信息安全专业　2021级本科B210314班　毕业论文答辩

基于多源数据的路由起源验证技术研究

南京邮电大学

答辩人：B21031402谢睿熙
导　师：　吴争

目录

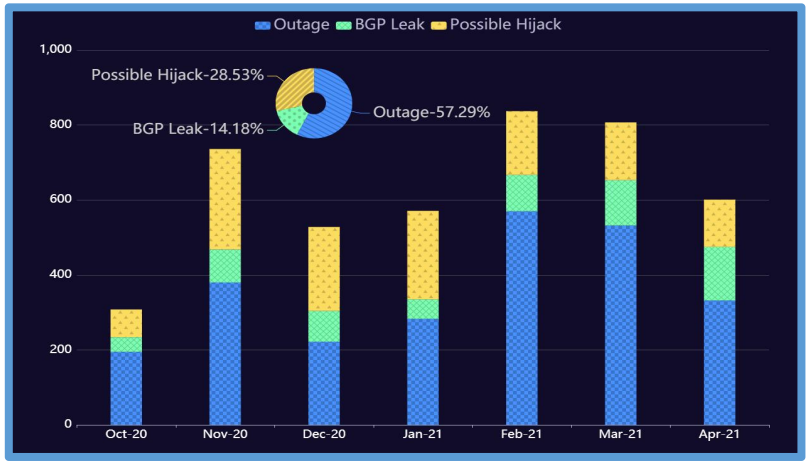
1

1.选题背景及意义

BGP协议

• BGP协议的核心地位

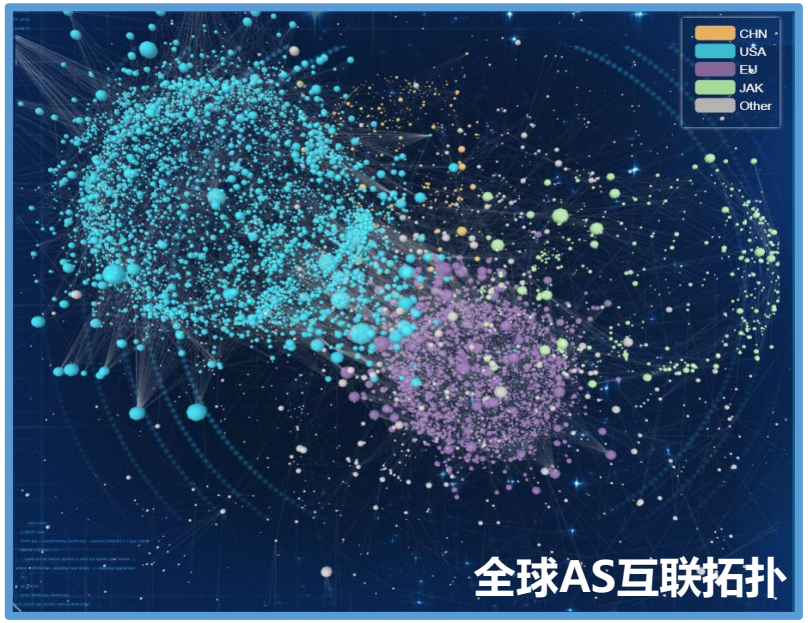
边界网关协议（BGP）是互联网域间路由的核心协议，负责**在自治系统（AS）之间交换路由信息**，构建全球路由网络，对互联网的稳定运行至关重要。



