计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20180127
55(9):1853-1870,2018

区块链数据分析:现状、趋势与挑战

陈伟利　　郑子彬
(中山大学数据科学与计算机学院　广州　510006)
(国家数字家庭工程技术研究中心(中山大学)　广州　510006)
(chenwli9@mail2.sysu.edu.cn)

Blockchain Data Analysis: A Review of Status, Trends and Challenges

Chen Weili and Zheng Zibin
(*School of Data and Computer Science, Sun Yat-sen University, Guangzhou* 510006)
(*National Engineering Research Center of Digital Life (Sun Yat-sen University), Guangzhou* 510006)

Abstract Blockchain technology is a new emerging technology that has the potential to revolutionize many traditional industries. Since the creation of Bitcoin, which represents blockchain 1.0, blockchain technology has been attracting extensive attention and a great amount of user transaction data has been accumulated. Furthermore, the birth of Ethereum, which represents blockchain 2.0, further enriches data type in blockchain. While the popularity of blockchain technology bringing about a lot of technical innovation, it also leads to many new problems, such as user privacy disclosure and illegal financial activities. However, the public accessible of blockchain data provides unprecedented opportunity for researchers to understand and resolve these problems through blockchain data analysis. Thus, it is of great significance to summarize the existing research problems, the results obtained, the possible research trends, and the challenges faced in blockchain data analysis. To this end, a comprehensive review and summary of the progress of blockchain data analysis is presented. The review begins by introducing the architecture and key techniques of blockchain technology and providing the main data types in blockchain with the corresponding analysis methods. Then, the current research progress in blockchain data analysis is summarized in seven research problems, which includes entity recognition, privacy disclosure risk analysis, network portrait, network visualization, market effect analysis, transaction pattern recognition, illegal behavior detection and analysis. Finally, the directions, prospects and challenges for future research are explored based on the shortcomings of current research.

Key words blockchain; data analysis; Bitcoin; Ethereum; smart contract

摘　要 区块链是一项具有颠覆许多传统行业的潜力的新兴技术.自以比特币为代表的区块链1.0诞生以来,区块链技术获得了广泛的关注,积累了大量的用户交易数据.而以以太坊为代表的区块链2.0的

收稿日期:2018-02-21;**修回日期**:2018-06-24
基金项目:国家重点研发计划项目(2016YFB1000101);国家自然科学基金优秀青年科学基金项目(61722214);广东省高等学校珠江学者岗位计划资助项目(2016);广东省创新团队项目(2016ZT06D211)
This work was supported by the National Key Research and Development Program of China (2016YFB1000101), the National Natural Science Foundation of China for Excellent Young Scientists (61722214), the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2016), and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06D211).
通信作者:郑子彬(zhzibin@mail.sysu.edu.cn)

诞生,更加丰富了区块链的数据类型.区块链技术的火热,催生了大量基于区块链的技术创新的同时也带来许多新的问题,如用户隐私泄露,非法金融活动等.而区块链数据公开的特性,为研究人员通过分析区块链数据了解和解决相关问题提供了前所未有的机会.因此,总结目前区块链数据存在的研究问题、取得的分析成果、可能的研究趋势以及面临的挑战具有重要意义.为此,全面回顾和总结了当前的区块链数据分析的成果,在介绍区块链技术架构和关键技术的基础上,分析了目前区块链系统中主要的数据类型,总结了目前区块链数据的分析方法,并就实体识别、隐私泄露风险分析、网络画像、网络可视化、市场效应分析、交易模式识别、非法行为检测与分析等7个问题总结了当前区块链数据分析的研究进展.最后针对目前区块链数据分析研究中存在的不足分析和展望了未来的研究方向以及面临的挑战.

关键词 区块链;数据分析;比特币;以太坊;智能合约

中图法分类号 TP391

区块链技术是一种新型的分布式账本技术,它可以在互不信任的环境下实现去信任中介的可信交易.与传统数据库技术相比,区块链技术具有防伪造、不可篡改以及能方便实现智能合约等特点,被誉为一种将引发社会变革的新型技术[1].国务院印发的《十三五国家信息规划》将区块链等相关技术列入强化超前布局的战略性前沿技术.著名信息技术研究分析公司Gartner①连续2年(2017年、2018年)将区块链及其相关技术列入十大战略科技.

区块链技术是比特币等新型数字货币的底层支撑技术.由于要在分布式环境中实现可信的交易,区块链技术大量使用密码学技术隐藏用户信息,同时所有交易信息则由分布式网络共同验证、存储.区块链技术可根据应用场景和网络加入许可机制的不同划分为公有链、联盟链和私有链[2].比特币[3]、以太坊[4]等对节点的加入与退出没有任何限制,是典型的公有链.传统数据库中的数据通常隶属于某家企业或机构,只有内部人员能够查看和分析,而公有区块链因为可以自由加入与退出,其中的数据即区块链数据可以方便地获取.这为数据分析人员通过获取公有链的交易数据,进而分析系统中的各种行为提供了前所未有的机会.

当前,各种公有链如比特币、以太坊等获得了大量用户的参与,积累了大量交易数据.以比特币为例,一份ARK投资公司和Coinbase联合发布的研究报告指出,截至2016年底,全球有超过1000万用户持有比特币,每天比特币的交易量达到2亿美元[5].大量用户的参与和活跃的用户交易使得基于区块链的数据分析成为一个重要且有价值的研究问题.随着区块链技术的发展,各行各业将区块链技术作为底层技术引入,势必导致大量的数据以区块链数据的形式存在,因而研究基于区块链的数据分析问题具有重要的理论和现实意义.

与典型数据形式相比,区块链数据具有2个重要的特点:1)在区块链系统中,尤其是在公有链中,用户都是匿名的,各种用户的属性数据(如性别、年龄等)都无从获得;2)区块链系统中用户通过交易形成联结,构成各种网络、数据之间彼此互相关联.因而,基于匿名网络的数据分析技术将成为区块链数据分析的重要技术.

目前,由于区块链技术尚处于初始阶段,缺乏大量成熟的区块链应用项目,因而基于区块链的数据分析亦处于探索阶段.区块链数据分析相关的文献主要针对目前相对成熟且已有足够数据的区块链如比特币、以太坊等.其中,又以比特币区块链因其创立时间较长、广受关注而成为区块链数据分析研究的重要对象.

目前,已有大量针对区块链技术不同角度的文献综述,如技术架构[6]、共识机制[7]、安全和隐私的问题[8-9]、攻击问题[9]、应用现状[10-11]、研究方向与挑战[2]、研究热点[12]等,但尚缺乏对区块链数据分析技术的进展的相关报道,为弥补这一缺失,本文对目前区块链(主要是比特币、以太坊)数据分析的相关文献进行了对比分析,概括出2类典型的区块链数据和相应的分析方法,并总结了区块链数据分析的七大研究问题和进展,希望能够给当前区块链技术的相关研究提供一定的参考与帮助.

1 区块链基础介绍

当前,我们处于一个信息泛滥的时代,各种数据充斥我们周围.然而,数据是否真实可信却不得而知.

① http://www.gartner.com

区块链技术因其特殊的机制,其中的数据具有“可信”的宝贵特征,这使得基于区块链数据分析的信息具有重要的价值.为理解为何区块链数据具有可信的特征,本节将介绍区块链的基本架构,并重点分析使得区块链数据可信的关键技术.

