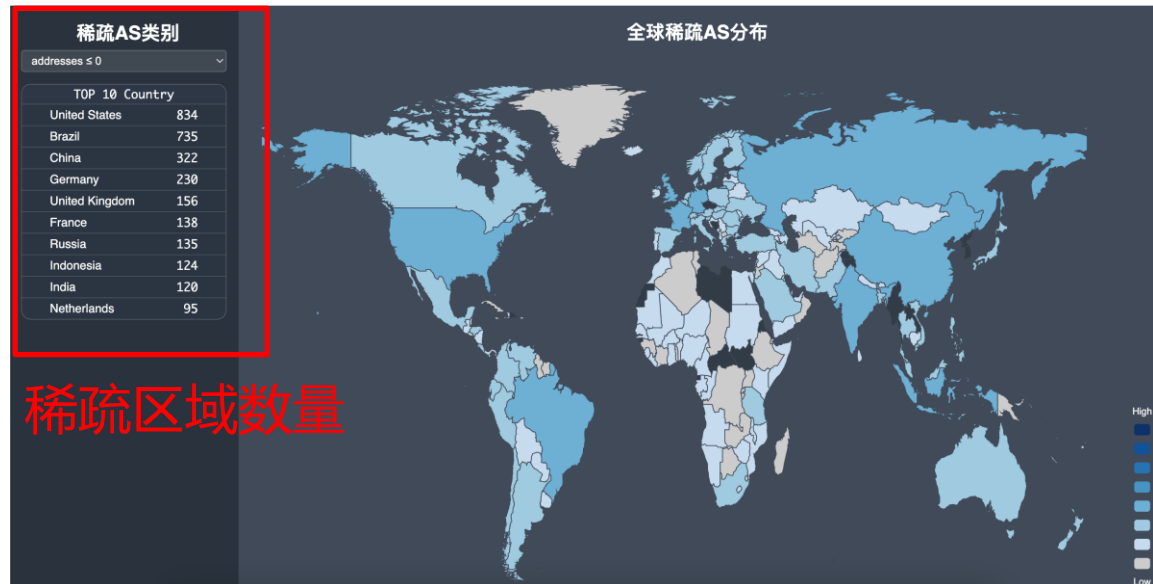
- 下一代互联网活跃地址探测的背景和意义
- 国内外研究进展
- 活跃地址探测技术体系设计
- 面向种子密集区域的基于实时模式生成的活跃地址探测
- 面向种子稀疏区域的基于多模式混合目标生成的活跃地址探测
- 基于被动分析和主动检测融合的别名前缀检测
- 相关实践与成果



全球活跃IPv6活跃地址探测系统



全球活跃IPv6地址探测系统架构

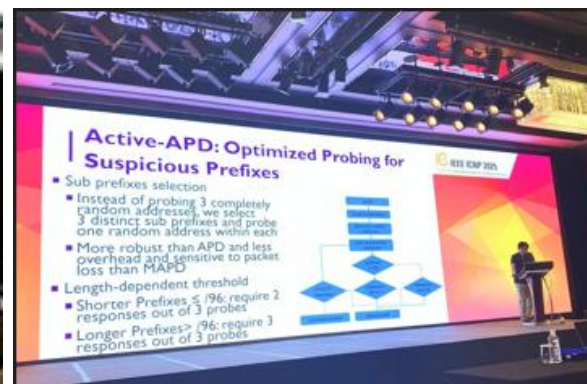


稀疏区域数量



相关实践与成果

成果连续在APNIC58, APNIC59, APNIC60, ICNP2025, IWQoS2025等会议汇报



连续在下一代互联网技术创新大赛获奖并入选教育部教育系统IPv6规模部署与应用优秀案例



获互联网域名管理技术国家工程实验室开放课题资助 (KF202502)

获国家实验室态势感知项目支持



感谢聆听

rengang@cernet.edu.cn