轻量级比特币交易溯源机制

高峰[1)] 毛洪亮[2)] 吴震[2)] 沈蒙[1)] 祝烈煌[1)] 李艳东[1)]

[1)](北京理工大学 计算机学院 北京 100081)

[2)](国家计算机网络应急技术处理协调中心 北京 100029)

摘 要 比特币技术发展迅速，交易规模逐步攀升，引起国内外广泛关注．比特币技术具备去中心化和匿名化特征，使得比特币交易者的身份难以识别，为不法行为（如毒品交易、比特币勒索病毒等）提供了隐匿空间．本文提出了一种针对比特币交易的溯源机制，能够追踪交易信息在网络层的传播路径，从而将交易中的匿名比特币地址和发起交易节点的 IP 地址相关联．通过设计一种基于主动嗅探的邻居节点识别方法，溯源机制支持轻量级监测，而且相比现有溯源技术具有更好的实用性．我们针对比特币系统开发了溯源程序，从有效性、准确率、适用范围等方面对其进行测试与分析评估．实验结果表明，比特币网络中有 69.9%的服务器节点适用于这种溯源机制，能够获得召回率 50%、准确率 31.25%的溯源精度，优于现有的交易溯源方法，具有较强的实践意义和使用价值．

关键词 比特币；对等网络；溯源；区块链；反匿名

中图法分类号 TP391 DOI 号

Lightweight Transaction Tracing Technology for Bitcoin

Gao Feng[1], Mao Hong-liang[2], Wu Zhen[2], Shen Meng[1], Zhu Lie-huang[1], and Li Yan-dong[1]

[1]（School of Computer Science, Beijing Institute of Technology, Beijing 100081）

[2]（National Computer Network Emergency Response Technical Team/Coordination Center of China(CNCERT/CC), Beijing 100029）

Abstract The rapid development of Bitcoin technology and the growing scale of Bitcoin transactions have drawn wide attention at home and abroad. Whereas, Bitcoin is often used by terrorists and criminals attracted to the anonymity of the currency, such as all deals on Silk Road were made in Bitcoin. Therefore, it is essential to supervise Bitcoin and track the source transaction when necessary. However, as Bitcoin technology has the characteristics of de-centralization, traditional financial supervision means cannot provide effective supervision. Philip Koshy et al. found some special trading patterns for originating node by analyzing the propagation law of currency transactions in the network layer, but the proportion of special deals is less than 9%. Alex Biryukov et al. take advantages of the information of neighbor nodes of Bitcoin peer to locate the originating node. This approach improves fault tolerance and accuracy (experiment shows the accuracy of 11%), but requires constantly sending information to all nodes, which can cause network congestion. There are also some methods of transaction data analysis. However, they usually only get the relationship between the addresses, but cannot directly obtain the corresponding identity information of the address. Therefore, it is necessary to design a new transaction tracking mechanism for Bitcoin architecture, which can detect with fewer resources and has higher tracking accuracy than existing mechanisms.

本课题得到国家重点研发计划 (2016YFB0800301)、国家自然科学基金(61602039)、北京市自然科学基金(4164098)、广西云计算与大数据协同创新中心(YD16E14)、CCF-启明星辰科研基金项目资助．高峰，男，1987年生，博士研究生，计算机学会(CCF)会员（41307G），主要研究领域为区块链技术与应用、网络安全．E-mail: gaofengbit@foxmail.com．毛洪亮，男，1990年生，博士，工程师，主要研究领域区块链技术与应用．E-mail: mhl@cert.org.cn．吴震，男，1976年生，博士，正高级工程师，主要研究领域为网络安全、区块链技术与应用．E-mail: wuzhen76@126.com．沈蒙(通信作者)，男，1988年生，博士，讲师，计算机学会(CCF)会员（E200038784M），主要研究领域为云计算隐私保护、区块链技术与应用．E-mail: shenmeng@bit.edu.cn．祝烈煌，男，1976年生，博士，教授，主要研究领域为密码学、网络与信息安全．E-mail: liehuangz@bit.edu.cn．李艳东，男，1991年生，硕士研究生，主要研究领域为区块链技术与应用、云计算安全和数据隐私．E-mail: leeyandong@foxmail.com．

In this paper, we optimize the existing bitcoin transaction traceability mechanism and propose a new neighbor node identification scheme based on active sniffing. Our scheme supports lightweight transaction traceability and has better traceability than existing schemes. In addition, by designing a matching value optimization method based on multiple detection, the traceability mechanism can gradually improve the traceability results through continuous monitoring and improve the traceability accuracy. The main contributions of this paper include three parts: at first, we designs a practical Bitcoin transaction tracking mechanism that can track the transmission of bitcoin transactions under the public Bitcoin network and associate the anonymous bitcoin transaction with the IP address of the transaction originating node. Secondly, for the first time, we propose a new method for neighbor node detection based on active sniffing, which can infer the neighbor nodes of a specific node by sending probe information. This method can obtain the topology information of any server node with less resources. Finally, we developed a prototype system for traceability mechanisms and tested the efficiency and accuracy on public Bitcoin network. The experiment results demonstrate that 69.9% of the backbone nodes in the Bitcoin network are suitable for the proposed tracing mechanism, with traceability recall rate of 50% and accuracy of 31.25%, which is superior to the current tracing methods and of great importance in practice. The proposed traceability mechanism can trace the transactions in Bitcoin networks and identify the transactions created by specific server nodes which can help to track down criminals who maliciously use bitcoin technology to deter Bitcoin-based crimes. Moreover, the traceability mechanism of this article is also applicable to altcoin based on Bitcoin code and other digital currencies based on Blockchain technology, and has a wide range of application scenarios.

Key words Bitcoin; peer-to-peer networking; tracing; Blockchain; de-anonymization

1 引言

比特币发展迅速，交易规模逐步攀升，在国内外引起极大关注[1-4]．截至 2017 年 10 月 6 日比特币单价超过 4370 美元，市值突破 725 亿美元①．比特币系统由于具有匿名化和去中心化的特征，已经成为洗钱、地下黑市等非法交易活动的温床[5]．例如，2017 年 5 月导致 100 多个国家和地区超过 10 万台电脑被感染的勒索病毒“WannaCry”就是利用比特币获取赎金，由于比特币交易难以追踪，各国执法机构都未能抓获病毒制造者．因此非常有必要研究针对比特币交易的溯源机制，即通过追踪比特币交易的传播路径，找到创建比特币交易的服务器节点，从而推测出恶意交易者的身份信息，遏制基于比特币的犯罪行为．

比特币是一种去中心化的数字货币技术，相对于传统金融系统，比特币交易具有较强的反溯源能力．比特币数字货币系统的特点包括：

1）地址匿名性．比特币地址是用户参与比特币交易时使用的账号．地址由用户自行创建，与身份信息无关，创建和使用过程不需要第三方参与．

2）交易分散性．比特币系统支持用户为每次交易生成不同的地址．因此用户的交易信息将分散在不同的匿名地址中，很难通过分析交易记录推测用户的身份特征．

3）网络结构去中心化．比特币系统采用 P2P（对等网络）协议组网，不存在中心节点．因此很难通过监测单一服务器追踪交易信息在网络中的传播路径．

由于以上特点，传统的交易溯源技术在比特币中难以适用．目前国内外针对比特币交易溯源的研究相对较少，现有研究主要分为两类：网络层溯源技术和交易数据分析技术．

网络层溯源技术是指通过搜集比特币网络层传输的信息，分析出比特币交易在网络中的传播路径，从而追踪产生该交易的服务器 IP 信息．此技术能够直接将匿名交易和交易始发节点的 IP 地址关联，实现溯源目的．然而，现有的网络层溯源技术准确率较低，而且通常需要较多的计算资源和存储资源，实用性较差．

交易分析技术是指通过分析比特币交易记录，发现不同交易地址之间的关联关系，从而推测出匿名用户的交易规律．这种方法只能分析出匿名用户的交易特征，不能直接获得交易者的身份信息．而

① CoinMKTCap, CrytoCurrency Market Capitalizations, https://coinmarketcap.com/ 2017-10-6

且比特币用户可以采用一次性地址策略、混币策略[6-10]等方法增加交易分析的难度.

针对以上问题，本文对现有的比特币交易溯源机制进行优化，提出一种新型的基于主动嗅探的邻居节点识别方法，支持轻量级交易溯源. 通过设计基于多次检测的匹配值优化方法，溯源机制能够通过持续监测逐渐优化溯源结果，提高溯源准确率.

本文的主要贡献如下：

1) 设计一种实用的比特币交易溯源机制，能够在真实环境下追踪比特币交易的传播路径，将交易中的匿名比特币地址和交易始发节点的 IP 地址相关联.

2) 首次提出一种基于主动嗅探的邻居节点识别方法，能够通过对特定节点发送探测信息，推测出特定节点的邻居节点，实现轻量级交易溯源.

3) 实现了交易溯源系统，并在真实的比特币网络中对方案进行验证测试. 实验结果表明，比特币网络中有 69.9%的服务器节点适用于这种溯源机制，能够获得召回率 50%、准确率 31.25%的溯源精度，优于现有的交易溯源方法.

本文第 2 节介绍比特币交易溯源机制涉及的背景知识；第 3 节介绍现有的交易溯源技术；第 4 节详细介绍溯源机制的系统架构和技术细节；第 5 节介绍溯源机制中的关键技术；第 6 节从有效性、准确率和适用范围 3 个方面对溯源技术进行测试和评估；第 7 节介绍溯源技术相关的研究方向；第 8 节进行总结.

2 问题定义

本节主要介绍比特币交易溯源机制涉及的背景知识. 其中，2.1 节介绍比特币交易溯源概念，2.2 节介绍溯源节点选择策略，2.3 节介绍溯源机制技术难点.

2.1 比特币交易溯源

比特币技术是一种去中心化数字货币技术，其中的代币被称为比特币[11-13]. 用户可以在全球任意位置的服务器上通过创建比特币交易和其他用户进行比特币代币的双向交易. 由于交易过程不需要第三方参与，而且交易双方使用的地址具有匿名性，因此很难发现比特币交易发起者的真实身份.