数据来源： BGPStream, https://www.bgpstream.com/

• BGP路由安全

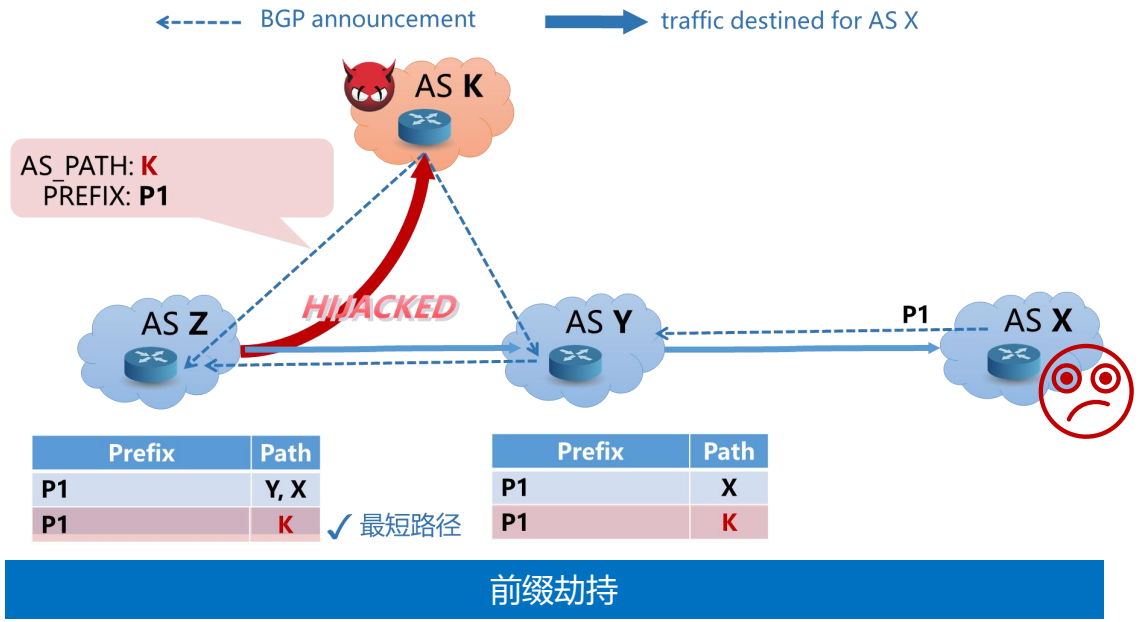
BGP在设计之初，没有充分考虑安全机制。例如，由于**缺乏路由认证机制**，使得路由消息容易被篡改。



前缀劫持

攻击机理：前缀劫持是常见的BGP异常情况，攻击者伪造路由信息，将流量重定向至恶意目的地。

重大事件：2008年巴基斯坦电信劫持YouTube流量，致其全球服务长时间不可用；2018年针对Amazon Route的BGP劫持，攻击者窃取大量加密货币。



BGP相关数据源对比

• RPKI特性

RPKI是基于公钥密码学的安全框架，用于解决**BGP路由起源认证**问题。网络运营商通过**证书和ROA记录**验证路由宣告合法性，**权威性高。**

• RIB表

RIB存储**路由信息表**，是BGP路由信息的**静态存储**机制。其覆盖范围受观测点地理位置限制，数据每5分钟更新一次，反映网络实际传播的路由信息，但准确性较低。

• IRR特性

IRR是分布式数据库，记录IP地址空间、AS号码持有信息及路由策略。但存在数据质量问题，**不同数据库数据可能不一致**，部分信息更新不及时，攻击者还可能注册虚假信息。

数据源	特点
RPKI	权威性高
RIB	数据量大
IRR	数据较为准确

2 2.研究现状与挑战

检测前缀劫持的方法

1

主动探测方法

方法：通过主动向目标前缀发送探测流量，检测其可达性和路径变化[1]。

优点：能**确定事件源头**和发展程度

缺点：BGP拓扑规模复杂，**开销大**，无法全量检测且缺乏实时性。

2

机器学习方法

方法：运用深度学习等算法对网络流量或路由数据建模[2]。

优点：能**自动学习**复杂网络行为模式，适应**网络变化**

缺点：大量标记数据，依赖数据质量和特征工程，且**部分模型难以解释**

3

路由逻辑方法

方法：基于**BGP路由信息**分析检测前缀劫持[3]。

优点：具有较好的**可解释性**和较快处理速度。

缺点：但**对数据可靠性依赖大**，数据源不准确或不完整会影响检测结果。

[1] Z. M. Mao et al. BGP Beacons. IMC 2003.
[2] Shone N et al. A deep learning approach to network intrusion detection. TETCI, 2018.
[3]Giotsas, V. et al. On the Incompleteness of BGP Data for AS-level Analysis. Sigcomm 2014.

