文章编号:1007-5321(2018)02-0027-05 **DOI**:10.13190/j.jbupt.2017-237

基于启发式的比特币地址聚类方法

毛洪亮[1], 吴 震[1], 贺 敏[1], 唐积强[1], 沈 蒙[2]
(1. 国家计算机网络应急技术处理协调中心, 北京 100029; 2. 北京理工大学 计算机学院, 北京 100081)

摘要: 针对比特币这种新型的数字货币,通过分析其交易规律和交易地址关系,综合多个交易聚类特征,提出一种基于启发式条件的聚类方法,能够对匿名比特币地址进行相关性聚类,从而发现被同一用户团体控制的地址群,有助于分析用户的交易特征,推测用户的真实身份. 设计了具体的聚类方案,分析了迭代次数对聚类效果和代价的影响. 大量的实验分析结果验证了该方法的准确性和全面性.

关 键 词: 比特币; 启发式; 聚类
中图分类号: TN911.4 **文献标志码**: A

Heuristic Approaches Based Clustering of Bitcoin Addresses

MAO Hong-liang[1], WU Zhen[1], HE Min[1], TANG Ji-qiang[1], SHEN Meng[2]
(1. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China;
2. Department of Computer Science, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Because of anonymity of Bitcoin accounts, Bitcoin may be popular in some illegal trades and black market, such as the Silk Road. The author proposed an improved heuristic approaches-based method to cluster Bitcoin addresses and identify different addresses controlled by the same user. Three heuristic evidences were employed jointly. Through an elaborately designed experimental analysis, the precision and recall of the proposed method was verified. Furthermore, the efficiency with different iterations was analyzed, which can provide guidance in designing efficient clustering algorithm.

Key words: Bitcoin; heuristic; cluster

比特币起源于化名为中本聪的文章——Bitcoin: A Peer-to-Peer Electronic Cash System,具有一个全球共享的分布式、去中心化、去信任的账本系统,由计算机生成的一串串复杂代码构成,是目前为止区块链技术最成功的应用. 比特币可以由任何人发送给任何一个其他的人,无论他们是否在同一个城市或国家. 比特币账号是匿名的,无法被审查. 近年来,比特币和类似的新型数字货币发展迅速. 目前比特币市值已超过 1 000 亿美元,2016 年我国比特币交易额达 4 万亿元,占全球总交易量的 90% 以上. 此外,各国央行积极研究数字货币,甚至计划发行法定数字货币. 然而,与传统的资金交易系统相比,比特币交易具有较强的匿名性,很难进行有效管理,因此正在被广泛用于一些违法行为和黑市交易,例如枪支贩卖和毒品交易等. 在基于区块链技术的数字货币中,通常采用基于公钥的钱包地址作为用户在区块链网络上的假名,不同用户之间的交易通过这种假名实现. 这种假名通常由用户自由生成,与用户身份特征无关,因此很难通过分析交易数据推测用户的身份信息. 此外,区块链数字货币系统允许用户自由生成多个钱包地址,用户可以采用不同的钱包地址进行交易,以便减少单个钱包地址携带的用户交易特征. 因此,通过分析交易记录从大量假名中找出隶属于同一个用户的假名,并分析出

收稿日期: 2017-12-06
作者简介: 毛洪亮(1990—), 男, 博士, E-mail: mhl@cert.org.cn.

特定用户的交易规律,有助于推测用户的身份信息,对于遏制各类基于比特币的犯罪行为有重要作用.

笔者主要的工作和创新点包括:

1)利用提出的启发式方法,对比特币地址进行聚类,发现被同一用户团体控制的地址群. 启发式方法[1-9]的启发式条件在全面性方面均存在一定局限性. 对此,笔者提出了具有3种条件的启发式方法,具有更好的全面性.

2)通过大量的实验分析验证了方法的准确性和全面性. 同时分析了迭代次数对聚类效果的影响. 结果表明,聚类程序的迭代次数越多,获得的数据越全面,相反,耗时也会增加,研究结果显示其以线性趋势增长.

1 比特币交易

比特币可以抽象为交易用户之间的交易链,其加密方案通过非对称加密体制进行识别. 其中,加密使用的公钥可以生成比特币地址,简称地址.

1.1 交易分类与过程

比特币交易是一个包含输入和输出的数据结构,是将一定数量的比特币从输入地址转移到输出地址的一串代码信息.