交易溯源技术希望追踪比特币交易在网络中的传播路径，最终发现交易的始发节点，即第一个在比特币网络转发此交易的服务器节点. 一旦将比特币交易和始发节点的 IP 地址关联，就可以将交易中的匿名账号和用户身份关联，有助于识别恶意交易者的身份信息、分析比特币资金流向.

2.2 溯源节点选择

比特币网络是一个由全球各地的服务器组成的对等网络，网络中的每个节点地位平等，不存在中心节点. 但是根据节点提供的服务不同，比特币节点可以分为服务器节点和客户端节点.

定义 1. 服务器节点. 既能实现基本交易功能，又能够为其他节点提供信息中转、交易验证等服务的比特币服务器.

定义 2. 客户端节点. 只能实现基本交易功能，不能对外提供服务的比特币服务器.

比特币网络中的服务器节点和客户端节点都可以创建比特币交易. 区别是：服务器节点是比特币网络中的骨干节点，需要对外提供信息中转、交易验证等服务，通常具备独立的公网 IP 并维护全部交易数据；客户端节点依附于特定的服务器节点，不需要对外提供服务，通常没有固定 IP，而且在线时间较短.

根据以上特点，可以分析出针对服务器节点进行交易溯源具有更好的效果. 原因包括：

第一，服务器节点本身也创建交易，而且由于服务器节点能够更好的保证交易可靠性，重要的交易通常采用服务器节点创建①. 服务器节点通常具有较为稳定的 IP 地址，有利于对溯源结果进行优化. 服务器节点是比特币网络中的骨干节点，所有交易信息都要通过各个服务器节点进行转发.

第二，针对服务器节点的溯源可以为追踪客户端节点奠定基础. 客户端创建的交易将首先通过特定的服务器节点进入比特币网络，如果能够追踪到交易首次进入比特币网络的服务器节点，将显著缩小客户端节点的过滤范围，有利于找到真实的客户端节点. 7.2 小节介绍一种从始发服务器节点找到创建交易的客户端节点的方法.

因此本文针对比特币服务器节点研究交易溯源机制，希望找到比特币交易的始发服务器节点.

2.3 溯源机制技术难点

相对于传统中心化系统，比特币技术支持匿名

① Kyle Torpey. You Really Should Run a Bitcoin Full Node: Here's Why. https://bitcoinmagazine.com/articles/you-really-should-run-full-bitcoin-node-heres-why/

交易，具有更好的隐私保护效果，但是比特币系统依然存在隐私泄露的缺陷．祝烈煌等人[14]分析了以比特币为代表的区块链技术在隐私保护方面存在的优势和缺陷，为本文的研究提供了理论支撑．Androulaki 等人[15]和 Monaco[16]通过实验证明结合线下数据或交易模式信息能够将用户身份和比特币地址相关联．

在比特币协议中，为了维持去中心化网络的稳定运行，比特币服务器节点通常需要对外提供服务，例如帮助客户端节点连接比特币网络，中转交易信息等．因此比特币服务器节点通常允许任意节点发起的连接请求，并会向这些连接节点广播交易信息．这种特征使得分析人员只需要利用比特币程序中的连接指令就可以对全球任意比特币服务器节点开展监测，搜集节点转发的交易信息，从而推测交易信息在网络中的传播路径，甚至找到交易的始发节点，实现针对匿名交易的溯源．

为了解决上述问题，比特币的开发者设计了黑名单机制和延迟转发机制来增加溯源难度．黑名单机制是指比特币节点会对网络中其他节点的行为进行评估，如果节点危害网络运行就会将其列入黑名单，拒绝此类节点的连接请求．Gervais 等人[17]详细介绍了比特币的黑名单机制．Huang 等人[18]提出一种基于行为模式聚类的恶意节点检测方法，能够快速定位、消除恶意节点；延迟转发机制是指不同的比特币节点在转发交易时会随机采用不同的延迟时间，模糊始发节点和非始发节点转发交易的区别．Biryukov 等人[19]详细介绍比特币的延迟转发机制．

针对上述背景，可以分析出比特币交易溯源机制的核心难点是满足以下目标：

1）设计轻量级数据搜集机制，能够有效搜集溯源所需的信息，获得比特币服务器节点网络拓扑情况．同时尽量减少对比特币网络正常运行的干扰．

2）设计具有通用性的交易匹配算法，能够根据搜集的数据筛选出由待监测节点始发的交易信息，能够应对延迟转发机制带来的干扰，对不同服务器节点具有通用性．

3 相关工作

针对比特币交易的溯源研究相对较少，现有的方法分为两种：交易关联分析和网络层溯源．

交易关联分析技术是通过分析比特币账本中的交易记录推测不同交易之间的关系，例如交易规律，资金流向等．Reid 和 Harrigan [20]针对维基解密公布的比特币地址进行数据分析，统计出维基解密网站公布的比特币地址的资金余额、资金来源和资金流向．Liao 等人[21]通过分析比特币交易数据，对勒索软件 CryptoLocker 的勒索过程进行了分析，找到了 968 个属于勒索组织的地址，鉴定出价值 1128.40 个比特币的赎金交易．Meiklejohn 等人[22]使用启发式的聚类分析技术识别出隶属于丝绸之路网站的多个比特币地址．Zhao[23]提出一种针对比特币交易数据的聚类过程，针对比特币全局账本中 35,587,286 个地址进行分析，得到 13,062,822 个不同的用户集合．基于交易关联分析的方法通常只能获得地址之间的关系，而不能直接获得用户身份信息．而且一旦用户会采用一次性地址策略或交易混淆策略，这种方法的准确度将受到显著影响．

网络层溯源技术是通过分析比特币网络层传输的交易信息，发现特定交易在比特币网络中的传播路径，进而推测交易的始发节点．Koshy 等人[24]通过分析比特币交易在网络层的传播规律，发现可以利用特殊交易模式寻找始发节点．例如，大部分正常交易会被多个节点转发一次，而交易格式存在问题的交易只会被始发节点转发一次，因此可以利用这种特征识别特殊交易的始发节点．但是由于特殊交易的比例较小（论文试验中特殊交易的比例低于 9%），此方法效果有限．Kaminsky①在 2011 年的黑帽大会上提出，“第一个告诉你交易的节点可能就是交易的始发节点”．分析人员只需尽可能多的连接比特币服务器节点并记录从不同节点转发的交易信息，然后即可判定首先转发信息到达探针的节点就是始发节点．这种方法只依赖首发节点作为判断特征，准确率较低．Biryukov 等人[19,25]提出基于邻居节点的交易溯源机制，通过将邻居节点作为判断依据，能够提高溯源准确率．但是方案需要持续向比特币网络中的所有节点发送信息，有可能对比特币网络造成严重干扰，实用性较差．

① Kaminsky D. Black Ops of TCP/IP 2011. https://dankaminsky.com/2011/08/05/bo2k11/

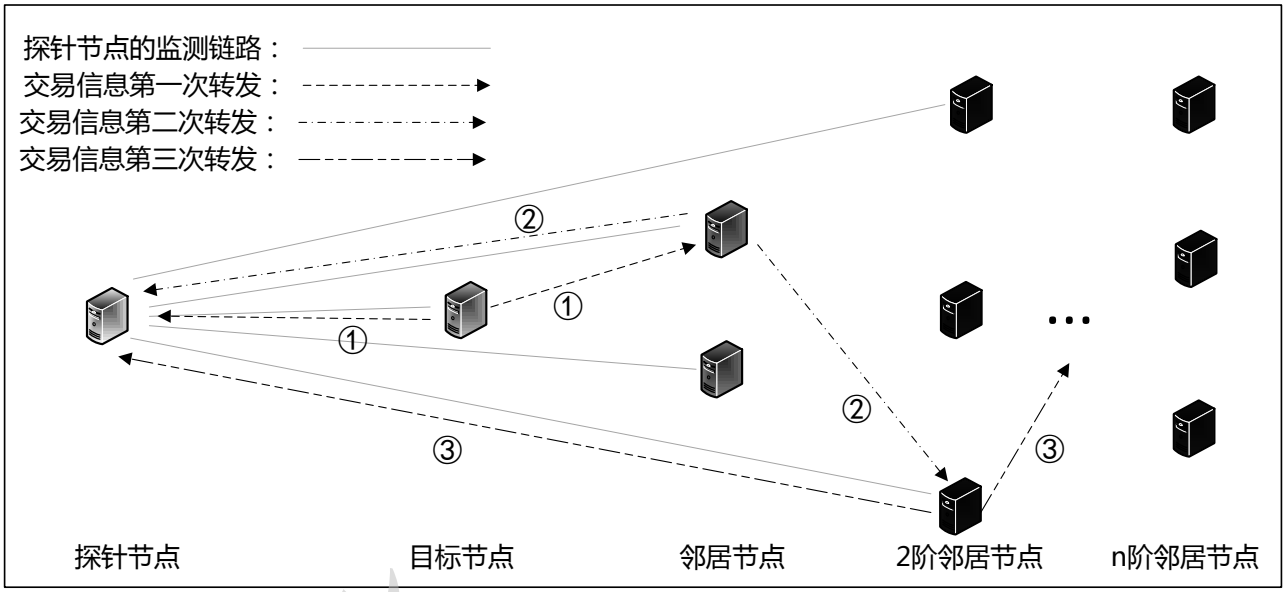


图 1 比特币交易溯源机制系统架构

现有的比特币交易溯源技术准确率较低，而且通常需要较多的计算资源和存储资源，很难在真实的环境中使用．我们希望设计一种轻量级的交易溯源机制，能够利用较少资源对比特币服务器节点进行交易溯源，并且具有较高的准确率和适用范围．