研究挑战

目标

保证异常检测的**实时性**和**准确性**。

挑战1

数据源的P/O对信息不完整，检测结果准确率低。

挑战2

多源数据存在冲突，检测结果出现误报。

挑战3

路由信息频繁变化，数据库需要及时更新。

1.数据不完整性

不同数据源**覆盖范围不同**，数据不完整

2.数据多起源

一个前缀对应多个 AS 的情况，如不**同数据源的P/O对信息不一致**。

3.数据时变性

BGP前缀**会频繁进行宣告和撤销操作**，给检测带来挑战。

贡献一

对**三方路由数据**进行全面测量

贡献二

提取**时空稳定性特征**，并利用云模型进行数据筛选

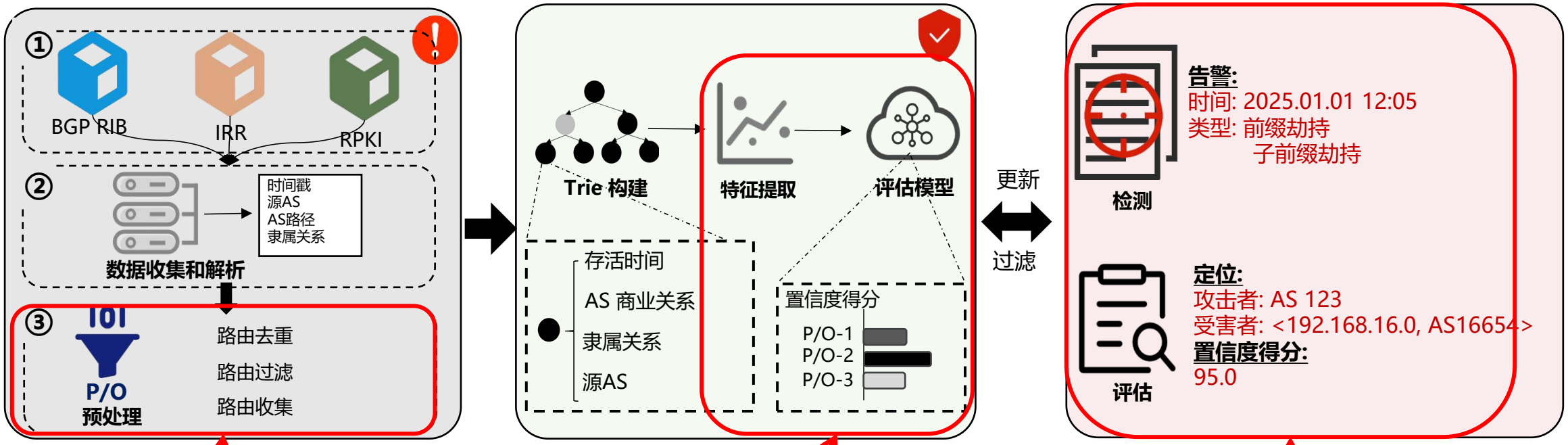
贡献三

双Trie结构进行异常检测，自动更新所提方法

3 3.方法设计与分析

南京邮电大学
Nanjing University of Posts and Telecommunications

总体框架与模块划分



贡献1　测量路由

贡献2　云模型与特征提取

贡献3　双Trie树检测机制

路由测量

- 数据源来自RouteViews RIB、RPKI ROA和IRR注册数据
- 根据固定窗口（5天），存储路由信息
- 数据结构为Trie树

三方数据源5天内宣布情况

{'RPKI': [1,0,0,0,0],'RIB': [1,1,1,1,1],'IRR':[1,0,0,0,0]

PeerAS:

{'34549':[1,1,1,1,1]，'58057':[1,1,1,1,1]，'37721':[1,1,1,1,1],'49432':[1,1,1,1],'38001':[1,1,1,1,1],'202365':[1,1,1,1,1],'22652':[1,1,1,1],44393':[1,1,1,1,1],'204092':[1,1,1,1,1],'15562':[1,1,1,1,1],'34880[1,1111],'55720':[1,1,1,1,1],，'50628':[1,1,1,1,1],'207564':[1,1,1,1,0],'45049':[1,1,1,1,1],'34927':[1,1,1,1,1],'328474':[1,1,1,1,0],'47251':[1,1,1,1,1],'44103':[1,1,1,1,1],'3333':[1,1,1,1,1],'59919':[1,1,1,1,1],'206499':[1,1,1,1,1],'852':[1,1,1,1,1]}}