交易类型分为产量交易、合成地址交易、通用地址交易.

1)产量交易

每个区块都对应一个产量交易,该类交易是没有输入交易的,挖出的新币是所有币的源头.

2)合成地址交易

该类交易的接收地址不是通常意义的地址,而是一个合成地址,以3开头,需要几对公私钥一起生成合成地址,在生成过程中可以指定几对公私钥中的几个签名以后,就可以消费该地址的比特币.

3)通用地址交易

该类是最常见的交易类型,由 N 个输入地址、M 个输出地址构成.

在比特币中,每一笔交易都是可追溯的. 交易的输入地址来源于之前一笔交易的输出,交易的输出地址又会在其他交易中作为输入,从而形成交易链. 根据交易之间的链式关系,分析人员可以获得任何一笔资金的使用情况和任何一个区块链地址的相关交易.

举例说明交易的简单过程,如图1所示. 假设由Alice、Bob和Mike 3个用户分别发起了3个交易A、B和C.

在交易A中,Alice发送2.5BTC给Mike,剩余的0.5BTC存到找零地址中,该找零地址属于Alice. 在交易B中,Bob发送2.0BTC给Mike. 在交易C中,Mike把从Alice和Bob那里收到4.5BTC中的4.0BTC发送给了别的用户,剩余的0.5个BTC存入找零地址,该交易中的2个输入地址和找零地址属于Mike.

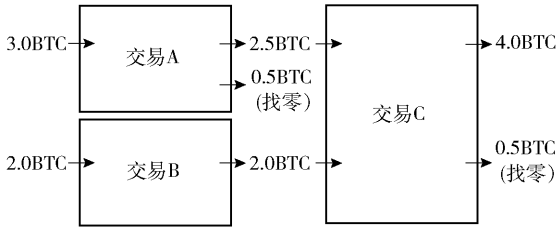

图1 比特币交易过程

1.2 地址归属定义

比特币地址类似于银行账号,任何人都可以产生无限数量的比特币地址. 因此,用户通常可以控制多个地址. 比特币交易由一组输入地址、一组输出地址和找零地址构成,输入地址属于付款人,输出地址属于收款人,找零地址用来存储付款人支付后剩余的金额,属于付款人.

定义1 用户集合 $U=\{u_1, u_2, \cdots, u_n\}$,比特币地址集合 $A=\{a_1, a_2, \cdots, a_n\}$,交易的集合 $T=\{t_1, t_2, \cdots, t_n\}$. 比特币交易的输入表示为 $\mathrm{Inputs}(t)$,交易的输出表示为 $\mathrm{Outputs}(t)$.

1.3 找零地址

在比特币交易中,有时作为输出的金额超过了用户想要支付的金额. 在这种情况下,比特币客户端会创建一个新的比特币地址,并把差额发送回这个地址,这就是比特币的找零机制.

找零地址和输入地址隶属于同一个用户. 找零地址是比特币系统自动产生的特殊地址,用于接收交易中的找零资金,并在以后的交易中由比特币程序自动选择作为输入地址. 因此,如果能够识别出找零地址,就能发现不同交易之间的关联关系,进而将多个交易中的比特币地址进行聚类.

2 启发式地址聚类

Man等[1-9]研究了在同一个交易中的所有输入地址都隶属于同一个用户集合(同一个人,或者一个机构)的启发式地址聚类方法,认为多输入交易

中的每个输入都需要单独签名，因此大多数多输入交易都是由同一个用户发起.

Meiklejohn 等[3-4]研究了找零地址、找零地址和输入地址隶属于同一个用户. 找零地址是比特币系统自动产生的特殊地址，用于接收交易中的找零资金，并在以后的交易中由比特币程序自动选择作为输入地址. 因此，如果能够识别出找零地址，就能发现不同地址之间的关联关系.

提出基于 3 种启发式条件的比特币地址聚类方法，通过结合多种启发式条件，有效提高聚类全面性，为推测交易背后的关联关系提供更好的基础条件. 需要说明的是，对于聚类算法的选择需要考虑两方面的原则：一是准确性，即分析得到的地址确实属于同一个用户团体控制；二是结果的全面性，即尽可能全地找到同一个用户团体控制的多个地址.