1.1 区块链架构

2008 年,化名为“中本聪”(Satoshi Nakamoto)的学者在密码学邮件组发表比特币奠基性论文[3],并于 2009 年 1 月实现了比特币的最初版本.在经过一段时间的运行之后,比特币开始走进大众的视野.由于比特币具有许多优良特性,其迅速成为金融市场的宠儿.区块链技术正是指比特币等加密货币的底层支撑技术.

目前,尚未形成行业认可的对区块链技术的统一定义.在 2016 年 10 月由中国工业和信息化部发布的《中国区块链技术和应用发展白皮书》将区块链技术描述为分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式.在区块链的技术架构方面,袁勇等人[13]将其划分为数据层、网络层、共识层、激励层、合约层和应用层.然而,随着区块链技术的不断发展,区块链的内涵与外延也在不断演化.比如一些实际商业应用中,并不需要币的存在,因而激励层也并不存在.文献[14]从隐私保护的角度将区块链的架构划分为 3 个层次:网络层、交易层和应用层.本文站在数据分析的角度,从数据的类型和环境出发,认为区块链可以描述为三横一纵的结构,如图 1 所示:

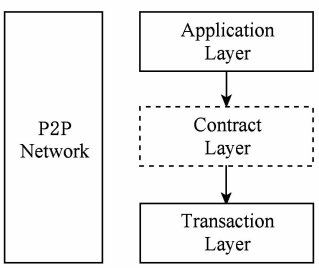

Fig. 1 The blockchain framework

图 1 区块链架构

三横既是对区块链数据类型的抽象,同时也代表区块链的 3 种发展阶段.最底层是交易(transaction)层,对应的是以比特币为代表的区块链 1.0 阶段.交易是改变区块链数据的手段,也是区块链中重要的基础数据.记录交易以及确保区块链数据本身的全局唯一性和不可篡改性是区块链 1.0 的核心.中间层是智能合约层,智能合约是密码学家尼克·萨博(Nick Szabo)于 1996 年提出的概念[15],旨在将合约条款电子化,当条件满足时,条款能自动执行,因而智能合约本质上是一段计算机程序.由于区块链数据具有不可篡改、完整可信等重要特征,基于区块链数据构建智能合约具有天然的优势.因而,区块链加智能合约被称为区块链 2.0.智能合约本身亦是构成区块链 2.0 的一种重要数据类型,我们称这种数据为合约数据.当然,智能合约的部署和运行离不开交易,同时也会产生交易数据.最顶层是应用层,应用层代表的是基于区块链的各种应用,对应区块链 3.0.应用可以建立在智能合约之上,实现许多复杂的自动化功能,也可以没有智能合约(因而合约层以虚线表示),如基于比特币的各种应用.目前,由于区块链技术尚处于发展初期,因而缺少与实际场景结合的应用数据.

一纵代表的是区块链的运行环境是分布式的.分布式环境穿越区块链的 3 个不同层次,表示 3 个层次都处于分布式环境中.比如同一个智能合约,作为一种数据存储在分布式环境中的每一个节点上,当接收到触发合约运行的交易时,每个节点基于本地的区块链数据运行相关合约,并将运行结果存入本地数据中,最后通过共识机制,使得本地的数据与整个分布式网络达到一致.在区块链的分布式网络环境中,通常存在着大量节点.不同的节点在网络中可能扮演着不同的角色.在比特币系统中,节点有钱包、挖矿(争夺记账权)、完整区块数据存储、路由 4 种角色[16].通常一个节点会因功能的不同实现不同的角色,比如一个轻量级的钱包节点不仅需要实现钱包功能,同时需要链接一定数量的节点以实现发布交易和验证交易的功能,即需要实现路由功能,但大部分情况下,由于资源的限制,并不需要实现完整区块数据存储,而只是存储区块链头部.我们把同时实现 4 种角色的节点称为全节点,通常一个挖矿节点会实现所有的角色而成为全节点.由于全节点实现了所有的角色,因而是整个网络中重要的支持和维护节点.本文中,当谈到节点时,若没有特别指明,都表示这类全节点.

1.2 区块链关键技术

区块链在本质上相当于一个去中心化的账本数据库,相对于传统数据库,其核心特征是“不可篡改”.正是不可篡改的特性,使得区块链数据具有“可信”的特征.在分布式环境中,实现一个不可篡改的账本,其关键的问题是数据如何组织以确保不可篡改以及如何在分布式环境中对账本状态达成共识.我们将解决这 2 个问题的技术概括为数据结构和共识机制.下面以比特币系统为例分别介绍这 2 个关键技术.

数据结构决定了区块链中账户和交易的组织形式. 在区块链系统中,通常采用 Merkle 树组织账本中所有的账户或发生的交易[16]. Merkle 树,又称 Hash 树,其上所有的值都是 Hash 值. 图 2 展示了比特币系统中采用的 Merkle 树结构. 在比特币系统中,当交易发生时,节点根据接收到交易的先后顺序或手续费高低等条件将交易排在一起,然后通过 Hash 运算得到每个交易的 Hash 值. 这个 Hash 值就是 Merkle 树的叶子节点值,此后通过将 Hash 值两两拼接在一起的再次 Hash 运算得到一个新的 Hash 值,依此自底向上通过不断地拼接 Hash 运算就可以得到 Merkle 树的树根节点值(Merkle root),这个根节点值代表了一段时间内被打包的所有交易的摘要信息. 当交易发生任何的篡改(如地址、数额等发生变化)或者交易组织的顺序发生任何改变,重复这个过程得到的新的树根节点必然与改变之前的树根节点完全不同. 与比特币的做法不同,在以太坊等区块链中,为了方便实现智能合约,引入了账户的概念,因而它们的做法是将账户组织到一个 Merkle 树上,账户代替了交易去做 Hash 运算,当账户状态发生变化时,对应的 Hash 值发生变化,最终导致 Merkle 树根节点值发生变化. 因此,不管是通过将账户还是交易组织到一棵 Merkle 树上,引进 Merkle 树的一个重要作用是当交易或账户状态被篡改时,重新计算对应区块中的 Merkle 树根节点必然跟篡改之前的不同,换句话说,Merkle 树根节点值可以看成是对账本当前状态形成了一个“快照”,这是区块链系统防篡改的第 1 步.

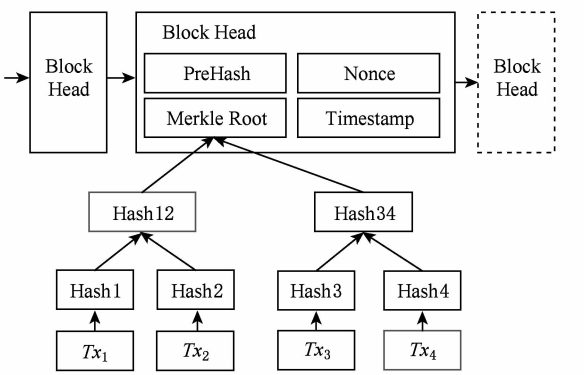

Fig. 2 Bitcoin blockchain data structure

图 2 比特币区块链结构

采用 Merkle 树组织交易的另一个重要作用是简化支付验证(simplified payment verification, SPV)[16]. 如图 2 所示,设想某轻量级钱包节点需要验证交易 4 的合法性,此时,该钱包节点只需要与邻近的数据节点(存储了所有区块数据的节点)通信获取对应区块的 Hash12 的值即可,而不需要区块中所有相关的交易信息. 因为有了这个值,该钱包节点即可通过 Hash 运算重新构造出区块的 Merkle 树根节点,与本地存储的对应区块头部中的根节点比对以验证其合法性.