4 轻量级交易溯源机制

本节主要介绍比特币交易溯源机制．其中，4.1 节介绍溯源机制的系统架构，4.2 节介绍溯源流程．

4.1 系统架构

为了对比特币交易实施有效溯源，我们在比特币网络中部署探针节点．探针节点可以搜集网络层中的传输信息，分析交易的传播路径，推测交易的始发节点．一旦找到交易的始发节点，就可以将交易中的比特币地址和始发节点的 IP 地址相关联．

比特币网络中大约有 7500 个服务器节点，每天约产生 28.8 万条交易记录（2017 年 10 月份数据[①]）．对全部节点和全部交易进行匹配需要巨大的计算资源和存储资源．因此我们的溯源机制是针对特定服务器节点进行交易溯源，即识别出由特定服务器节点始发的交易信息．这种溯源机制支持通过增加硬件扩大监测范围，具有较好的可扩展性．

图 1 展示比特币交易溯源机制的系统架构．图中的目标节点是指被监测的服务器节点．邻居节点是指目标节点在比特币网络中相邻的节点，比特币网络中，每个节点维持不超过 125 个邻居节点．节点通过和自己的邻居节点交换信息从而与比特币网络保持同步．2 阶邻居节点是指邻居节点的邻居节点．依次类推，从目标节点的视角出发，整个比特币网络可以被分为邻居节点、2 阶邻居节点、…、n 阶邻居节点．

图 1 中的探针节点是指部署了溯源程序的节点．探针节点负责从比特币网络传输的交易信息中筛选出由目标节点始发的交易信息．图中的灰色线条代表探针节点和其他节点建立的连接．探针节点通过向其他节点发送连接请求来建立连接．

探针节点围绕目标节点建立监测网络，将持续搜集比特币网络中的交易信息．比特币交易在网络层采用泛洪的方式传输，即首先将交易转发给邻居节点，然后邻居节点继续将交易转发给自己的邻居节点．以此类推，直到将交易传到比特币网络中的所有节点．因此，针对一个比特币交易，探针节点将从不同节点收到多个版本．虽然每个版本的交易内容相同，但是由于不同节点转发的交易时间不同（越靠近始发节点的节点转发交易越早），每个版本的交易信息到达探针节点的时间不同．因此，探针节点可以根据时间排序推测交易信息在网络中的传播路径．

例如，图 1 中目标节点始发的交易信息首先转发给探针节点和右上侧的邻居节点，这是交易信息在网络中的第一次转发．然后，邻居节点将交易信息转发给探针节点和下侧的 2 阶邻居节点，这是第

① Bitnodes, Global bitcoin nodes distribution, https://bitnodes.21.co/

二次转发．最后，2 阶邻居节点将信息转发给探针节点和右侧的节点，这是第三次转发．

在不考虑网络延迟等干扰因素的条件下，这 3 次转发的交易信息将依次到达探针节点．因此，探针节点可以分析出这条交易信息的传播路径为（目标节点，邻居节点，2 阶邻居节点）．传播路径与目标节点的网络拓扑相同，因此可以猜测这条交易是由目标节点始发．

上述例子中，我们假定已知目标节点的 IP 地址和邻居节点的 IP 地址．在实际溯源中，这些信息是未知的．溯源系统需要首先选定目标节点并获得其 IP 地址．然后需要识别出目标节点的邻居节点．

此外，在实际运行中不同节点和探针节点之间的网络延迟不同．这有可能导致不同节点转发交易的时间顺序和交易信息到达探针节点的顺序不同，干扰推测结果．因此，交易溯源机制需要研究如何降低干扰因素的影响，提高推测准确率．

4.2 交易溯源流程

比特币交易溯源机制的核心思路是利用探针节点搜集比特币网络中传输的交易信息，然后通过分析信息传播规律推测交易信息的始发节点．

溯源机制的流程如图 2 所示．步骤如下：

1) 使用探针节点搜集比特币网络中服务器节点的信息，确定溯源目标．

2) 根据目标节点的 IP 地址建立监测网络．

3) 使用探针搜集比特币网络中的交易信息，同时采用主动嗅探算法推测目标节点的邻居节点．

4) 根据搜集的交易信息和推测的邻居节点信息为每一条交易计算匹配值．

5) 输出匹配值超过阈值的交易信息．

下面具体介绍每一步的操作过程．

4.2.1 确定溯源目标

溯源机制首先需要选定待监测的目标节点，并获得节点的 IP 地址．

探针节点可以利用比特币技术中的节点发现机制搜集比特币网络中服务器节点的信息．首先，探针节点将连接比特币种子节点（硬编码在比特币客户端程序中的节点 IP 地址），然后向种子节点索取邻居节点信息．通过递归索取，探针节点最终将获得比特币网络中大部分在线服务器节点的信息．这些信息的格式如表 1 所示．

探针节点搜集的服务器节点信息中包含 IP 地址、地理位置、组织名称和经纬度等信息．分析人员可以从中选择感兴趣的目标节点进行交易溯源．

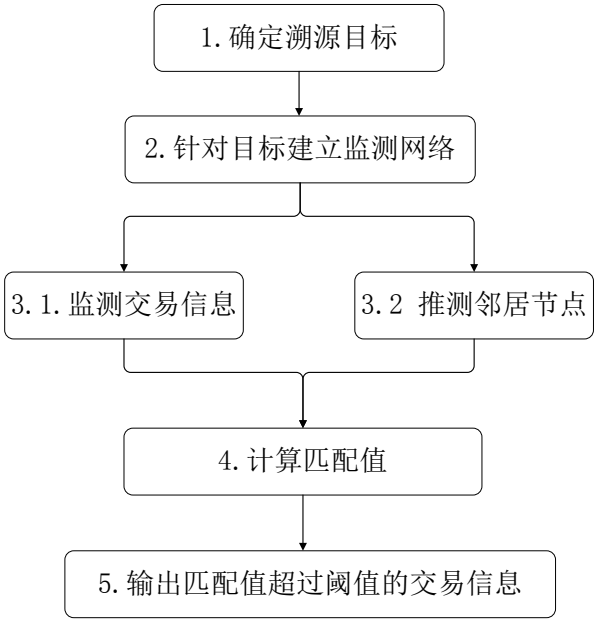


图 2 比特币交易溯源机制流程

4.2.2 建立监测网络

为了分析比特币网络中交易的传播路径，我们基于比特币开源代码[①]开发了探针程序，能够对服务器节点进行监测，搜集节点转发的交易信息．

探针节点对服务器节点的监测是通过模拟比特币节点之间的连接实现的．比特币节点通过保存邻居节点的 IP 地址和端口来实现连接，节点之间相互转发交易信息和区块信息，还会通过 Ping 等方式监测节点在线状态．比特币协议中为了降低节点的消耗，规定邻居节点的数量上限是 125 个．在我们的探针程序中，我们设置探针程序只接收信息而不转发信息，从而降低了节点的通信代价，因此探针程序可保持的邻居数量显著上升．在实验中探针节点可以连接 7000 个邻居节点，即通过 1 台探针节点就可以对几乎所有的比特币服务器节点开展监测．

监测网络的组建步骤如下：

1） 探针节点向目标节点发送连接请求．

2） 连接成功后，向目标节点发送“GETADDR”指令，索取目标节点存储的节点地址列表．

3） 探针节点对获得的地址列表中的所有节点重复步骤 1 和步骤 2．

通过上述步骤探针节点将围绕目标节点建立监测网络，能够搜集所有连接节点转发的交易信息．

4.2.3 推测邻居节点

邻居节点是比特币服务器节点与外界通信的

① Bitcoin. Bitcoin Core integration/staging tree. https://github.com/bitcoin/bitcoin

表 1 比特币服务器节点信息

IP	地理位置	组织名称	经纬度
83.xx.xx.217	荷兰 阿姆斯特丹	Ziggo	52.0452, 4.6556
59.xx.xx.134	中国 杭州	Hangzhou Alibaba Advertising Co.,Ltd.	30.2936, 120.1614
52.xx.xx.40	美国 洛杉矶	Amazon.com, Inc.	45.8696, -119.688
104.xx.xx.235	美国 洛杉矶	Google Inc.	37.4192, -122.0574

接口．服务器节点创建的交易将首先发送给自己的邻居节点，因此准确推测出邻居节点有利于分析交易与服务器节点的匹配关系．

推测邻居节点主要依靠邻居节点在由服务器节点始发的交易中的排序特征．如果已知服务器节点的始发交易，则可以推测这些交易中排序靠前的节点是服务器节点的邻居节点．

但是在实际的溯源环境中，很难区分非可控服务器节点的始发交易和转发交易，因此邻居节点推测是溯源技术中主要的技术难点．Biryukov 等人[19]提出了一种通过搜集地址信息判断邻居节点的机制，但是这种机制需要使用探针程序和每个服务器节点建立 50 个以上的连接，而且需要持续向所有节点转发信息，会对比特币网络造成巨大压力，容易引起网络故障．

为了实现轻量级溯源，我们设计了一种主动嗅探技术，可以利用较少资源推测目标节点的邻居节点．主动嗅探技术将在 5.2 小节介绍．

4.2.4 监测交易信息

探针节点启动后将记录到达探针节点的每一条交易信息，包括交易的哈希值、来源 IP 和到达时间．探针节点将搜集到的交易信息按照交易哈希分类，并对每一类交易按照交易到达探针的时间进行排序，然后从排序中挑选出目标节点和邻居节点的排序信息组成交易传播路径．

交易传播路径是一个 10 元组，包括[TXID，R_t，R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8]．其中 TXID 代表交易哈希，R_t 代表目标节点转发的交易被探针检测到的时间排序．R_1-R_8 代表 8 个邻居节点的排序．例如，当目标节点转发的交易信息第一个到达探针节点时，则 R_t=1．当搜集到的交易中没有找到邻居节点转发的交易时，设定排序值为 MAX，代表无穷大．

4.2.5 计算匹配值

匹配值反映了交易传播路径和目标节点网络拓扑的吻合度．匹配值越高，交易信息越有可能是由目标节点始发．理想情况交易的始发节点首先转发交易，始发节点的邻居节点将第二批转发交易，其他 n 阶邻居节点依次推后．因此，可以依据特定交易的传播路径是否满足上述规则来判断特定交易是否由目标节点始发．