5天内观测到该前缀的AS号

IPTrie节点中存储IP前缀的信息

路由测量

1.验证数据源多起源

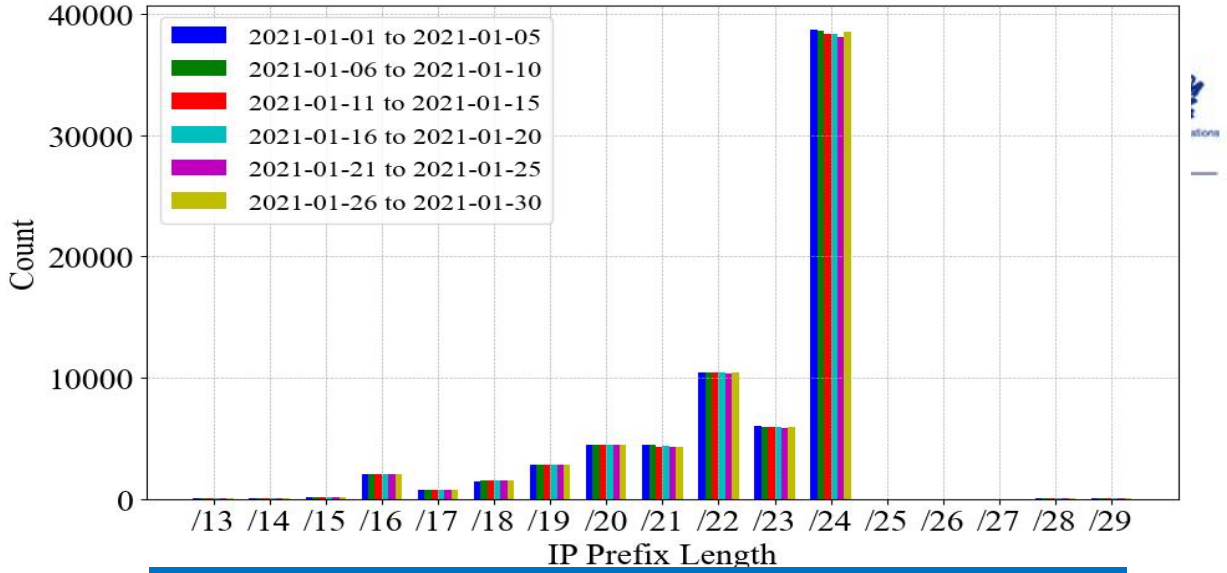
一个前缀对应多个 AS 的情况

2.验证数据源不完整性

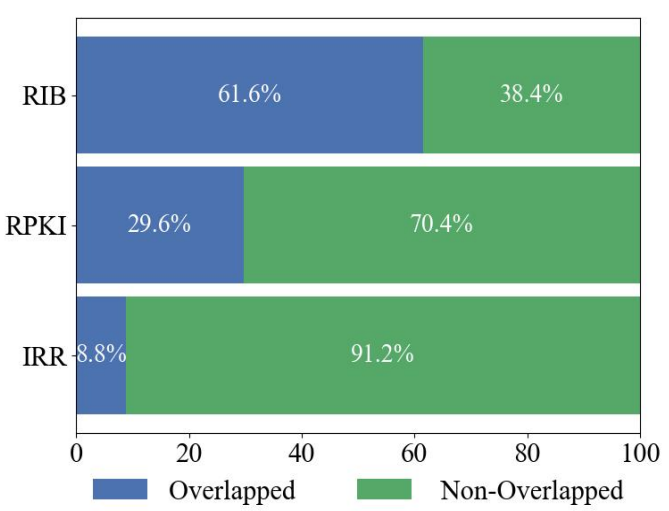
RIB覆盖广但不准确，IRR数据最少

3.验证数据源时变性

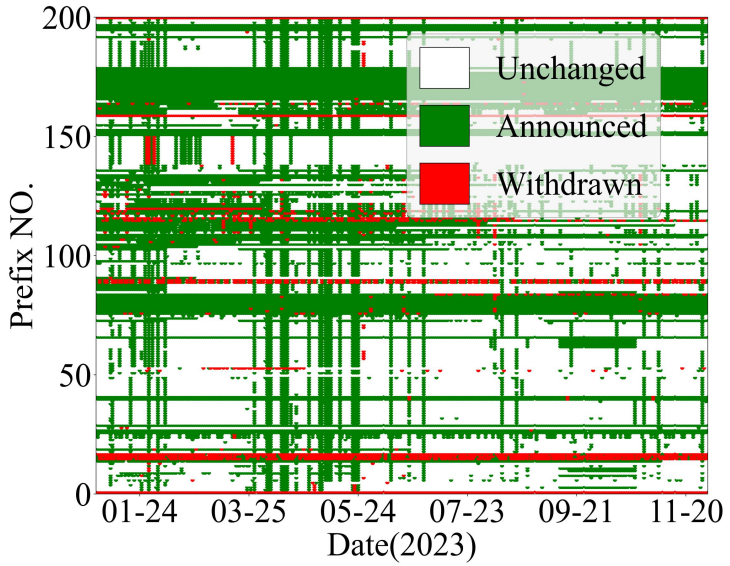
对200个前缀一年的UPDATES表检测发现其变化规律不同



MOAS前缀分布情况



RPKI、IRR、RIB三方数据重合的比例



2021年1月内路由更新情况

云模型与特征提取方法

1.提取空间一致性特征

$$S_p^t = h \cdot w_t^T$$

空间一致性表示P/O对被观测点观测到的次数

空间稳定性

2.提取时间持续性特征

$$S_p^s = g$$

时间持续性表示AS对P/O对5天内持续宣告的天数

时间稳定性

3.提取数据源隶属度特征

$$S_p^m = m \cdot w_d^T$$

对于来自不同数据源（RPKI、IRR、RIB）的P/O对

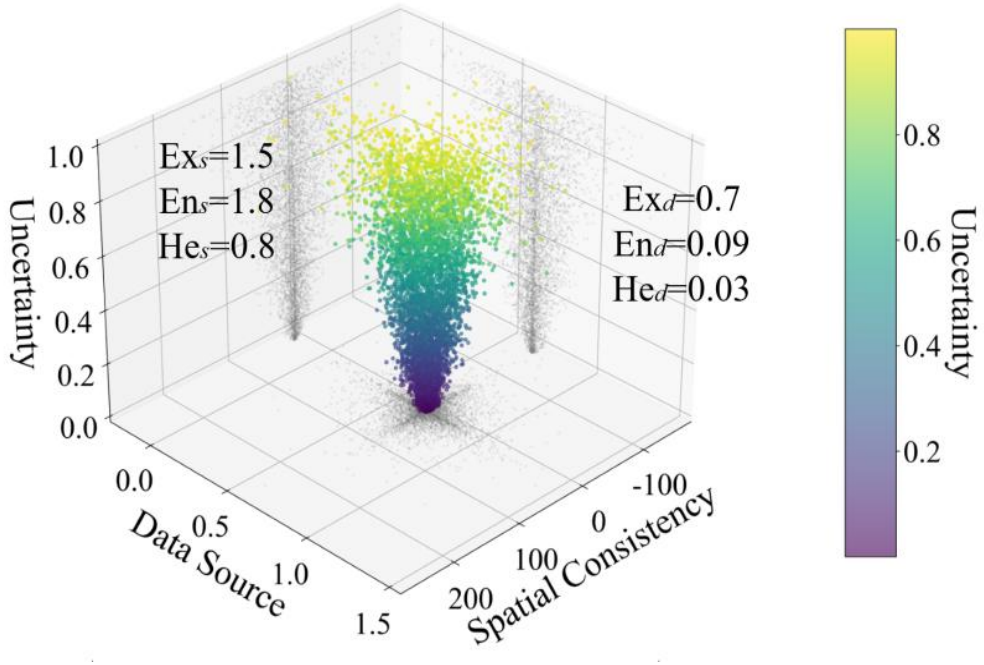