决定区块链数据不可篡改的第 2 步也是最重要的一步是链式结构和共识机制. 在以比特币为代表的区块链中,当节点接收到系统中发生的交易时,首先验证交易的合法性,如是否存在“双花”问题等,确认不合法的交易被直接抛弃,而合法交易将广播给相邻节点同时存储在本地的区块中. 由于系统存在网络延迟,不同节点接收到交易的顺序可能不同,而且由于交易支付的手续费差异,不同的节点打包交易的顺序可能不同. 因而,不同的节点计算的系统当前状态的“快照”可能都不一样,那么如何在分布式环境下维护一个完整和唯一的账本即形成唯一的“快照”呢?一个普遍的做法是:根据某种条件,在分布式网络中挑选一个具有“优先记账权”的节点,然后网络中所有的节点都与之保持同步即可. 我们称选择具有优先记账权节点的机制为共识机制.

在比特币系统中,共识的过程可以划分为 3 个阶段:1)各个节点根据协议将接收到的交易打包进区块,并产生本节点对本地账本当前状态的快照;2)节点将当前状态的快照信息与其他信息一起填入区块头并找到一个合适的随机数,使得该区块头的 Hash 值小于某个给定的数或者该值的高位有足够的零. 由于 Hash 运算具有随机性和不可预测性,这个合适的随机数可能需要重复多次 Hash 运算才能找到,这一过程就是常说的“挖矿”. 在填入区块头的各种值中,有一个重要的值是上一个区块的 Hash 值,通过这个值,区块之间形成了一个“链”式结构(如图 2 所示). 一个节点选择将某个区块的 Hash 值存入当前区块然后去挖矿,代表着该节点接受了这个区块以及该区块链接的所有之前的区块的交易. 当然,每个节点都可以自由地选择从哪个区块开始挖矿,但系统规定节点应该选择接着最长的链去挖. 最后,最先找到合适的随机数的节点将其对应的区块广播给网络中相邻的节点,每个节点在接收到最新的块时,首先验证其合法性(如是否从最长的链开始挖,区块头 Hash 值是否满足条件等),在确认合法之后,节点将该区块广播给相邻节点,并以该新的区块为始,构造、寻找下一个合法区块. 当这个新

的区块被全网50%以上节点接受之后,找到该区块的节点相当于获得了优先记账权.

为了激励节点之间争夺优先记账权,每个区块的第1笔交易通常是一笔比特币的奖励交易,节点将自己作为奖励的接受方,但只有获得优先记账权之后,该奖励所得才能花费.此外,获得优先记账权的节点提交的区块中包含的所有交易的手续费也将同时作为该节点的回报.由于挖矿的过程需要不断的通过Hash运算寻找合适的随机数,因而节点间争夺记账权,本质上是Hash运算能力的竞争.由于这一过程需要一定的工作量,因而比特币共识机制也被称为工作量证明机制(proof of work, PoW).在实践中,不同的区块链系统可能采用不同的共识机制或不同的参数设置.

从以上的介绍可以看出,在比特币系统中,交易被组织进区块,同时区块又通过Hash运算串成了一个"链"式数据结构,再加上系统总是在该"链"的最新端链接新的区块.因而,随着交易和区块的不断增加,包含在历史区块中的交易信息被不断确认和锁定,任何试图篡改某笔交易的尝试,都必须重新计算这笔交易所在的以及后续所有的区块,并必须让重新计算的区块链最长以使得系统中其他节点接受新的链.但要做到这一点,需要攻击者具备系统51%以上的算力.因而,通过Merkle树,链式结构和共识机制,比特币系统实现了防篡改.当然交易不一定组织成链式结构,有向无环图(directed acyclic graph, DAG)也是一种重要的组织方式[17].

2 区块链数据类型

了解区块链的数据类型是进行区块链数据分析的前提,因此,本章主要介绍区块链数据的特点及其形式.区块链技术有3个不同的层次,但这3个层次并不是递进的关系,而是同时在发展着.目前,以比特币为代表的区块链1.0和以以太坊为代表的区块链2.0广受市场关注,发展相对成熟.因此本节主要介绍目前这2个层次中最重要的数据类型:交易数据和合约数据.

2.1 交易数据

1.2节中已经谈到,区块链数据相当于一个账本,交易是改变账本状态的基本事件.交易通常发生在账户之间,在区块链系统中,为了匿名性,通常用地址代表账户,地址是由字母、数字组成的字符串,如1BvBMSEYstWetqTFn5Au4m4GFg7xJaNVN[16].一个地址可能锁定了一定数据量的币(如比特币地址)或拥有自己的存储空间和代码(如以太坊合约地址).尽管地址是区块链系统中通用的数据载体,但不同的区块链系统可能采用不同的账本组织方式.在比特币系统中,并没有通常意义下的"账户"概念,作为一种强化匿名的手段,一个用户可以拥有任意多个地址,一个交易也可能同时涉及多个地址.事实上,比特币客户端会在交易中自动生成新的地址用于接受交易找零.图3展示了一个典型的比特币交易,在该交易中(Tx 为交易ID),有2个输入地址(A_1,A_2)上的币(输入额分别为5个币和6个币)被转移到了另外3个输出地址上(A_4,A_5,A_6),输出额分别为3,7,0.8个币.在一个交易中,输入总额与输出总额的差,称为该交易的手续费.图3对应的交易,手续费是0.2个币.手续费通常是由交易发起方在构造交易时指定的.一个地址上未被花费的交易输出数额称为一个未花费交易(unspent transaction output, UTXO).通常,一个地址上的所有的UTXO都对应着此前某些交易的输出额.因而,比特币地址上存储的UTXO,都可以按照产生该UTXO的交易回溯,直至挖矿所得.而挖矿所得是经过全网验证和接受的,由此,比特币系统实现了币的防伪造.

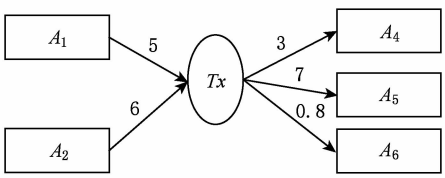

Fig. 3 A typical Bitcoin transaction

图3 一个典型的比特币交易

比特币系统的数据组织方式保留了交易过程中的所有细节信息,通过这些信息可以方便的验证一个交易中币的来源.然而,这一机制导致在比特币系统中对诸如验证一个用户的账户余额等这种简单操作变得低效,因为一个用户可能拥有多个地址,每个地址上又拥有多次交易产生的UTXO,要验证账户余额,需要验证所有地址上的所有UTXO的来源.这种低效使得基于比特币系统实现智能合约变得复杂.因而以太坊等实现智能合约的平台引入了账户的概念.账户在形式上仍是地址,但账户引入了存储空间用于记录账户余额、交易次数、代码等.因而,与比特币系统不同,以太坊等的账户可以看成是银行卡账户的一个类似物.