实际环境中，由于比特币程序采用随机延迟的干扰技术，以及存在网络延迟等干扰因素，排序可能与理想情况不同．为了减少干扰，我们设计匹配值（G）代表交易和目标节点的关联关系．匹配值计算方法如公式（1）所示．

$$G = (\frac{1.5}{R_t} + \sum_{i=1}^{n}(\frac{1}{R_i}) / (\frac{8389}{2520})) \quad (1)$$

G 代表匹配值．R_t 指目标节点的排序．R_1-R_n 指 n 个邻居节点的排序，n 的取值等于找到的邻居节点个数．排序越大对匹配值的作用越小，为简化计算当排序大于 100 时默认分数值为 0．目标节点转发的交易只需要 1 次转发即可到达探针，而邻居节点需要经过 2 次转发（转发路径：目标节点-邻居节点-探针节点），受网络延迟的干扰较大，因此公式中目标节点的系数为 1.5，邻居节点的系数为 1．8389/2520 是理想情况下得分的最大值，即目标节点第一个到达，8 个邻居节点第一批到达的情况：1.5+1/2+1/3+1/4+1/5+1/6+1/7+1/8+1/9=8389/2520．通过除以最大值，可以保证 G 值的取值为（0,1）．

根据公式（1）计算的匹配值 G 反映了特定交易和目标节点的关联关系．在实际运行时，由于干扰条件的影响，单次计算的匹配值可能出现异常值．为了提高推测的准确率，可以通过多次监测优化匹配值，剔除异常数据的干扰．我们设计公式（2）计算优化后的匹配值．

$$G_{opt} = \frac{\sum_{i=1}^{n}(G_i)}{\mathrm{n}} \quad (2)$$

其中 G_{opt} 代表优化后的匹配值．n 代表匹配次数，G_i 代表第 i 次的匹配值．

计算优化后的匹配值包括两种情况：

1）多探针优化．使用 x 个探针同时监测，针对每个交易都将产生 x 个匹配值．

2）多次交易优化．具有相同输入地址的交易很可能是由同一个节点始发的，因此可以假设具有相同输入地址的 y 笔交易是由相同服务器节点创建的交易，得到 y 个匹配值．

将上述两种情况的得到的(x+y)个匹配值代入公式（2）计算出的结果就是优化匹配值．优化匹配值能够更精确的反映交易和目标节点的关系，减少因网络延迟等干扰因素带来的误差．

4.2.6 输出疑似交易

G_{opt} 值反映了交易和目标节点的关联关系．G_{opt} 值越高，说明此交易的传播路径和目标节点的网络拓扑越吻合，越有可能是目标节点的始发交易．为了从大量交易中筛选出由目标节点始发的交易，我们设计阈值作为评判标准，当 G_{opt} 的值大于阈值时认为此交易是疑似交易，即由溯源系统推测的可能由目标节点始发的交易．

阈值的取值通过实验获得，即挑选能够准确区分始发交易和非始发交易的值作为阈值．在挑选阈值时需要考虑节点的服务器此外，阈值的取值直接影响溯源精度．阈值较大时捕获的交易较少，准确率较高；阈值较小时捕获的交易较多，准确率较低．因此，阈值的选择需要综合考虑多种因素，5.3 小节介绍阈值取值方法．

5 关键技术

5.1 交易排序准确率分析

网络层交易溯源的本质是根据不同节点发送交易到达探针的时间排序推测交易的传播路径．理想条件下，始发节点的交易最早到达探针，邻居节点的交易第 2 批到达探针，之后的 n 阶邻居节点到达探针的排序将随距离增加．实际环境中不同节点转发交易到达探针的时间排序受网络延迟、延迟转发策略等多种因素的影响，有可能出现距离远的节点转发的交易提前到达的情况．为了准确分析交易排序与节点网络拓扑的吻合度，我们将综合考虑多种影响因素，计算交易排序准确率．表 2 分析不同节点发送的交易到达探针的排序的影响条件．

比特币服务器节点连接到比特币网络后，会持续向邻居节点转发交易信息．转发策略分为两种：

1）如果交易是由服务器节点创建的交易（即始发交易），则节点首先判断交易哈希的末尾是否是“00”（概率为 1/4）．如果是，则将交易信息立即发送给邻居节点，然后每隔 100 毫秒从邻居节点中抽取一个节点进行转发；否则直接每隔 100 毫秒从邻居节点中抽取一个节点进行转发．已经发送的节点不再转发，直到所有邻居节点都转发完毕．

表 2 交易排序相关参数

参数	参数特征
T	目标节点发送的交易到达探针的时间
T_n	非目标节点发送的交易到达探针的时间.n 代表交易中转的次数，n 从 1 开始
N_m	目标节点邻居节点的个数（1<= N_m <=125）
N_{fm}	非目标节点邻居节点的个数（1<= N_{fm} <=125）
S_{mt}	目标节点转发交易给探针的次序（0<= S_{mt} <= N_m −1）
S_{mft}	目标节点转发交易给非探针的次序（0<= S_{mft} <= N_{fm} −1）
S_{fmt}	非目标节点转发交易给探针的次序（0<= S_{fmt} <= N_{fm} −1）
S_{fmft}	非目标节点转发交易给非探针的次序（0<= S_{fmft} <= N_{fm} −1）
t_1	从节点转发交易信息到探针收到交易信息的时间间隔为 t_1．不同节点的 t_1 根据网络延迟各不相同，在估算时采用平均值．
t_2	从节点收到交易信息到节点开始中转交易的时间间隔为 t_2．不同节点的 t_2 根据节点性、网络延迟各不相同，在估算时采用平均值．

2）如果交易是服务器节点从其他节点接收的交易，则服务器节点经过一系列验证后，将把合法交易转发给所有邻居节点．转发规则是每隔 100 毫秒从邻居节点中抽取一个节点进行转发．

针对非可控服务器节点，很难直接区分始发交易和非始发交易，本文中主要采用主动嗅探算法，利用可控节点单点连接服务器节点的方式，使服务器节点被动产生“始发”交易，这种情况下的交易只会采用第（2）种转发策略，因此我们的估算值只针对 3/4 的交易有效，存在一定误差．

根据上述规则，我们可以对目标节点和非目标节点（包括邻居节点以及其他 n 阶邻居节点）发送的交易到达探针的时间进行估算．公式为：

$$T=100* S_{mt} + t_1 \quad (3)$$

$$T_1=100* S_{mft} + t_1+ t_2+100* S_{fmt} + t_1 \quad (4)$$

$$T_n=100* S_{mft} + t_1+ t_2+(100* S_{fmft} + t_1+ t_2)*(n-1)+100* S_{fmt} + t_1 \quad (5)$$

公式（3）代表从目标节点发送的交易到达探针的时间．由于目标节点直接将交易发送给探针，因此不需要额外的中转，时间 T 只包括比特币系统的随机延迟时间（100* S_{mt}）和目标节点与探针之间的传输时间 t_1．

公式（4）代表从邻居节点发送的交易到达探针

的时间，邻居节点和目标节点之间的距离为 1，即交易只需经过 1 次中转就可以到达探针．因此时间 T1 包括 2 个阶段：目标节点发送交易到达邻居节点的时间（100* S_{mft} + t_1+ t_2），其中 t_2 代表邻居节点收到交易到开始转发交易的时间；邻居节点到达探针节点的时间（100* S_{fmt} + t_1）．

公式（5）代表从其他 n 阶邻居节点发送的交易到达探针的时间，此类节点和目标节点之间的距离为 n，即需要经过 n 次中转才能到达探针．因此时间 T_n 包括 n+1 个阶段：目标节点发送交易到达邻居节点的时间（100* S_{mft} + t_1+ t_2）；邻居节点到达 n 阶邻居节点的时间（100* S_{fmft} + t_1+ t_2）*（n-1）；第 n 阶邻居节点到达探针节点的时间（100* S_{fmt} + t_1）．

由于每增加一轮中转都需要增加 1 组时间间隔（100* S_{fmft} + t_1+ t_2），因此 n 值越大，该交易首先到达探针节点的概率越小．因此，我们以 T 和 T_1 为例进行分析．当满足不等式（6）时，探针节点能够最早捕获到目标节点始发的交易，即交易排序满足网络拓扑规律．

$$T< T_1 \quad (6)$$

$$100* S_{mt} + t_1 < 100* S_{mft} + t_1 + t_2+100* S_{fmt} + t_1 \quad (7)$$

$$(S_{mt} - S_{mft} - S_{fmt})< (t_1+ t_2)/100 \quad (8)$$

公式（6）表示只有当目标节点的时间 T 小于邻居节点的时间 T_1 时，探针节点才能正确识别始发交易．公式（7）是带入公式（3）和公式（4）之后的中间结果．公式（8）是经过规整后的结果．

公式（8）中的 t_2 是指从节点收到交易信息到节点开始转发交易的时间间隔，t_2 的平均值可以使用探针节点获得，即利用探针节点向所有服务器节点发送交易，搜集每次的时间间隔，然后计算平均值．公式（8）中的 t_1 代表从节点转发交易信息到探针收到交易信息的时间间隔，此时间无法直接测得，我们使用可控节点转发交易信息到达探针的时间作为近似值.这种 t_1 的取值与实际情况存在误差，但是根据实验测试可知 t_2 远大于 t_1，因此 t_1 取值的误差造成的影响较小．经过实验测试，t_2 平均取值约为 400 毫秒，t_1 平均取值约为 58 毫秒．