数据源稳定性

云模型与特征提取方法

引入云模型理论

- **期望** $\overline{S}_t = \frac{1}{a}\sum_{i=1}^{a} s_i^t$
- **熵** $\hat{En}_t = \sqrt{\frac{\pi}{2}} \times \frac{1}{a}\sum_{i=1}^{a}(s_i^t - \overline{S}^t)$
- **超熵** $\hat{He}_t = \sqrt{B_t^2 - \hat{En}_t^{\,2}}$
- **云滴不确定度** $\hat{y}_i = e^{-\frac{(s^t - Ex_t)^2}{2\cdot(En_t')^2} - \frac{(s^s - Ex_s)^2}{2\cdot(En_s')^2} - \frac{(s^m - Ex_m)^2}{2\cdot(En_m')^2}}$

进行定性概念和定量数值的双向转换。



数据源与空间一致性的不确定度3D云模型

双Trie树异常检测机制

1.双Trie树的功能分工

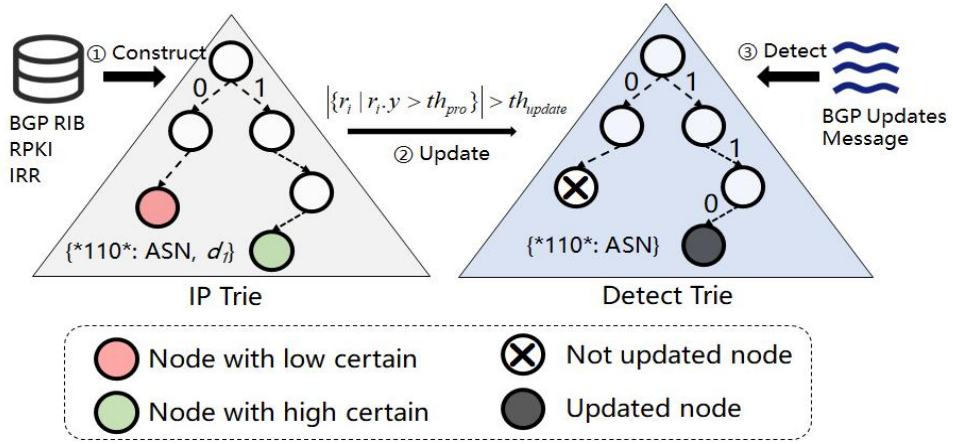
IPTrie用于**存储全量的前缀及**其相关信息，并实时**评估其不确定度。**