2.1　启发式条件

启发式 1：多输入交易地址聚类

当用户发生支付行为时，支付额度超过了用户 u 的钱包中每一个可用地址中比特币的额度，为了避免执行多笔交易以完成支付，从而造成交易费用方面的持续损失，比特币用户 u 会从钱包中选择多个比特币地址，把它们聚合的价值进行匹配支付，实现多输入交易. 由于比特币交易中使用每一个地址中的资金都需要单独签名，所以，通常认为一个多输入交易中的所有输入地址来源于同一个用户. 也就是如果 2 个或多个地址是同一个交易的输入，那么认为它们被同一个用户控制，即对任意交易 t，所有的 $a \in \text{inputs}(t)$ 被同一个用户控制.

在不考虑用户为规避聚类分析特意采用“混币”服务的情况，多输入地址聚类的准确率可以达到 100%.

启发式 2：产量交易地址聚类

1.1 节所述的产量交易是指比特币系统中创建比特币代币的交易. 区块链上的每个区块都对应一个产量交易，该类交易没有输入地址，只有输出地址. 产量交易中创建的代币是所有比特币交易的源头，新创建的代币将作为奖励发送给“矿工”，即交易中的输出地址. 由于挖矿的本质是在一台服务器上运行比特币挖矿程序. 所以，可以认为一个产量交易中的输出地址是由同一个用户进行配置.

如果一个或多个地址是同一个产量交易的输出，那么认为它们被同一个用户控制，即对任意产量交易，所有的 $a \in \text{outputs}(t)$ 被同一个用户控制.

对于用户自行挖矿模式的情况，产量交易地址聚类的准确率可达 100%. 对于“矿池”模式，多数情况下，出块奖励会在产量交易中转入“矿主”的私有收益地址，然后根据矿池用户的算力贡献进行二次收益分配，因此可以认为产量交易输出地址属于同一用户.

启发式 3：基于找零地址的聚类

找零地址是比特币交易中用于接收零钱（输入金额大于输出金额的部分）的地址. 此地址是由输入用户指定或系统自动生成，因此找零地址和输入地址属于同一个用户. 找零地址将在以后的交易中作为输入地址. 因此，结合启发式 1 的条件，通过将找零地址作为连接纽带，可以将 2 个交易中的输入地址聚类为同一个用户控制的地址群.

如果交易 t 产生一次性的找零地址，则该找零地址和交易输入地址由同一个用户控制，即对任意交易 t，$\text{inputs}(t)$ 的控制者，同样控制一次性找零地址 $a \in \text{outputs}(t)$.

由于比特币协议的变化，基于找零地址的聚类在准确率方面无法保证 100%（准确率具体指标会在后续工作中量化分析，本文通过实验结果定性验证），但是在全面性方面可以作为有效的补充.

2.2　找零地址识别算法

找零地址识别算法的核心工作是识别出输出地址中的找零地址. 找零地址的特征包括：作为输出地址的情况通常只会出现一次；找零地址不会同时出现在输入地址和输出地址；输出地址中不能只有找零地址.

找零地址识别算法[3]：若一个地址 a 满足下面的条件，那么该地址 a 是交易 t 的一次性找零地址：

1）a 只用作一次交易 t 的输出.

2）交易 t 不是产量交易.

3）对于 $a' \in \text{outputs}(t)$ 不存在 $a' \in \text{inputs}(t)$，即交易 t 不是“自我找零”交易.

4）对于 $a' \in \text{outputs}(t)$ 不存在 $a' \neq a$ 但 a' 只用作一次交易的输出.

2.3　聚类流程

比特币地址聚类方案总体框架如图 2 所示. 首先输入要查询的比特币地址；然后分别利用启发式 1、2 和 3 进行判断，将查找到的同一用户的比特币地址存储到地址集合中，利用相同的方法对新搜寻到的地址再次利用启发式方法进行查找其关联的地址，并存入地址集合. 迭代次数越多查找到的地址

就会越多,即结果的全面性越好. 但是,迭代次数的增加会显著降低聚类效率.