基于上述的假设（t_2=400，t_1=58），可以将公式（8）转化为：

$$(S_{mt} - S_{mft} - S_{fmt}) < 4.5 \quad (9)$$

假设 P 代表目标节点转发的交易第一个到达探针的概率，则概率 P 实际上等于 S_{mt} 、S_{mft} 、S_{fmt} 三个变量满足公式（9）的概率．S_{mt} 、S_{mft} 、S_{fmt} 分别代表目标节点转发给探针节点的排序、目标节点转发给非探针节点的排序和非目标节点转发给探针节点的排序．根据比特币节点转发规则，排序主要取决于节点的邻居节点的个数．例如，目标节点有 N_m 个邻居节点，则 S_{mt} 的取值应满足：1<= S_{mt} <=125，而且 S_{mt} 取任意值的概率均为 1/ N_m．

在计算概率时我们从 S_{mt} 的角度出发，分别计算 S_{mt} 取不同值时满足公式（9）的概率．假设目标节点的交易首先到达探针的概率是 P，则 P 值就等于目标节点按照不同次序转发给探针并第一个到达探针的概率之和．假设目标节点在第 n 次挑选中将交易转发给探针，则这种情况下交易第一个到达探针的概率假设为 p_n，则 P 值可以使用公示（10）计算：

$$P=p_0+p_1+p_2+\ldots+p_n \ (0<= n <= N_m -1) \quad (10)$$

分别计算 p_n 的取值：

首先，当 S_{mt} 取值为（0,1,2,3,4）时，无论 S_{mft} 和 S_{fmt} 取何值，公式（9）恒成立．因此满足条件的概率 p_0=p_1= p_2=p_3=p_4=1/N_m．

其次，当 S_{mt} 取值为 5 时，只有当所有距离为 1 的非目标节点的 S_{mft} 和 S_{fmt} 都不满足（0,0）时，公式才成立，因此满足条件（9）的概率 p_5=(1/N_m)*(1-1/ (N_{fm}*N_m))$^{(Nm-1)}$．

当 S_{mt} 取值为 6 时，只有当 S_{mft} 和 S_{fmt} 满足（0,0）（1,0）（0,1）时，公式才不成立，因此满足条件（9）的概率 p_6=(1/N_m)*(1-3/ (N_{fm}*N_m))$^{(Nm-1)}$．

以此类推，当 S_{mt} 取值为 n 时，满足条件的概率 p_n 的取值的计算公式为：

$$P_n = (\frac{1}{N_m}) * (1 - \frac{X}{N_{fm} * N_m})^{(N_m - 1)} \quad (11)$$

其中 X=(n^2-7n+12)/2，X 推导过程见附录 A．

将公式（11）带入公式（10）可以得到：

$$P = \frac{5}{N_m} + \sum_{i=5}^{n} (\frac{1}{N_m} * (1 - \frac{i^2 - 7i + 12}{2 * N_{fm} * N_m})^{(N_m - 1)}) \quad (12)$$

根据公式(12)可知，P 的取值主要和节点邻居节点个数（N_m，N_{fm}）有关．在比特币网络中，比特币节点的邻居节点的个数各不相同，与网络状态、节点运行时间有关，但是邻居节点个数在分布上具有规律，根据 Biryukov 等人[19]的研究，80%以上的节点拥有的邻居节点数量低于 80 个．因此在计算概率 P 时可以假设比特币节点的邻居节点个数为 80．节点的邻居节点个数越多，比特币交易延

迟转发策略带来的干扰越大，因此假设邻居节点个数为 80，能够使我们推测的准确率在 80%的情况下比实际结果更保守．

将 N_m=N_{fm}=80 带入公式(12),则可以计算出 P 的近似值：P=0.189．P 值计算过程见附录 B．

P 值代表针对 80%的节点，在平均网络状态条件下（t_2=400，t_1=58），使用 1 个探针能够首先接收到目标节点转发交易的概率值是 0.189．在实际环境中，由于网络状态、邻居节点个数都不一样，针对不同服务器节点的 P 值各不相同．此处计算的 P 值可以作为平均值，用于估算溯源准确率．此外，P 值还可以用于估算邻居节点推测的准确率．

5.2 主动嗅探技术

为了实现轻量级溯源，我们设计了一种主动嗅探技术，可以利用较少资源推测目标节点的邻居节点．主动嗅探技术的流程如下：

1）部署一台可控节点，使用比特币系统命令“connect ip : port”，使可控节点只连接目标节点，即目标节点是可控节点唯一的邻居节点．

2）使用可控节点创建比特币交易并记录交易 ID．

3）利用探针搜集比特币网络中的交易信息．根据第 2 步记录的交易 ID 筛选出由可控节点创建的交易的传播路径．

4）针对第 3 步获得的交易传播路径，研究节点排序规律，并根据排序结果为每个转发交易的节点计算分值．排序越靠前，分值越高．

5）重复 2-4 步骤，分值累加．分值前 8 的节点为邻居节点．

上述流程中，由于可控节点只有目标节点这 1 个邻居节点，因此可控节点创建的交易将由目标节点首先转发到比特币网络．利用这种特点，可以针对任意服务器节点多次测试始发交易的传播路径，从而推测邻居节点．邻居节点推测的准确率在 5.3 节讨论．

主动嗅探技术的优点包括：

1)对比特币网络的影响较小，不容易被发现．首先，可控节点建立的单点连接和其他节点建立的连接在连接形式上没有区别，区别仅在于其他节点是随机连接多个节点作为邻居节点，可控节点只定向连接目标节点作为邻居节点．由于比特币网路中每个节点的网络拓扑只有节点自己知道，因此这种区别从目标节点和比特币系统的角度很难发现．其次，可控节点在执行主动嗅探机制时，虽然需要向目标节点转发多条交易信息，但是可控节点转发的交易在内容上与正常交易没有区别，而且转发交易的频率远低于比特币节点正常转发交易的频率，因此很难被识别．例如比特币服务器节点在正常运行时会持续向邻居节点转发交易信息，每分钟平均约为 150 条，而可控节点通常每分钟转发 10 条交易就足以满足嗅探机制的需要．此外，可以采用更换 IP 的方式使用具有不同 IP 地址的可控节点进行主动嗅探，进一步降低被发现的概率．

2）消耗资源较少．主动嗅探技术的主要消耗在于创建比特币交易的花费．由于探测过程在交易转播时已经完成，为了进行探测而发送的交易不需要被挖矿节点验证，因此，可以通过设置较低的手续费降低探测开销．经试验，每次测试所需的比特币低于 0.000003 个比特币（按照 2017 年 9 月份 2 万元每个比特币计算，价值约等于 0.06 元）．

3）主动嗅探技术可以用于选取阈值．交易溯源技术的准确率受到网络状态等多种因素的影响，针对可控节点计算的阈值并不一定适用于其他服务器节点．通过利用主动嗅探算法，可以为任意服务器节点创建始发交易，从而推测出针对特定服务器节点的阈值，提高溯源精度．

5.3 邻居节点推测分析

通过使用主动嗅探技术可以针对任意服务器节点创建始发交易，然后根据交易传播规律筛选出邻居节点．比特币网络中，每个节点的邻居节点数量小于 125 个，其中邻居节点中包含 8 个输出节点．这 8 个输出节点拥有较好的网络状态，实验发现绝大多数的首发节点都来自于这个 8 个节点．因此，我们希望挑选出 8 个邻居节点用于计算匹配值．

在由目标节点始发的交易中，邻居节点是第一批收到信息的节点，如果将邻居节点作为一个整体，则邻居节点发出的交易首先到达探针节点的概率与 5.1 节推测目标节点首发概率的推测过程类似，因此可以近似认为每次交易排序中首发节点是邻居节点的概率为 18.9%．根据此概率，可以计算出当有 40 次交易排序时，可以找到 40 个首发节点，其中大约包括 8 个邻居节点．将这 40 个疑似邻居节点带入公式（1）和公式（2）可以计算出匹配值的近似值．由于当前比特币服务器节点超过 7500 个，因此 32 个错误邻居节点带来的影响很小，低于 32/7500．

在实验中，可以通过增加探针的方式减少交易次数．例如，我们的试验中采用 2 探针模式，则发

送 20 笔交易，就可以得到 40 个疑似邻居节点.

5.4 阈值选择方法

阈值用于判断交易是否由目标节点始发．当交易的匹配值超过阈值时，即认为这笔交易属于目标节点始发的交易.

阈值的取值通过实验获得．在实验环境中，可以直接利用可控节点发送交易，然后测试始发交易和非始发交易匹配值的分布，选取能够有效筛选出始发交易的阈值．但是由于每个服务器节点的网络状态各不相同，实验环境中计算的阈值只能作为参考数据，并不能准确反映不同节点的真实情况，溯源机制需要对不同节点计算针对性的阈值．在实际溯源环境中，目标节点通常是不可控节点，不能直接操纵目标节点发送交易．因此我们采用主动嗅探机制使目标节点被动产生始发交易，即使用可控节点单点连接目标节点并发送交易，由于可控节点仅和目标节点连接，可控节点发送的交易将被目标节点最先广播到比特币网络，从探针的视角观察，此类交易和目标节点实际创建的交易大致相同，因此可以利用这种机制为不可控节点挑选阈值.

选择阈值的最优解对于提高溯源精度非常重要．阈值越高，筛选出的疑似交易的准确率越高，但是会导致很多交易被漏报；阈值越低，筛选出的疑似交易的准确率越低，但是捕获的交易数量越高．为了选择合适的阈值，我们使用准确率、召回率和 F 值作为评估条件.

定义 3．准确率．准确率 = 疑似交易中正确交易条数 / 疑似交易总数.

定义 4．召回率．召回率 =疑似交易中正确交易条数 / 发出的交易数量.

定义 5．F 值．F 值 = 准确率 * 召回率 * 2 / (准确率 + 召回率).

F 值是准确率和召回率的调和平均值，能够综合反映溯源效果,因此我们将 F 值作为评判标准.在选定阈值最优解时，我们通过实验针对不同的阈值计算准确率、召回率和 F 值，然后选择 F 值最高的阈值作为阈值的最优解.