DetectTrie存储经过验证后的**高可信P/O对**，用于**快速判断**前缀**是否**为**合法**路由。

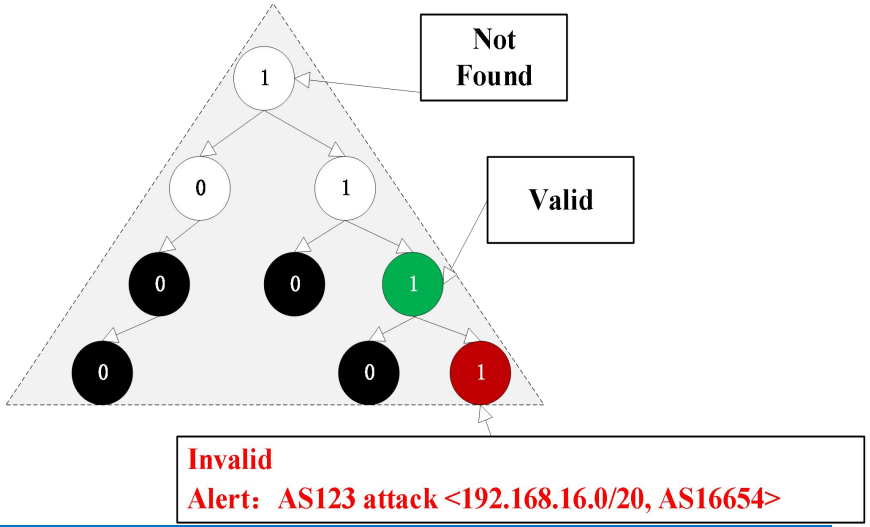
2.异常判定逻辑

合法：前缀与合法路由**符合匹配**条件

劫持：前缀不符合合法路由规则，如**覆盖关系不匹配或ASN不一致。**



一个时间窗口内路由信息处理



Trie树搜索IP前缀

4 4.实验结果与对比

实验环境与数据集

实验数据来源

实验数据主要来源于RouteViews RIB、RPKI ROA和IRR注册数据。

数据来源	数据时间范围	采集点	数据量
RIB表	2023-2024	RRC00 (路由表较为完整，对等邻居较多)	100GB
RPKI	2023-2024	RIPE、AFRINIC、APNIC、LACNIC、ARIN (五大互联网注册机构)	7GB
IRR	2023-2024	RADB	5.7GB