以太坊是一个基于区块链的智能合约平台,它提供了一种图灵完备的语言,使得基于以太坊可以

方便的实现去中心化的应用(DApp). 在以太坊中,账户分为 2 类:1)与银行账户类似的普通账户,用于记录用户参与交易的账户余额、交易次数等信息;2)智能合约账户,记录着合约的字节码等信息.

在以太坊系统中,由于有 2 类账户,其交易形式与比特币系统不同:当交易发生在普通账户之间时,交易与通常的银行转账交易是类似的,通常发生在 2 个账户之间,不存在多个输入或输出的情况;但当交易涉及合约对象时,情况较复杂. 此时,一笔交易可能是转账,对应一定数额的输入,也可能是调用合约的某个函数,或两者兼而有之. 图 4 给出了一个典型的涉及智能合约的交易. 在该交易中,普通账户 A_1 向智能合约账户 S_1 转账 5 个以太币,这是一个普通的交易(normal transaction),但该交易触发了智能合约向地址 A_2,A_3 的转账操作和调用智能合约 S_2 的操作,而对智能合约 S_2 的调用又进一步触发了其对地址 A_4 的转账操作. 所有后续的这些操作都是由第 1 个普通的交易触发的,因而称为触发的交易(fired transaction)或内部交易(internal transaction). 在以太坊中,一个普通的交易可能触发上百个内部交易,因而,事先无法确定交易能否成功. 一般情况下,普通交易都会记录在区块链上,但触发的交易,以及交易成功与否则需要执行合约代码才能确定. 尽管一个普通的交易,可能触发多笔交易,但他们具有相同的交易 ID.

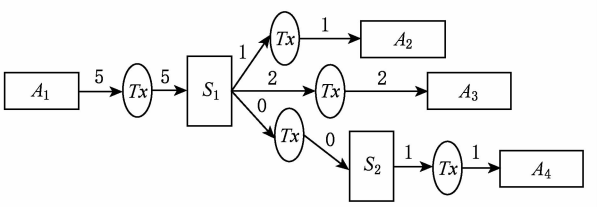

Fig. 4 A typical smart contract transaction

图 4 一个典型的智能合约交易

虽然交易在不同的区块链系统中可能具有不同的组织方式,但从数据的角度说,他们都具有相似的形式. 一个交易至少包含 5 个不可缺少的部分,即输入地址、输入额、输出地址、输出额以及交易发生的时间戳. 在不同的区块链系统中,这 5 个部分可能具有不同的要求,如在比特币系统中输入和输出地址可以是多个,而在以太坊系统中输入额等于输出额,同时一个交易可能对应多条记录.

2.2 合约数据

智能合约是区块链 2.0 的核心要素. 智能合约本质上是一段可以根据预先指定的条件被触发执行的代码. 基于不同的区块链系统,智能合约可以有不同的实现方法. 由于在目前的区块链技术中,以太坊是最流行的智能合约平台,因而以下主要介绍以太坊中智能合约数据.

以太坊提供了一个图灵完备的虚拟机(EVM)用于实现智能合约. 为了方便编写智能合约,以太坊提供了多种高级语言,其中 Solidity[18] 应用最为广泛. Solidity 是一门类似于 JavaScript 的面向以太坊虚拟机的智能合约编写语言. 目前,在以太坊平台上,采用 Solidity 编写的智能合约有超过 200 万个,而且每天都在不断地增加.

通常智能合约会涉及到 2 类数据:1)实现合约相关的代码数据;2)合约在运行过程中被触发的交易数据. 由于触发的交易数据并没有记录在区块链上,因而,需要运行相关的合约才能得到相应的数据. 一个替代的做法是通过网络爬取,目前,可以在 etherscan.io 上获取以太坊系统上智能合约相关的内部交易数据,但由于 API 的限制,只能获取每个账户最近 10 000 笔的交易.

代码数据是合约数据最重要的形式,因为代码的逻辑决定了内部交易的产生. 代码数据有 2 种存在形式:源代码和字节码. 源代码是以高级语言如 Solidity 编写的,可以通过阅读了解智能合约功能的文本数据;而字节码则是只对虚拟机有意义的数字串. 由于匿名性的需要,在以太坊中部署一个智能合约只需要提供相应的字节码即可,提供源代码只是方便使用者验证智能合约的内容,而并不是系统的要求. 目前以太坊平台上的超过 200 万智能合约中,只有 1 万多个(不到 1%)是可以查看源代码的,这使得基于代码数据存在大量研究问题. 由于代码数据中,大量存在的是字节码数据,除了可以通过字符相似度角度去挖掘合约间可能的关系外[19],能用的方法很少. 一个通常的做法是将字节码通过工具反编译为虚拟机的操作码. 以太坊黄皮书[20] 给出了虚拟机中的各种字节码对应的操作码及其代表的意义. 比如字节码 0x01 对应的操作码为 ADD,表示的是将 2 个操作数相加的运算. 在将字节码转换为操作码之后,可以方便地对合约代码数据进行分析,如合约是否存在漏洞[21]、合约能否优化[22]、合约以及区块链的性能评估等[23-24]. 目前关于智能合约的研究尚处于起步阶段,虽然由于以太坊的 DAO 攻击事件,吸引了一些对智能合约漏洞问题的研究,但大量其他问题仍有待探索,如这些合约都实现了什么功能、合约之间存在什么关系、是否有合约利用匿名的特性从事违法行为等.

3 研究现状与进展

本节主要介绍区块链数据分析主要回答的问题及进展情况.通过分析目前的相关文献,我们将当前区块链数据分析的研究概括为实体识别、隐私泄露风险分析、网络画像、网络可视化、交易模式识别、市场效应分析、非法行为检测与分析等 7 个研究问题.图 5 给出了这 7 个研究问题的关系,图 5 中实线箭头表示箭头发出端所代表的研究问题支持了箭头结束端所代表问题的研究,或者发出端是结束端的基础;虚线箭头表示发出端所代表研究问题在某一个角度的特殊化即是结束端所代表问题,即整体与部分的关系;而双向箭头表示两端的研究问题是同一个问题的不同侧面.