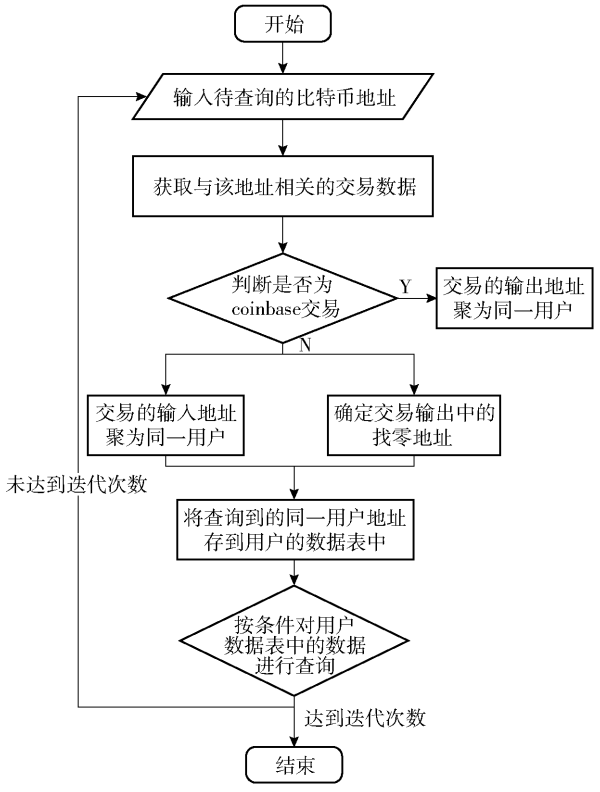

图2 聚类方案流程

方案过程如下:

1) 获得数据集

步骤1 输入需要查询的比特币地址.

步骤2 从整个比特币区块链交易数据中获得与该地址相关联的交易数据.

2) 获得同一用户控制的比特币地址

步骤1 产量交易地址聚类,即判断输入的地址是否为产量交易,如果是产量交易,则将交易中的所有输出地址归为同一用户.

步骤2 如果不是产量交易,则执行多输入交易地址聚类,即同一个交易的多个输入地址属于同一个用户.

步骤3 在步骤2的基础上执行找零地址识别算法,判定找零地址与输入地址属于同一用户. 本步骤找到的找零地址将存到同一用户地址数据表中,在下一轮聚类中作为输入地址. 如前所述,通过将找零地址作为连接纽带,能够将多个不同交易中的输入地址进行聚类.

重复执行1)和2),直到满足用户设定的迭代次数. 重复执行的次数越多,搜索到的关联地址就越多,耗时也就越多.

3 实验结果

方案使用Python语言实现,比特币区块数据存储的数据库为SQL Server,实验系统环境为Windows 7,CPU为i5-6200U,内存为8 GB.

同时,为了验证提出的启发式聚类方法的全面性和准确性,笔者采用2个方案进行验证.

方案1 利用笔者掌握的比特币客户端发送交易,与本文方法获得实验数据进行对比,以验证其结果的准确性和全面性.

方案2 通过同类网站,如walletexplorer.com已经聚类好的数据集,与笔者采用的方法获得的实验数据进行对比,验证其结果的准确性和全面性.

在实施方案1的过程中,使用2个笔者的地址进行多次交易进行测试;在实施方案2的过程中,使用500个地址组成的地址集合进行测试.

3.1 使用不同启发式的结果对比

以比特币地址"11g＊＊＊YZo"为例,分别使用启发式1、启发式1+启发式2、启发式1+启发式2+启发式3进行地址聚类,聚类到同一用户的比特币地址数量为6、10、15,如图3所示.

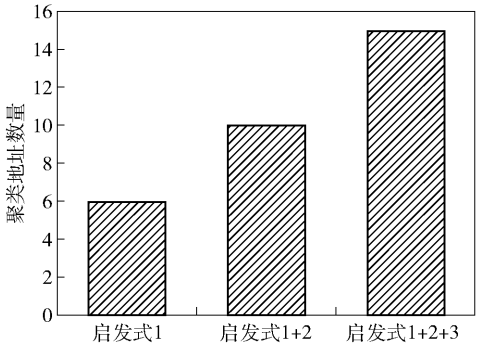

图3 使用不同启发式条件所得结果

3.2 性能分析

通过方案1对2个笔者掌握的比特币地址各进行了10次交易,这2个地址的各自交易分别涉及笔者控制的其他交易地址数量为22个和40个. 聚类后的实验结果与实际情况完全吻合,结果如表1所示.

图4所示为聚类地址数与迭代数的关系. 纵坐标表示查找到的与上述比特币地址属于同一用户控制的比特币地址数量. 从图4可以看出,聚类地址数量与迭代次数基本呈线性趋势增长.