6 实验验证

基于比特币交易溯源机制我们搭建一套交易溯源系统，可以针对比特币公共网络测试溯源效果．溯源系统中包括 2 个探针节点和 1 个可控节点．其中，探针节点部署了溯源程序，能够有效搜集比特币网络中的交易信息；可控节点部署比特币程序，能够自主创建交易、查看可控节点的邻居节点等信息．探针节点和可控节点的配置如下：

探针节点：16 Intel(R) Xeon(R) CPU E5-2682 v4 @ 2.50GHz，128GB RAM，500GB HDD. IP 位置：美国．操作系统：Debian 8.6 64 位.

可控节点：2 Intel(R) Xeon(R) CPU E5-2682 v4 @ 2.50GHz，4G RAM，250GB HDD.IP 位置：中国．操作系统：windows server 2012 64 位.

我们进行三类实验，测试方案的有效性和适用性．6.1 小节针对可控的目标节点测试溯源有效性；6.2 小节比较本文算法和同类算法的准确率；6.3 小节针对比特币网络中不可控的真实服务器节点进行交易溯源，测试在实际环境中的适用范围.

6.1 溯源有效性分析

本实验的目的是研究溯源机制的有效性，测试实际环境中的准确率和理论分析值的差距．本实验利用可控节点作为目标节点．实验时已知可控节点的邻居节点信息和始发交易的 ID.

实验中首先使用目标节点发送 10 次比特币交易，利用 1 个探针节点搜集交易传播路径．搜集的数据如表 3 所示．表 3 中记录了 20 条交易的测试结果.其中 1-10 号交易是由我们控制的目标节点始发，11-20 号交易是由比特币网络中的其他节点始发的交易(这 10 条交易是随机挑选的交易信息，用于展示始发交易和非始发交易的区别).每一行数据包括此交易的传播路径和计算的匹配值．其中，“Target”代表目标节点转发交易的排序，“N1-N8”代表 8 个邻居节点转发交易的排序，“Goal”代表根据公式（1）和公式（2）计算的匹配值.

通过观察表 3 的数据,我们可以发现一些规律：

1）1-10 号交易是由目标节点始发，可以发现目标节点有 3 次（第 4,7,8 号交易）的时间戳排序处于第 1 位，有 2 次（第 3,5 号交易）处于前 9 位（不包括首发）.而在 11-20 号交易中，目标节点的排序都超出前 9 位．这表明目标节点的排序在判断交易是否始发时能够起到重要作用.

2）1-10 号交易中邻居节点有 3 次（第 2,5,9 号交易）的时间戳排序处于第 1 位，而在 11-20 号交易中，邻居节点的排序都不在首位．这表明邻居节点的排序也可以用于区分始发交易和非始发交易.

3）表中最后一列“Goal”是交易对应的匹配值．通过分析 1-20 号交易的匹配值，可以计算出当选取 0.108345 作为阈值时，能够识别出所有的始发

表 3 交易排序数据

序号	Target	N1	N2	N3	N4	N5	N6	N7	N8	Goal
1	79	**3**	81	46	26	87	41	**5**	39	0.206188
2	49	80	**1**	**5**	**2**	12	22	92	46	0.572102
3	**8**	74	92	MAX	**5**	**2**	67	72	42	0.289731
4	**1**	42	45	**3**	64	81	**2**	10	MAX	0.753187
5	**2**	54	**1**	68	53	10	58	**5**	MAX	0.636634
6	71	34	**8**	81	43	18	78	69	15	0.108345
7	**1**	52	56	MAX	69	80	MAX	28	63	0.485336
8	**1**	34	79	MAX	29	MAX	54	23	75	0.496215
9	24	53	19	44	49	MAX	45	**1**	20	0.375299
10	28	26	11	17	35	40	56	23	87	0.110595
11	51	47	82	96	64	54	33	55	48	0.053098
12	49	65	**8**	50	38	60	58	43	10	0.11249
13	16	55	93	90	15	91	41	18	MAX	0.087534
14	42	78	89	87	20	44	68	**6**	62	0.102582
15	70	59	12	**8**	54	10	58	94	61	0.123012
16	87	47	90	MAX	11	14	88	57	91	0.075688
17	50	MAX	86	82	13	64	22	59	84	0.066291
18	46	MAX	18	45	**5**	70	72	76	**8**	0.143203
19	21	MAX	70	22	16	30	36	82	35	0.08878
20	20	48	84	56	51	36	79	49	22	0.075549

交易，即漏报个数为 0．此时误报个数为 3 个（第 12,15,18 号交易）. 出现误报的原因是由于网络延迟等干扰因素，导致邻居节点转发的交易信息落后于其它节点．此处的阈值取值只考虑召回率，实际测试时可以根据 5.1 小节介绍的阈值选取方法设置合适的阈值．

为了准确研究节点排序特征，我们进行 50 次实验，并设计了两项统计量：目标节点转发的交易第一个到达探针的概率；8 个邻居节点转发的交易第一个到达探针的概率．结果如表 4 所示．

表 4 节点排序首位概率表

序号	交易数量	目标节点首发	邻居节点首发
第 1 组	10*2	3 次	5 次
第 2 组	10*2	4 次	2 次
第 3 组	10*2	3 次	3 次
第 4 组	10*2	3 次	3 次
第 5 组	10*2	2 次	5 次
概率		15%	18%

本实验开展 5 组测试，每组测试发送 10 笔交易，由于有 2 个探针，每组测试产生 20 条交易路径，即有 20 个首发节点．表中分别记录了由目标节点和邻居节点发出的交易位于首发的个数和概率．其中，目标节点首发的概率约为 15%，此概率与 5.1 小节理论推导的概率 18.9%接近，存在的差距是由于网络延迟、节点个数不同导致．邻居节点首发的概率约为 18%，此概率值可以用于推测邻居节点．

此外，我们的实验还发现目标节点和邻居节点排在前 8 位和前 100 位（不包括第 1 位）的概率非常小，分别为 3/160 和 6/2000．这主要是由于探针连接的节点非常多（7000 左右），不同节点转发交易的时间间隔很小．因此在计算匹配值时优先考虑目标节点和邻居节点首发的情况．

6.2 溯源准确率对比分析

本实验的目的是将 2 种现有的比特币溯源技术和我们的方案进行比较，测试溯源准确率．

实验中的第 1 种方案（FirstReach）是由 Kaminsky[①]提出．判定条件是交易最先到达探针的

1 Kaminsky D. Black Ops of TCP/IP 2011. https://dankaminsky.com/2011/08/05/bo2k11/

节点为始发节点，即只有当目标节点的排序是第 1 位时，才推测这条交易是由目标节点始发.

第 2 种方案(Neighbor)是由 Biryukov 等人[19]提出. 判定条件是邻居节点排序落在前 8 位的数量超过 2 个时，则认为此交易是由目标节点始发. 此方案在推测邻居节点时需要持续向所有节点发送信息，有可能造成网络拥塞.

第 3 种方案(Combine)采用本文设计的溯源机制，利用 2 个探针搜集的数据计算优化匹配值. 根据优化匹配值筛选由目标节点始发的交易.

为了减少对比特币网络的干扰，本实验中不考虑方案不同的资源需求，假定 3 种方案得到的交易传播路径相同，而且能够准确获得节点的邻居节点信息.在这种条件下测试 3 种方案的推测准确率.试验中使用目标节点主动发送 10 笔交易，期间探针程序收到 12,378 笔交易. 溯源程序对每笔交易生成传播路径. 3 种方案采用各自的判定条件，对所有传播路径进行分析，推测出由目标节点始发的交易信息. 推测结果如表 5 所示. 表中的“正确结果”代表溯源机制推测结果中正确的数量，“输出结果”代表推测结果的总数量. 准确率代表溯源机制推测结果中正确交易占总交易数量的比例.

表 5 不同方案溯源准确率对比

方案	正确结果	输出结果	准确率	召回率
FirstReach	4	107	3.7%	40%
Neighbor	3	33	9.1%	30%
Combine	6	17	35.3%	60%

如表 5 所示，FirstReach 方案输出了 107 个推测结果，其中只有 4 个交易是正确的. 这种方案只认定最早到达探针的交易为始发交易，在实际环境中很容易受到干扰因素的影响，准确率较低，不具备实用性.

Neighbor 方案虽然只找出了 3 条正确交易，但是输出结果数量较少，准确率得到显著提高. 这说明方案采用 9 个节点进行估算的策略能够过滤掉大多数异常交易.

Combine 方案采用多探针架构，利用优化匹配值减少网络延迟带来的干扰. 当阈值设置为 0. 375 时，达到召回率 60%、准确率 35.3%的溯源精度，召回率和准确率都显著高于现有的 FirstReach 和 Neighbor 方案.

6.3 溯源适用范围分析

实验 3 利用探针节点对比特币网络中的非可控服务器节点进行探测，测试溯源机制针对非可控服务器节点的适用范围和溯源有效性.

为了测试溯源机制的适用范围，我们利用探针程序对比特币骨干网络进行探测，统计了连续 5 天中可以用探针节点探测到的比特币服务器节点的数量，数量平均值为 5368. 我们选定 Bitnodes① 网站同时期统计的服务器节点数量作为基准，服务器节点数量平均值为 7675. 因此，可以计算出探针的适用范围约为 69.9%. 这代表我们开发的探针程序能够对当前比特币网络中 69.9%的非可控服务器节点开展交易溯源.

经过分析，不能检测的节点包括 3 类：采用 IPV6 协议的节点、采用 TOR 加密协议的节点，以及经过特殊设置对访问 IP 进行设置的节点. 针对此类节点，可以通过扩展探针协议兼容性等方法提高适用范围.

为了测试溯源机制对非可控节点的有效性，我们从探针搜集的服务器节点中挑选一个服务器节点作为目标节点. 此节点的 ip 信息是（47.xx.xx.18）.

按照 4.2 节介绍的溯源流程，我们对此目标节点开展溯源. 试验中使用 2 个探针对比特币网络开展监测，使用 1 个可控节点作为主动嗅探技术中的可控节点.