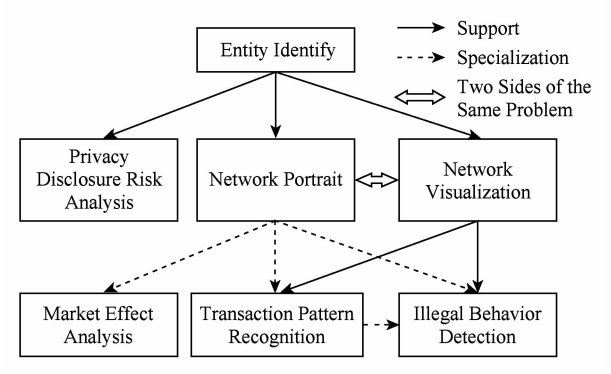

Fig. 5 Research problems and their relationship

图 5 研究问题及其相互关系

3.1 实体识别

由于在比特币的交易中,用户都是匿名的,而一个交易又可能涉及到多个输入和输出,因而一个自然的问题是:能否从交易记录中识别出用户,即哪些地址是属于同一个用户的.由于无法验证识别的是否是一个用户,在文献中通常认为识别的是实体(entity).一个实体可能是一个用户或者一个机构等[25];反之,一个用户或机构也可能控制着多个实体.在文献中,通常采用启发式方法识别潜在的实体,主要可以分为 2 种:共同输入法和找零地址法.所谓的共同输入是指在同一个交易中,将输入端的地址识别为属于同一个实体.因为在比特币系统中要花费一个地址上的币需要提供该地址对应的私钥,而通常用户并不会分享他们的私钥,因而可以认为一个交易的输入端的地址都在同一个实体的控制之下.以图 6 为例,在时刻 T_1,地址 A_1 和 A_2 因为同在交易 Tx_1 的输入端,可以认为是属于同一个实体,而到时刻 T_2,地址 A_3,A_4,A_5 因同时出现在交易 Tx_2 的输入端,可以识别为一个新的实体.到了时刻 T_3,地址 A_2,A_6 同时出现在交易 Tx_3 中,因而地址 A_2 所在的实体增加地址 A_6.这种动态地构造实体的算法虽然思想简单,但实现较复杂,文献[26]另辟蹊径,将 Petri 网的理论引入分析比特币交易.通过将地址和交易转为 Petri 网的矩阵表示,把比特币中许多问题(如实体识别)的分析转换为对矩阵的分析.由于矩阵运算易于理解和实现,该方法可以快速和方便地分析许多问题.但该方法的问题是随着交易的增多,矩阵的维数过大(因为一个地址对应矩阵的一行,一个交易对应矩阵的一列),这是该方法在实际中的重要缺陷.

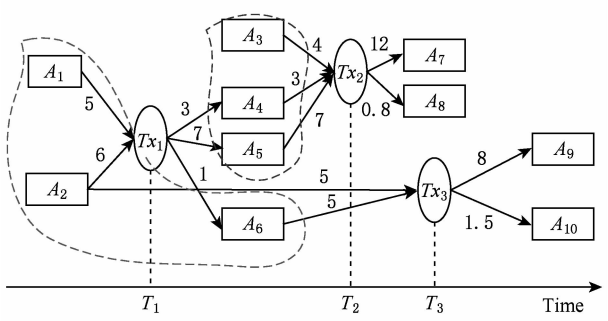

Fig. 6 A typical Bitcoin transaction fragment

图 6 一个典型的比特币交易片段

不管采用何种识别方法,一个明显的结论是随着时间的推移和交易的增加,在比特币系统中可以找到更多的实体,同时实体所包含的地址数量也可能相应地增加.实体的数量可以大致反应出系统中用户的数量,而实体的大小则一定程度上反映出实体的类型.如一般用户对应的实体通常较小,而网络钱包对应的实体可能包含大量的地址.值得注意的是,这一启发式方法并不总是成立的,事实上,由于这一启发式方法破坏了比特币的匿名性,因而多种针对这一方法的进一步提高匿名性的方法被提出,如 CoinJoin[27],MixCoin[28],BlindCoin[29].此外,文献[30]探讨了这种方法是否有效的问题.作者首先通过这种方法,找到许多包含至少 1 000 个地址的超级实体,其次,通过分析这些超级实体发现,它们都与某些服务有关,比如钱包服务,原因是这些服务通常重用一个地址为客户提供服务.重用地址不仅造成超级实体的存在,也会泄露服务提供者和用户的许多隐私.但作者指出,一个服务完全可以避免这种问题的存在,如一个有名的比特币钱包和交易提供商 Coinbase,并没有对应的超级实体存在.这说明,仅仅根据共同输入方法并不能有效确定实体与

用户之间的对应关系,一个用户可能存在多个与之对应的实体.

由于比特币钱包自动生成找零地址,因而在一个交易的多个输出地址中识别出找零地址是提高基于共同输入法建立实体的聚合度的重要方法.提高实体聚合度是指当采用一种方法识别的2个实体在另一种方法下被识别为同一个实体.关于找零地址的识别有多个启发式方法,最直接的是把一个交易的2个输出地址中唯一的新地址视为找零地址[31].这种方法正是利用了比特币系统自动生成新地址的机制,此后,文献[32]将输出地址为2个的限制条件推广为多个输出地址.另一个判断找零地址的方法是基于对输出数额的判断.如果一个交易中有2个输出数额,其中一个比另一个多出3个以上的小数位,那么小数位较多的输出认为是找零输出[32].比如一个交易中2个输出数额分别为0.03和0.013 345,那么可以认为0.013 345对应着找零输出.理由是用户在花费比特币时倾向于支付易辨别的数额.以上的启发式方法都是建立在某些假设的基础上,但这些假设是否可靠,以及根据这些假设构建的实体是否是对应着真实世界的用户无法验证.尽管如此,这些启发式方法仍然在分析比特币用户的行为上发挥了重要作用.

在共同输入和找零地址2类启发式方法的基础上,文献[33]提出通过社区发现的方法进一步聚合实体.作者通过实验对比了不同的启发式方法在识别实体的效果上的差别,实验结果表明单纯使用共同输入法可以获得较好的识别准确度但查全率相对较低,而如果只使用找零地址则准确度是最差的.然而综合运用所有的启发式方法,虽然准确度会下降很大,但查全率却可以达到最高水平.这说明利用社区发现的方法查找到的社区属于同一个实体的可能性不大,但通过这种方法可以发现普通启发式方法识别不了的潜在的实体.

3.2 隐私泄露风险分析

实体识别的目标是找到系统中属于同一实体控制的所有地址,虽然这对系统来说是一个风险因素,但由于地址只是一个无意义的假名,人们无法将其与真实的用户对应起来.但一个显然的结论是:如果我们能够获取某个用户在系统中的一个地址信息,那么这个用户的其他地址信息以及该用户在系统中的交易行为、账户余额等隐私信息就可能全部泄露,隐私泄露风险分析的目标在于回答,我们如何将系统中的实体对应到真实的实体,以及如果我们拥有了用户的一些额外信息,能在多大程度上获悉用户的地址?

针对如何获取实体的真实信息,文献[34]在利用启发式方法将12 056 684个地址聚合为3 383 904个实体的基础上,通过向接受比特币支付的网络商家购买产品和服务的方式给其中2 197个实体打上了某种标签.通过交易给地址打上标签的方式虽然可以很精确,但成本很高.由于人们可能在不经意间暴露自己的地址信息,比如论坛上某些用户贴出自己的比特币地址,以及某些机构公布自己的比特币捐献地址等,因而通过各种社区、网络获取地址的标签信息是一种成本相对较低的方法[32,35].此外,部署相关节点监听交易在比特币网络中的传播在一定程度上能够获取交易发出者的IP地址[36],从而暴露地址对应的位置信息.将实体打上某种标签之后,就可以深入分析比特币系统中实体的类别、分布及经济行为特征等.