实验环境

服务器： AMD EPYC 7K62，48 核处理器，代码运行在 Linux 操作系统， 64GB 内存。

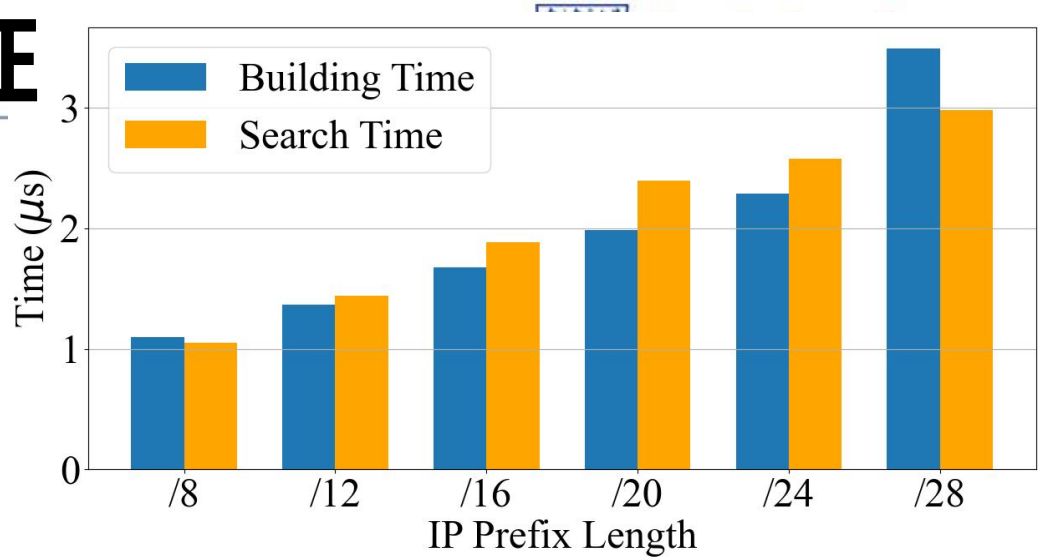
代码语言：Python

基准方法	分类
Artemis	基于路由逻辑
BEAM	基于机器学习
BGPviewer	基于机器学习
BGPvector	基于机器学习

实验1. 鲁棒性验证

1.实时性

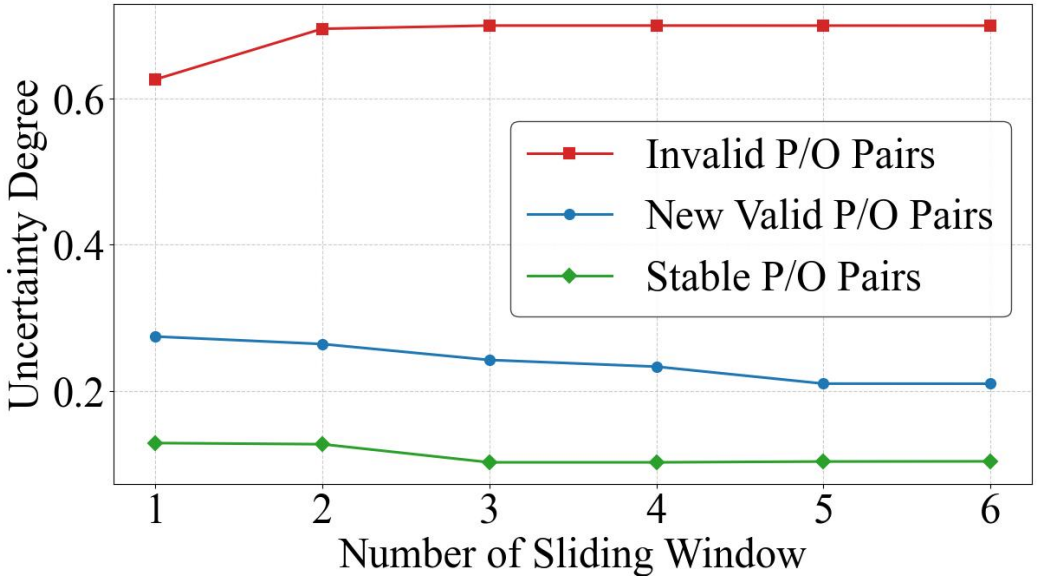
在不同的前缀长度下都能够进行实时地异常检测。



不同前缀长度下 Trie 树的构建和检测速度

2.稳定性

云模型能够较为有效的区分合法和不合法 P/O 对，且不会受到不合法 P/O 对的负面影响，体现出较好的稳定性。



P/O 对稳定性观测

实验2.基准方法对比

标注数据集

标注数据集包含10个真实劫持事件，选取依据为权威新闻标注。在**已知发生时间、攻击AS、受害AS以及前缀**的情况下，对每条BGP路由更新消息进行标注，保证了数据集的准确性。