监测环境建立以后，首先使用可控节点和目标节点建立单点连接. 然后使用可控节点创建 20 笔交易，并将 2 个探针搜集的 40 个首发节点作为疑似邻居节点. 接下来，我们利用可控节点发送 10 笔交易，并利用公式 1 和公式 2 分别计算每笔交易的匹配值. 经过实验，有 2 笔交易未被探针捕获，因此利用被捕获的 8 笔交易的匹配值作为阈值，分别计算准确率和召回率. 结果如图 3 所示.

图 3 反映溯源召回率和准确率随着阈值变化的情况. 虚线代表召回率的变化曲线，实线代表准确率的变化曲线. 曲线中的点代表阈值取上述 8 笔交易匹配值时的召回率和准确率. 每个点的上方标注当前匹配值 G 和 F 值的取值. G=0.3004 和 G=0.0205 各自代表 2 笔交易，因此图中一共有 6 个点.

从图中可以观察到准确率和召回率在交界处能够取得较为平衡的结果. 当阈值设置为 0.1752 时，召回率为 50%、准确率为 31.25%，此时 F 值获得最高值 0.3846. 因此，可以将 0.1752 设置为阈

① Bitnodes, Global bitcoin nodes distribution, https://bitnodes.21.co/

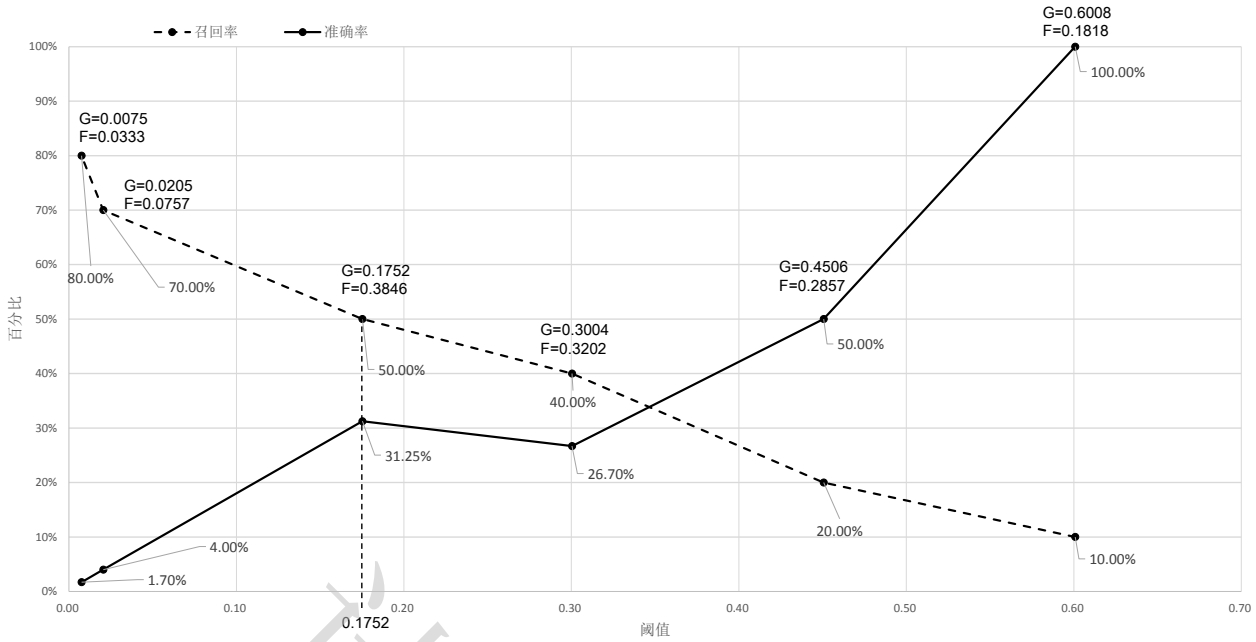


图 3 准确率与召回率随阈值变化散点图

值最优解，用于筛选可能由目标节点始发的交易信息，此时的溯源精度就是阈值最优解对应的准确率和召回率.

通过采用主动嗅探机制，可以针对不同的非可控服务器节点设置不同的阈值最优解，获得最优的溯源精度.

7 讨论

比特币交易溯源技术能够将匿名交易和发起者的真实身份相关联，为监测比特币交易、遏制基于比特币的犯罪活动提供了一种可行的解决方案. 本文的交易溯源技术在准确率、适用范围上还有很大提升空间，本节讨论溯源机制的改进思路和应用前景.

7.1 溯源技术应用范围

目前区块链技术发展迅速，除了比特币以外，还有很多类似比特币的区块链数字货币应用和非交易型数据区块链应用，主要可以分为三类.

1）基于比特币代码的山寨币：莱特币、狗狗币等. 此类数字货币应用利用比特币的开源代码建立，在底层协议和代码层面与比特币没有区别，因此，本文溯源机制可以直接应用到此类山寨币.

2）基于区块链技术的竞争币：以太坊、Zcash[26]等. 此类数字货币应用在代码层面与比特币不同，但是在网络层都采用 P2P 通信协议，而且都采用类似的交易转发策略. 因此，本文溯源机制在进行针对性改进之后也能适用.

3）针对非交易型数据区块链技术：例如超级账本[27]. 此类区块链技术的应用背景不是数字货币，甚至没有内置的代币. 在组网模式上通常采用联盟链和私有链，通常对联网节点设置身份鉴别机制，只允许通过验证的节点接入网络. 因此，本文溯源机制很难适用此类区块链应用.

7.2 客户端节点溯源技术

本文的溯源机制主要是针对服务器节点进行交易溯源. 针对客户端节点（以及其他不接受输入请求的节点，例如部署于 NAT 服务后的节点），本方案只能溯源到客户端节点创建的交易首次进入比特币网络的服务器节点的 IP，这与真正的始发节点还相隔 1 跳的距离.

为了解决最后一跳的问题，可以在本文溯源机制的基础上结合流量分析技术，从始发服务器节点追踪到真正创建交易的客户端节点. 首先采用溯源机制找到交易信息进入比特币网络的第一个服务器节点的 IP，然后针对服务器节点的入口流量进行分析，找到客户端节点 IP 地址. 例如，溯源技术针对交易哈希值为 tx1 的交易进行溯源，找到始发服务器节点的 IP 地址为 ip1，然后分析人员可以搜集服务器节点（ip1）的入口流量，从中筛选出符合特定格式的 IP 数据报（特定格式：IP 数据报中的目的地址是 ip1, IP 负载中包含字符串 tx1）. 找到的 IP

数据报的源 ip 地址就是客户端节点的 IP.

基于流量过滤的方法能够解决溯源最后一跳的问题，发现客户端节点的 IP 地址. 但是这种方法需要获得服务器节点的流量数据，具有一定的门槛. 根据《中华人民共和国网络安全法》第二十一条规定“网络运营者应当监测、记录网络运行状态、网络安全事件的技术措施，并按照规定留存相关的网络日志不少于六个月”. 因此，当发生基于比特币的犯罪行为时，执法机构有能力获得国内服务器的入口流量. 针对国外服务器，可以通过分析跨境流量的方式发现跨境的交易信息.

7.3 溯源技术应用前景

本文的溯源机制能够针对比特币网络中的服务器节点进行交易溯源，识别出由特定服务器节点创建的交易. 如果具有较多的服务器资源，可以通过部署多个探针同时监测大量比特币服务器节点，然后根据溯源结果为交易中的比特币地址赋予 IP 标签. 基于 IP 标签，可以推测匿名比特币地址对应的用户身份信息，为追踪恶意使用比特币技术的犯罪分子、遏制基于比特币的犯罪行为（例如比特币勒索病毒 WannaCry）提供技术支持.

8 总结

本文分析了比特币交易溯源的技术难点，提出了一种轻量级比特币交易溯源机制，能够追踪比特币交易的传播路径，将交易中的匿名地址和始发服务器节点的 IP 相关联. 这种关联关系可以用于发现恶意使用比特币交易的用户身份信息、追踪资金流向，为解决比特币勒索等非法比特币交易问题提供一种新的思路. 本文通过实验验证了溯源机制可以对真实环境中的非可控比特币服务器节点开展交易溯源，溯源准确率高于现有技术. 而且，本文分析了溯源机制对基于比特币代码的山寨币，以及基于区块链技术的其他数字货币同样适用，具有较强的实用价值.

参 考 文 献

[1] Yuan Yong, Wang Fei-Yue. Blockchain: The State of the Art and Future Trends. Acta Automatica Sinica, 2016, 42(4): 481-494 (in Chinese)
（袁勇，王飞跃. 区块链技术发展现状与展望. 自动化学报，2016, 42(4): 481-494）

[2] He Pu, Yu Ge, Zhang Yan-feng, et al. Survey on Blockchain Technology and Its Application Prospect. Computer Science, 2017, 44(4): 1-7 (in Chinese)
（何蒲，于戈，张岩峰，等. 区块链技术与应用前瞻综述. 计算机科学, 2017, 44(4): 1-7）

[3] Luo Qiang, Zhang Rui. Bitcoin. Beijing: China Machine Press，2014 (in Chinese)
（罗强，张睿. 比特币. 北京：机械工业出版社, 2014）

[4] Zhao Kuo, Xing Yong-heng. Security survey of internet of things driven by blockchain technology. Netinfo Security, 2017(5): 1-6 (in Chinese)
（赵阔，邢永恒. 区块链技术驱动下的物联网安全研究综述. 信息网络安全, 2017(5): 1-6）

[5] Moser M, Bohme R, Breuker D. An inquiry into money laundering tools in the Bitcoin ecosystem//Proceedings of the eCrime Researcher Symposium. Birmingham, England, 2014: 1-14.