针对基于额外信息,能在多大程度上获取地址信息的问题,目前的分析方法大致是:首先根据获取的实体信息,将交易数据用一个五元组(S,R,M,V,T)表示.一个五元组表示交易发送者S向接收者R在时刻T发送了总量为M的币,对应法币价值为V.其次,假定五元组中发送者是无法获知的,但其他4个信息是有可能暴露的,因而,可以借由其他4个信息,通过查询区块链数据获取交易发送者的地址信息.这个假定是合理的,因为首先用户在支付时,保护自己的账户信息是自然的,但一笔交易的其他要素则完全有可能暴露,比如2个人先后向同一家接收加密货币支付的咖啡店购买咖啡,那么前面顾客的其他信息(如支付金额、时间等)很容易被后面顾客获悉,此外,商家有可能将自己接受支付的地址信息放置在公共场所.基于上述的2个假设,文献[35]在分析比特币账本后指出,如果用户意外泄露支付的金额信息,那么窃听者可以通过分析在相应时间窗内,查找具有近似金额的所有可能交易,获取相应的地址.在2012年之前,大约可以找到10个可能的地址.当然,如果再结合一些其他信息,找到确定地址的可能性将大大增加.文献[37]基于Ripple网络的信息更全面的分析了不同信息泄露之后用户地址泄露的可能性.结果表明,如果泄露所有其他4个信息,那么发送者信息能够以99%以上可能性确定,即使不知道接收者和金额信息,仍能以90%以上的可能性确定发送者地址.这些研究结果表明,基于加密货币的支付存在暴露用户隐私的重大缺陷.

综合实体识别和隐私风险分析方法可以发现,在以比特币为代表的区块链系统中,存在许多分析用户隐私的方法,因而有许多基于比特币的服务提供各种保护用户隐私和钱包安全的方法.但一项调查了990名比特币持有者使用隐私保护策略情况的研究表明[38],只有46%的使用者会使用基于网络的隐私和安全保护服务,且其中的一半仅使用一种服务.同时,许多的参与者,对比特币可能涉及的隐私泄露问题缺乏很好的理解,只使用了相关服务的部分功能.此外,有22%的用户声称由于安全问题及操作问题丢失过比特币.这一现象表明,比特币的参与者中大部分缺乏对这一新兴事物的准确认识,因而容易引致各种违法犯罪行为.

3.3 网络画像

目前,比特币主链已挖出超过50万个区块,包含超过150 GB的交易数据.面对大量的交易数据,一个自然的问题是,数据中包含多少用户?这些用户有什么特征?这个巨大的支付网络是否具有一般的复杂网络的特征?比特币作为一种"资产",它是如何在用户之间分配的,是否满足一般的经济学规律等等.我们将这类研究整个网络的一些特征的研究概括为网络画像.下面主要介绍比特币网络画像的一些进展.

1) 活跃度画像.针对比特币网络是如何从最初的一个程序员"实验",发展成为一个庞大的加密货币帝国的问题,许多文献从网络活跃度及其动态变化的角度进行了分析.文献[25]基于共同输入启发式方法分析了2012年5月之前的比特币中所有交易数据,从3 730 218个地址中发现拥有2个以上地址的实体1 851 544个.基于识别的实体信息,作者分析早期比特币网络中币的活动特性,结果发现,早期的比特币系统中大量的币都没有进入流通领域,有大约78%的币都处于"休眠"状态,因为大量地址只接受比特币,而从不使用.文献[39]分析了2013年1月之前的所有比特币交易数据,并根据数据集中地址数量和比特币价格的不同将比特币划分为初创阶段和交易阶段.初创阶段从比特币2009年运行开始至2010年秋,这一阶段的特征是网络活跃度不高,网络特性不稳定,因而可以认为处于试验阶段.此后,比特币交易开始活跃,网络特征趋于稳定,进入交易阶段.通过对网络中地址的数量和使用情况的统计分析,他们发现,在交易阶段地址的使用情况(出入度)服从典型的幂律分布.文献[40]分析了比特币系统中常见的一些指标如每日活跃用户数、每日交易额与交易量等.统计结果显示,这些反映比特币活跃度的指标都在随着时间指数式地增长.此外,针对比特币的流动性问题,文献[40]从地址和币2个角度进行了分析,结果发现在比特币系统中,真正用于流通的比特币只有不到50%,而流通中的币大部分额度很小,且大部分小额交易都与赌博活动有关.与此类似,文献[41]则深入分析了比特币系统中各种指标如地址数量、活跃地址数量、实体数量、实体大小、币的流动性等随时间变化的函数关系.

2) 服务画像.比特币匿名又庞大的网络使得人们对比特币中到底有些什么服务,进行什么交易充满疑惑.为了回答这一问题,文献[40]首先基于启发式方法识别实体,同时通过利用购买物品和服务的方式进一步合并实体信息,从而获取大量基于比特币的服务对应的地址信息如矿池、钱包、换币服务、赌博等.他们分析了比特币系统中的流行服务和非法活动,发现在早期比特币的使用中,接近50%的交易都跟一个著名的赌博服务——中本聪骰子(Satoshi Dice)有关.文献[41]在此基础上通过获取Blockchain.info提供的交易的IP地址信息进一步将交易与对应的地理信息融合起来,分析了比特币的使用现状.该文给出了除赌博以外最热的24种比特币服务,并分析了各种服务在每个时间段的活跃情况.值得一提的是,由于引入了地理位置信息,该文指出比特币交易最活跃的区域是欧洲、美国、中国东部沿海地区、澳大利亚、巴西、加拿大南部和俄罗斯西部地区.文献[42]通过分析2009—2015年比特币的交易数据,通过启发式方法将地址信息汇集为实体,然后筛选出超大规模实体,并通过给实体赋予标签,将实体分为正常交易者、赌博者、黑市交易者和其他4种类型.最后,通过分析超大规模实体的交易模式和实体间的支付关系及其变化,将比特币系统划分为3个不同阶段,分别为概念验证阶段即挖矿阶段、非法交易(赌博、黑市)阶段和成熟阶段即交换阶段.与大多数研究针对网络整体进行分析不同,文献[43]分析了著名比特币黑市丝绸之路(Silk Road).通过爬取该网站每日出售的各种商品信息,发现绝大部分出售商品都是管控物品和毒品.文献[35]通过构造的实体网络图分析了比特币交易网络在一天中的典型网络形状,发现网络中蕴含着社区、超大交易子网络以及呈放射状的子网络等.此外,通过利用PageRank算法选出重要节点,并结合网络信息,其成功找到许多重要机构如"丝绸之路"[43]的地址.

由于实际上很难给所有地址打上合适的标签,文献[32]采用了一个完全不同的思路分析比特币中的服务.首先,基于构造的加权实体网络(节点之间的权可以有多种构成方式,如交易次数,交易量的大小等),通过运用网络聚类算法[44]识别网络中存在的社区,虽然发现有大量的实体并不能归结到某个社区中,但该方法很容易发现存在内部交易的社区,如在由 10～99 个实体构成的社区中,40%的实体与社区中其他实体有多笔交易,且社区中 68%的交易都是发生在社区包含的实体之间.其次,通过搜集 Bitcointalk.org 上用户自己公布的地址和用户所在的国家地区信息,作者建立了一个具有位置信息的地址样本,然后通过抽取实体的一些统计特征如每天交易次数、交易对手数目等特征,训练了一个随机森林模型[45],通过该模型,作者将所有实体纳入 4 个不同区域.最后,基于通过模型给出的标签,作者分析了比特币在这 4 个区域中的交易特征以及比特币的跨区域流动特性.