开放数据集

在该数据集中，定义违背了 RPKI 中 P/O 对的报警为正确报警，其余报警为错误报警。

基于 2024 年 12 月下半月的未标注的 BGP 路由数据进行异常检测，将所提方法与其他方法进行对比。

标注数据集实验对比

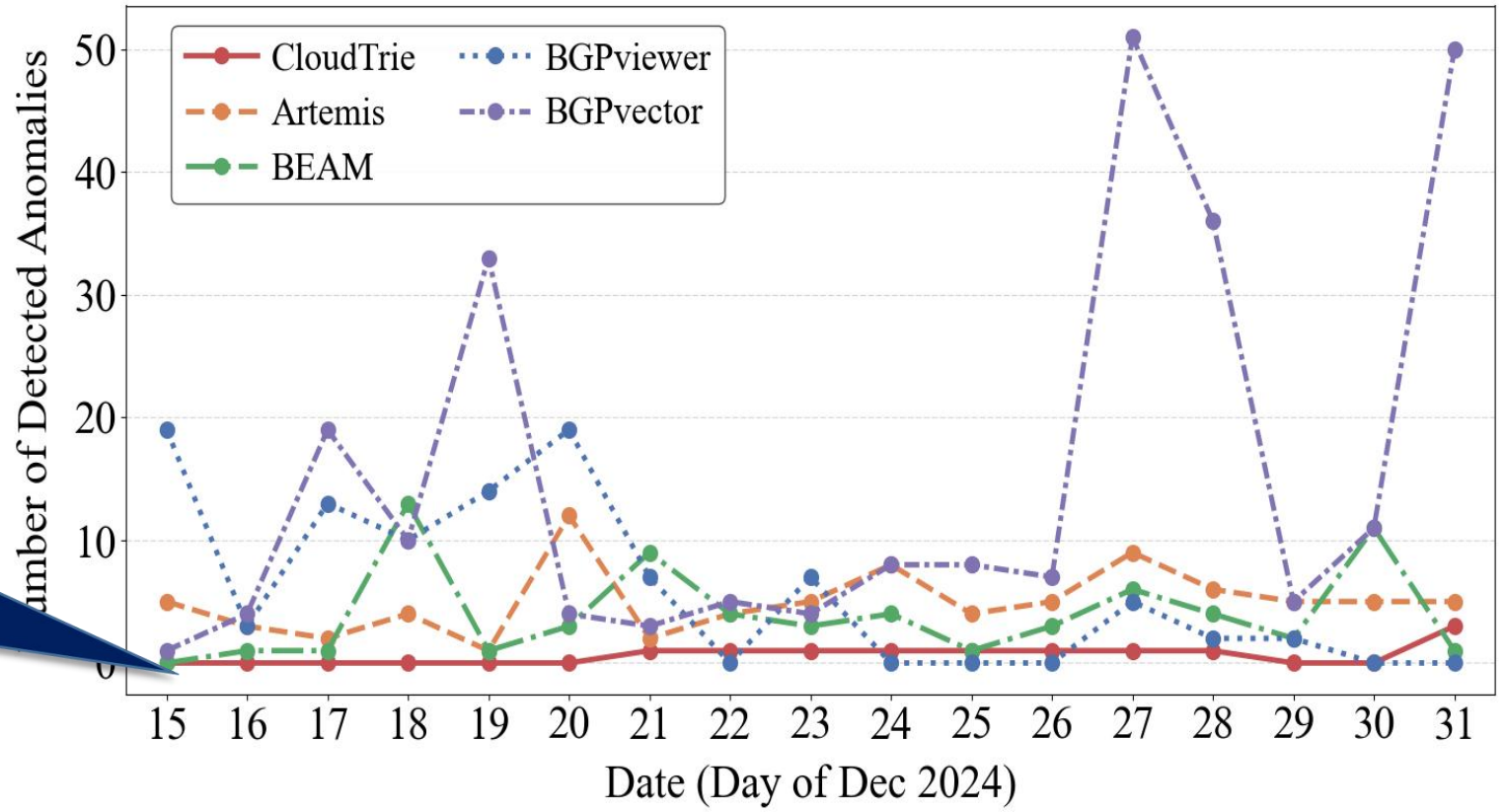
Dateset	# Alarms(#False Alarm)				
	CloudTrie	Artemis	BEAM	BGPVector	BGPviewer
Iran_hijack	16 (0)	16 (0)	255 (223)	25 (25)	4 (2)
PJSC_Rostelecom	8 (6)	0 (0)	1 (0)	1 (0)	1 (0)
DV_LINK	8 (0)	8 (0)			)
Bitcanal_Jingdong	7 (1)	7 (1)			)
Iran_Telegram	3 (0)	3 (0)			)
FIBRA_PLUS_TELECOM	4 (3)	[illegible]			)
Campana_MYTHIC	3 (2)	4 (4)	84 (76)	2 (2)	1 (0)
Nigeria_JSC_Ukraine	2 (1)	2 (1)	14 (1)	2 (2)	3 (2)
Russia_Ukraine	3 (3)	3 (3)	48 (47)	3 (3)	2 (1)
RU_AS_hijack_twitter	4 (0)	6 (2)	6 (6)	2 (2)	10 (9)
FP Rate	14.07%	26.43%	65.89%	24.92%	34.24%

CloudTrie检测的**准确率较高**，检测出10个事件中的**9个**异常事件。与其他方法相比，CloudTrie的**误报次数最少**，**假阳率最低**。

开放数据集实验对比

CloudTrie 具有**较低的报警次数**

表现最为稳定



在开放环境异常检测结果对比

5

5.总结

南京邮电大学
Nanjing University of Posts and Telecommunications

5.总结

多源数据测量

对RIB、RPKI和IRR三类数据源进行全面测量，揭示了BGP路由数据具有不完整性、MOAS现象普遍性以及时变性等特性。

云模型应用

基于云模型概念，设计实时IP路由可信度评估模型。综合时间、空间和数据隶属三个维度，提取稳定性特征评估路由信息准确性。

双Trie树机制

构造在线和离线两棵Trie树，克服BGP路由数据的时变性问题。基于固定窗口机制与双Trie树协同更新方法，使检测准确率提升10%，误报率降低10%。

- **现有成果：**

论文：面向不确定信源的前缀劫持检测方法，论文准备投稿《中国科学-信息科学》。

专利：一种面向不确定信源的前缀劫持检测方法，专利在受理中。

- **未来研究方向：**

未来，随着区块链、联邦学习等技术的引入，数据源的可靠性与隐私性会进一步提升，而增量学习与图神经网络的结合或将赋予模型更强的自适应能力。

信息安全专业　2021级本科B210314班　毕业论文答辩

恳请老师批评指正！

南京邮电大学

答辩人：B21031402谢睿熙

导　师：　吴争