[6] Bonneau J, Narayanan A, Miller A, et al. Mixcoin: Anonymity for Bitcoin with Accountable Mixes//Proceedings of the 18th International Conference on Financial Cryptography and Data Security Financial. Christ Church, Barbados, 2014: 486–504

[7] Ruffing T, Moreno-Sanchez P, Kate A. CoinShuffle: Practical Decentralized Coin Mixing for Bitcoin//Proceedings of the19th European Symposium on Research in Computer Security. Wroclaw, Poland, 2014: 345-364

[8] Bissias G, Ozisik A P, Levine B N, et al. Sybil-Resistant Mixing for Bitcoin//Proceedings of the 2015 ACM Workshop on Privacy in the Electronic Society. New York, USA, 2014: 149–158

[9] Ziegeldorf J H, Grossmann F, Henze M, et al. CoinParty: Secure Multi-Party Mixing of Bitcoins//Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. San Antonio, USA, 2015: 75-86

[10] Valenta L, Rowan B. Blindcoin: Blinded, Accountable Mixes for Bitcoin. Financial Cryptography and Data Security. Berlin,Germany: Springer, 2015:112-126

[11] Garay J, Kiayias A, Leonardos N. The Bitcoin Backbone Protocol with Chains of Variable Difficulty//Proceedings of the 37th International Cryptology Conference. Santa Barbara, USA, 2017: 291-323

[12] Bonneau J, Miller A, Clark J, et al. SoK: Research Perspectives and Challenges for Bitcoin and Cryptocurrencies//Proceedings of the 36th IEEE Symposium on Security and Privacy. California, USA, 2015: 104-121

[13] Zohar A. Bitcoin: under the hood. Communications of the ACM, 2015, 58(9): 104-113

[14] Zhu Liehuang, Gao Feng, Shen Meng, et al. Survey on Privacy Preserving Techniques for Blockchain Technology. Journal of Computer Research and Development, 2017, 54(10): 2170-2186 (in Chinese)
（祝烈煌,高峰,沈蒙，等. 区块链隐私保护研究综述. 计算机研究与

发展, 2017, 54(10): 2170-2186）

[15] Androulaki E, Karame G O, Roeschlin M, et al. Evaluating User Privacy in Bitcoin//Proceedings of the International Conference on Financial Cryptography and Data Security. Okinawa, Japan, 2013: 34–51

[16] Monaco J V. Identifying bitcoin users by transaction behavior //Proceedings of the 2015 International Society for Optics and Photonics Defense, Security, and Sensing. Baltimore, USA, 2015: 33-47

[17] Gervais A, Ritzdorf H, Karame G O, et al. Tampering with the Delivery of Blocks and Transactions in Bitcoin//Proceedings of the 22nd ACM Conference on Computer and Communications Security. Colorado, USA, 2015: 692-705

[18] Huang Bu-tian, Liu Zheng-guang, Chen Jianhai, et al. Behavior pattern clustering in blockchain networks. Multimedia Tools & Applications, 2017, 76(19): 20099-20110

[19] Biryukov A, Khovratovich D, and Pustogarov I. Deanonymisation of Clients in Bitcoin P2P Network//Proceedings of the 21st ACM Conference on Computer and Communications Security. New York, USA, 2014: 15–29

[20] Reid F, Harrigan M. An Analysis of Anonymity in the Bitcoin System//Proceedings of the Third International Conference on Privacy, Security, Risk and Trust. Massachusetts, USA, 2011: 1318—1326

[21] Liao K, Zhao Z, Doupe A, et al. Behind closed doors: measurement and analysis of CryptoLocker ransoms in Bitcoin//Proceedings of the Symposium on Electronic Crime Research. Toronto, Canada, 2016: 1—13

[22] Meiklejohn S, Pomarole M, Jordan G, et al. A fistful of bitcoins: characterizing payments among men with no names//Proceedings of the Conference on Internet Measurement. Barcelona, Spain, 2013: 127—140

[23] Zhao C. Graph-based forensic investigation of Bitcoin transactions. [M.S. dissertation]. Iowa: Iowa State University, 2014

[24] Koshy P, Koshy D, Mcdaniel P. An Analysis of Anonymity in Bitcoin Using P2P Network Traffic. Financial Cryptography and Data Security. Berlin ,Germany：Springer, 2014: 469-485

[25] Ivan P. Deanonymisation techniques for Tor and Bitcoin [Ph. D. dissertation]. University of Luxembourg, Luxembourg, 2015

[26] Sasson E B, Chiesa A, Garman C, et al. Zerocash: Decentralized Anonymous Payments from Bitcoin//Proceedings of the 35th IEEE Symposium on Security and Privacy. California, USA, 2014: 459-474.

[27] Dinh T T A, Wang J, Chen G, et al. BLOCKBENCH: A Framework for Analyzing Private Blockchains//Proceedings of the 2017 ACM International Conference on Management of Data. Chicago, USA, 2017: 1085-1100

附录A

公式（11）中X的计算方法.

满足条件的取值公式是：$(S_{mt} - S_{mft} - S_{fmt}) < 4.5$ （9）

因此，不满足条件的取值公式是：$S_{mft} + S_{fmt} > S_{mt} - 4.5$

当S_{mt}=5 时，不满足条件的差值为 0，即S_{mft}和S_{fmt}的组合为（0,0），个数为 1.

当S_{mt}=6 时，差值为 1，因此组合为（0,0）（0,1）（1,0），个数为 3.

当S_{mt}=n时，差值为t, t=n-5，此时组合的个数为$(t^2+3t+2)/2$，将t=n-5 带入，可以计算出$X=(n^2-7n+12)/2$.

附录B

公式（12）中的推导方法.

$P= p_0 + p_1 + p_2+\cdots+ p_n$， $p_0=p_1=p_2=p_3=p_4=1/Nm$

$p_5=(1/ N_m)*(1-1/ (N_{fm}*N_m))^{(Nm-1)}$)

$p_6=(1/ N_m)*(1-3/ (N_{fm} * N_m))^{(Nm-1)}$)

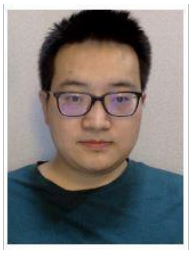
…

$p_n=(1/N_m)*(1-((n^2-7n+12)/2)/ (N_{fm}*N_m))^{(Nm-1)}$)

通过观察，可以发现 p_n的概率公式中包含小于 1 的小数的指数乘法（指数等于 N_m-1），根据指数乘法的定义，当 n 值越大，指数乘法中的小数取值越小，指数乘法的值会迅速变小. 因此可以预测当 n 的取值大于某个值是，p_n的值可以忽律. 将 $N_m=N_{fm}=80$(平均值)带入公式，依次计算 P_n，P_n 取值如下表所示.

P_n	值	P_n	值
p_5	0.012	p_{17}	0．004
p_6	0.012	p_{18}	0．003
p_7	0.011	p_{19}	0．003
p_8	0.011	p_{20}	0．002
p_9	0.010	p_{21}	0．002
p_{10}	0.010	p_{22}	0．001
p_{11}	0.009	p_{23}	0．001
p_{12}	0.008	p_{24}	0．001
p_{13}	0.007	p_{25}	0．001
p_{14}	0.006	p_{26}	0．001
p_{15}	0.006	p_{27}	0．000
p_{16}	0.005	…	0．000

可以看出，P_{27}之后的取值都为 0．因此可以计算出 $p_5 + p_6$



+.. +p_n= p_5 + p_6 +…+ + p_{26}=0.126.

Gao Feng, born in 1987. Ph.D. candidate. His research interests include blockchain application and network security.

Mao Hong-liang, born in 1990. Ph. D. His research interest is in blockchain application.

Wu Zhen, born in 1976. Ph. D. His research interest is in cyber security and blockchain application.

则，P=p_1+p_2+…+p_n=5/80+0.126=0.189.

Shen Meng, born in 1988. Ph. D. , assistant professor. His research interests include network security and privacy-preserving algorithms in cloud computing.

Zhu Lie-huang, born in 1976. Ph. D. , professor. His research interests include cryptography, network and information security.

Li Yan-dong, born in 1991. M. S. candidate. His research interests include blockchain application, cloud computing security and data privacy.

Background

The rapid development of Bitcoin technology and the growing scale of Bitcoin transactions have drawn wide attention at home and abroad. Whereas, Bitcoin is often used by terrorists and criminals attracted to the anonymity of the currency, such as all deals on Silk Road were made in Bitcoin. Therefore, it is essential to supervise Bitcoin and track the source transaction when necessary. However, as Bitcoin technology has the characteristics of de-centralization, traditional financial supervision means cannot provide effective supervision. Philip Koshy et al. [3] found some special trading patterns for originating node by analyzing the propagation law of currency transactions in the network layer, but the proportion of special deals is less than 9%. Alex Biryukov et al. [4] take advantages of the information of neighbor nodes of Bitcoin peer to locate the originating node. This approach improves fault tolerance and accuracy (experiment shows the accuracy of 11%), but requires constantly sending information to all nodes, which can cause network congestion. There are also some methods of transaction data analysis. However, they usually only get the relationship between the addresses, but cannot directly obtain the corresponding identity information of the address.

In this paper, we propose a transaction tracking mechanism in the Bitcoin network layer to track the propagation path of an anonymous transaction, thereby associating the transaction with the IP address of the originating node. By designing an active sniffing algorithm, our traceability mechanism supports lightweight monitoring and has a better practicality than traditional tracing mechanisms. We further design a mechanism which optimize the accuracy of network layer tracking by using address clustering information. We developed a prototype system for traceability mechanisms and tested the efficiency and accuracy on public Bitcoin network. The experiment results demonstrate that 69.9% of the backbone nodes in the Bitcoin network are suitable for the proposed tracing mechanism, with traceability recall rate of 50% and accuracy of 31.25%, which is superior to the current tracing methods and of great importance in practice.

This work is partially supported by the National Key Research and Development Program of China (No.2016YFB0800301), the Beijing Natural Science Foundation (No. 4164098), the National Natural Science Foundation of China (No. 61602039), the Guangxi Cooperative Innovation Center of cloud computing and Big Data (No. YD16E14), the CCF-Venustech Open Research Fund. The purpose of this project is to explore the monitoring method of Bitcoin, so as to provide a guarantee for the healthy development of finance.