3) 网络特性.在早期关于比特币网络的研究中,发现了交易网络中存在幂律特征[39],由于在 2013 年之后,比特币交易量出现了爆炸式增长,文献[46]基于最新的数据分析了比特币系统中实体的数量和实体包含的地址数量的分布情况,发现实体的规模大致服从幂律分布,但明显存在一些异常值;针对这一现象,文献[47]重点分析了这些异常值,并证明了这些异常值源于用户的刻意行为;文献[48]利用最新的数据从复杂网络的角度分析了比特币网络的各种经典网络特征,如稠化过程、网络直径及平均距离、网络聚类系数、度分布以及一些中心化指标,并分析了这些指标随着时间的变化规律,结果表明随着时间的推移,比特币网络表现出小世界网络的特征.

在复杂网络分析中,偏好依附(preferential attachment)常用于解释各种网络属性是幂律分布的原因.而在经济学上,偏好依附又被称为"马太效应"或者富者愈富现象.由于比特币可以看成某种资产,因而检验其中是否存在这种现象是一个有趣的话题,一些文献对不同时期比特币系统中的马太效应进行了分析[39,47],结果表明在比特币网络中存在明显的财富聚集效应,即富者愈富现象.

3.4 网络可视化

随着区块链技术的火热,区块链中存入的交易数据不断增加,面对一个庞大且不断快速增长的交易网络,研究其可视化工具是一个重要的方向.目前,已有不少的研究对这一问题展开了探索.文献[49]在介绍比特币、以太坊等区块链特性的基础上,重点分析了区块链交易中蕴含的特殊图结构,可作为区块链图挖掘的入门材料;文献[50]介绍了一个比特币交易网络的可视化系统 BitConeView,利用该系统可以实时地、直观地追踪比特币中交易.特别是该框架中定义了"纯度"(purity)的概念,从而可以方便的找出混币交易,实现实时监控比特币网络中潜在的洗钱行为;GraphSense[51]是一个区块链网络的图形化分析工具,它不仅可以用于追踪资金流,实现自动的实体识别,同时可以用于搜索网络路径和特殊的交易模式;文献[52]描述了一个比特币交易的可视化监控系统,该系统聚焦于识别比特币交易中各种人为或算法的行为,通过该系统可以快速的识别比特币系统中异常的交易模式(如洗钱)和各种攻击区块链行为(如寄生虫交易攻击)等;与大多数研究聚焦比特币网络不同,文献[53]介绍了一个基于 Scala 的开源的分析框架,该框架可以用统一的方法同时分析比特币和以太坊这 2 个当前最重要的区块链.

3.5 市场效应分析

由于以比特币为代表的加密货币不仅有记录系统交易的区块链数据,一个更吸引眼球的数据则是加密货币与法币之间的兑换价格(为描述方便,以下简称加密货币的价格).截至写作当前,coinmarketcap.com 上记录的各种加密货币有 1 494 种,对应的交易所有 8 165 个,整个市值超过 5 600 亿美元,其中占据主导地位的比特币有接近 2 000 亿美元市值.市值排名前 3 的加密货币是比特币、以太币和 Ripple 币,其市值之和占整个加密货币市场的 60%以上.图 7 展示了比特币从 2013 年 4 月以来的每日价格、市值及成交量.从图 7 中可以明显地看出,比特币价格在短短的几年里从最初的几乎毫无价值到最高突破 2 万美元,而最近又因为某些原因跌到接近 1 万美元,可以说比特币等加密货币的价格具有极强的波动性.这种极强的波动性一方面吸引了经济学家们从金融学的角度探讨比特币是不是货币等问题[54-57](由于这方面的探讨超出了本文的范围,这里不做更多分析总结);另一方面,也更偏向数据分析的是,解释这种极端波动性的背后驱动因素是什么?本节拟对这一问题的研究进展做简单总结.

为了回答价格的驱动因素、分析价格波动的原因,在现有的研究中一个通常的思路有 3 步骤:1)寻找价格的潜在影响因素;2)利用各种影响因素构造

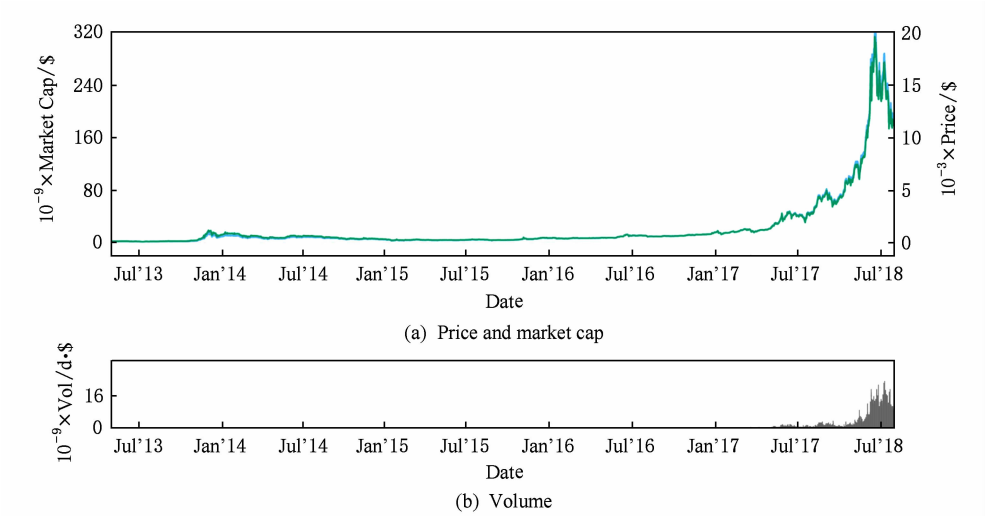

Fig. 7 Bitcoin price, transaction volume and market cap

图 7 比特币市场价格、成交量和市值图

时间序列,通过与价格时间序列构建相关统计、计量分析模型;3)结合模型估计的结果,分析价格与影响因素间的关系.

价格的影响因素是多方面的,因而准确全面地寻找价格的影响因素是成功解释价格波动原因的前提.目前,广泛使用的价格影响因素可以分为 6 类:1)矿工因素.矿工是维护加密货币的主要力量,也是加密货币的最初持有者,其行为必然对价格造成影响,有不少的指标可以反映矿工的行为如 Hash 比率(代表矿工的算力投入)、交易手续费、挖矿回报等.2)系统因素.加密货币系统的设置是影响货币价格的重要因素.以比特币为例,其币的供应量已被预先设定为接近 2 100 万个币,其挖矿难度每 2 016 个区块调整一次,挖矿回报每 21 万个区块减少一半,这些系统的设定和动态变化会影响加密货币的供应,因而最终也会影响其价格.3)用户因素.由于比特币已经从早期的矿工时代和以赌博、黑市交易等为主的非法时代进入成熟阶段[38],其参与者日趋广泛和多样,因而,用户的参与度和参与方式也必然是价格的影响因素.这部分的主要数据指标来源于区块链数据,如地址数量、实体数量、交易量以及交易额等.此外,以实体构建交易网络之后,从复杂网络的角度可以得到许多反映网络中用户参与交易的活跃度和交易模式的指标.4)政策、事件因素[58-59].由于加密货币的特殊性,使得其受政策影响非常大,此外发生在加密货币领域的典型事件,如漏洞攻击(DAO 攻击[60])、平台倒闭等[61],也必然对价格产生巨大影响.5)网络因素.网络搜索热度等反映的是普通网民对加密货币的"追捧"程度,也反映出加密货币潜在的用户规模及市场情绪,从而也是一个潜在的指标.6)竞争、替代因素[62].如前所述,目前加密货币市场有 1 494 种竞争币.在这上千种加密货币中,有些币具有完全不同的理念,但更多的币可能只是经典币在某种程度上的改进.因而,币之间的价格也必然存在着相互影响、相互替代的关系.此外,加密货币作为一种资产,与传统的资产如黄金、石油等之间也可能存在竞争或替代关系.