表 1　测试地址实验结果 1

测试地址示例	本文方法（迭代 10 次）	准确率/%
1Pu＊＊＊MaP	22	100
1Dc＊＊＊C5y	40	100

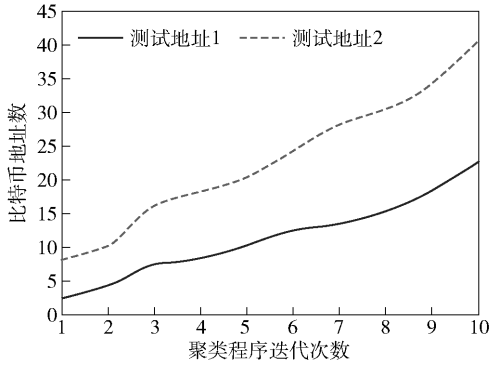

图 4　改变迭代次数结果对比

从图 5 可以看出，聚类耗时与迭代次数呈线性趋势增长.

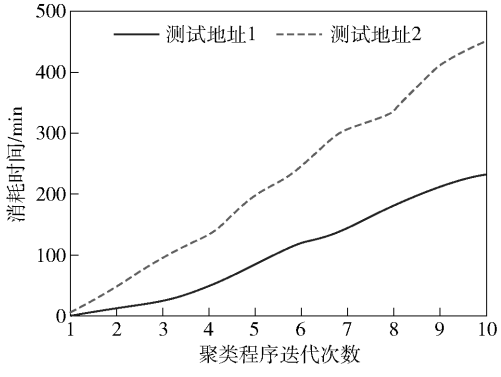

图 5　不同迭代次数的耗时走势

3.3　对比分析

通过方案 2 对随机选取的 3 个比特币地址进行测试. 利用图 2 所示的方案迭代 10 次获得的结果比 walletexplorer. com 网站聚类的结果更加全面，其结果如表 2 所示.

表 2　测试地址实验结果 2

测试地址示例	WalletExplorer 网站数据/个	本文方法/个
1BZ＊＊＊vao	5	6
12t＊＊＊2jd	17	20
167＊＊＊MML	219	264

4　结束语

针对利用比特币匿名特性进行非法交易的问题，深入分析比特币的交易规律，实现了一种启发式的比特币地址聚类方法，能够获得隶属于同一个用户的比特币地址群. 本文提出了具体的聚类流程，并通过大量的实验验证了该方法的全面性和准确性，分析了迭代次数对聚类效率的影响. 本方法对于区块链数字货币交易监管有着一定的理论意义和应用价值.

参考文献：

[1] Man H A, Liu J K, Fang J, et al. A new payment system for enhancing location privacy of electric vehicles[J]. IEEE Transactions on Vehicular Technology, 2014, 63(1): 3-18.

[2] Reid F, Harrigan M. An analysis of anonymity in the Bitcoin system[C]// IEEE Third International Conference on Privacy, Security, Risk and Trust. Boston: IEEE, 2011: 1318-1326.

[3] Meiklejohn S, Pomarole M, Jordan G, et al. A fistful of Bitcoins: characterizing payments among men with no names[J]. Communications of the ACM, 2016, 59(4): 86-93.

[4] Androulaki E, Karame G O, Roeschlin M, et al. Evaluating user privacy in Bitcoin[C]// International Conference on Financial Cryptography and Data Security. Berlin: Springer, 2013: 34-51.

[5] Monaco J V. Identifying Bitcoin users by transaction behavior[J]. Proc SPIE, 2015, 9457: 1-15.

[6] Liao K, Zhao Z, Doupe A, et al. Behind closed doors: measurement and analysis of CryptoLocker ransoms in Bitcoin[C]// APWG Symposium on Electronic Crime Research. Toronto: IEEE, 2016: 1-13.

[7] Ron D, Shamir A. Quantitative analysis of the full Bitcoin transaction graph[C]// International Conference on Financial Cryptography and Data Security. Berlin: Springer, 2013: 6-24.

[8] Huang B, Liu Z, Chen J, et al. Behavior pattern clustering in blockchain networks[J]. Multimedia Tools & Applications, 2017, 76(19): 20099-20110.

[9] Spagnuolo M, Maggi F, Zanero S. BitIodine: extracting intelligence from the Bitcoin network[C]// International Conference on Financial Cryptography and Data Security. Berlin: Springer, 2014: 457-468.