文献[63]采用 ARDL 模型分析了比特币价格的决定因素,认为比特币只是一个投机泡沫,远没有投资价值;文献[64]通过从交易网络中根据活跃度的不同提取了 2 类节点,构建了 2 个不同的子网络,并以这 2 个子网络每日的交易构建矩阵,进而采用主成分分析的方法构建特征序列,最后发现比特币价格与交易网络存在明显的相关关系;文献[58]采用小波分析的方法,研究了交易、技术、兴趣、避险、以及中国政策等价格影响因素,发现在不同时期比特币价格与其影响因素间存在不同方向的相关关系;文献[62]选取了市值靠前的 10 种加密货币,发现它们的价格表现出明显的非正态和长尾特征,并采用 Copula[65]方法捕捉价格间的相关关系,认为加密货币价格间存在着强化和替代 2 种不同的关系;一项对已经倒闭的比特币交易所 Mt. Gox 泄露的用户交易数据的分析指出[66],在 2013 年底,Mt. Gox 通过 2 个机器人操纵比特币价格,导致比特币价格在短时间内从 150 多美元上涨到突破 1 000 美元.

3.6 交易模式识别

与传统的银行等支付系统不同,比特币是一个匿名、无中心的支付系统.在匿名的情况下,人类的支付行为具有什么特点是一个有趣的问题.此外,匿名的特性导致基于比特币存在许多非法行为,如洗钱、诈骗等,能否从区块链的交易记录中识别特殊的模式而发现相关的非法行为是一个有价值的问题.解决这些问题的关键在于识别和分析比特币中的交易模式,下面介绍一些相关研究.

文献[25]基于建立的实体网络,通过选择大于5 000个币的交易构建了一个大额交易子网络.通过分析这个子网络,发现在早期的比特币交易中有许多特殊的交易模式,如分叉和自循环,即一个实体将一个地址上的比特币通过交易分割到不同的地址上,如此反复但最终所有的币又都汇集到该实体的某个地址上.如果说这种交易很有可能与早期的洗钱行为有关,那么另一种交易则似乎是刻意为之.文献[25]发现在交易网络中存在二叉树形式的交易模式,即一个地址将其上的币等额地存入2个地址,这2个地址又将存入的币等额地分别存入2个新的地址,如此反复使得交易网络变成一个类似二叉树的结构.

文献[34]发现一种被称为"剥离链"(peeling chain)的交易模式,该模式主要出现在拥有大额比特币的地址上,在交易过程中该地址每次支付小额比特币给某个地址,看上去像每次从原地址中"剥"去小部分,剩下部分则转入一个只用2次的找零地址(一次用于接收比特币,另一次则用于全部转出),这个过程可能重复成百上千次,直至找零地址上余额很少.这样,通过许多只用2次的找零地址,这些交易形成一条"链".文献[34]指出,具有这类模式的交易可能对应许多交易类型,如用户从钱包服务中取钱或矿池给参与者支付收益等.

文献[47]发现某些交易的输出中绝大部分是0.000 01比特币,作者将这类交易定义为"伪刷屏交易"(pseudo-spam transaction).一个"伪刷屏交易"是一个特殊的交易,该交易只有一个输入,但包含3个以上的输出,其中除不多于一个输出外,其余均为固定数额,如0.000 01比特币.通过分析大量的"伪刷屏交易",该文作者指出这种交易模式背后有2种极有可能的原因:去匿名攻击和发广告.去匿名攻击是指发送者通过给某个地址发送0.000 01个比特币,获取控制这个地址的实体的其他地址信息,因为这个数额太小,要花费这笔比特币必须与其他地址混合共同输入一个交易,从而暴露了实体的其他地址信息.而发广告是指在交易中包含一些广告信息使之永久的记录在区块链上或者在短时间内使之充斥整个网络达成某种目的.类似于"剥离链",文献[47]发现网络中存在大量"伪刷屏链",具体是指"伪刷屏交易"通过例外的那个输出串在一起的交易链.在这些"伪刷屏链"中最常见的输出是0.000 01比特币,占到所有这类交易的43.8%,而剩下最常见的输出是(1 000,7 800,10 000,100 000,200 000,500 000,1 000 000)等.在文献[67]中作者进一步分析了这种类似的交易行为,并指出正是这类特殊的交易行为导致了实体交易网络中入度分布的离群值.

3.7 非法行为检测与分析

由于区块链匿名的特性,很难得知交易参与者的身份信息,比特币等加密货币成为犯罪分子的天堂.文献[68]基于搜索数据将比特币的使用者分为计算机技术狂热者、投机客、自由主义者和犯罪分子等4类.在这4类参与者中,犯罪分子的行为破坏了技术发展的生态,导致了许多社会、经济问题.目前存在许多基于区块链技术的非法行为,如赌博[42]、洗钱[69]、贩卖违禁品[43]、诈骗[70]等.然而这些暴露的非法行为只是冰山一角,可能存在大量的非法行为并不为人所知.因而,基于区块链数据识别其中存在的非法行为不仅是促进区块链技术健康发展的需要,也能给区块链行业的监管、立法等提供参考.下面重点介绍基于区块链技术的洗钱和诈骗的相关研究.

1) 洗钱.历史上最大的跨国洗钱机构Liberty Reserve的取缔使人们看到了数字货币用于洗钱的可能性[71].而比特币等加密货币的天然匿名特性,使得比特币等成为潜在的洗钱工具.洗钱的最终目的是使得非法所得无所追踪.但基于前面介绍各种的启发式方法,尤其是共同输入方法,还是可以方便地将比特币中的实体识别出来.因而,一旦某个地址被确认为非法地址,其交易以及该地址对应实体的其他地址都能被追踪.因而,基于比特币的"混币"服务[27-29]成为洗钱的一个重要工具.混币服务的想法非常简单,多个人同时输入某个交易使得基于共同输入的启发式方法失效.图8是一个简化的混币交易,假定一个混币服务提供商M提供混币服务,3个用户(分别用3个地址A_1,A_2,A_3代替)希望混币以增强匿名性,他们分别给M的3个地址M_1,M_2,M_3存入1个币,同时提供各自的新地址A_4,A_5,A_6用于接受返回的币.M在M_1,M_2,M_3这些地址中随

机选择,将币返回给 3 个用户的新地址.只要 M 不用同一个地址接受和返回比特币,资金流关系就断了.在这里 M 虽然用了 3 个地址接受混币输入,但事实上,他可以在同一笔交易中用一个地址接受所有输入.当然,混币的方法还有很多,但可以明确的是:混币之后,不仅基于共同输入的启发式方法失效,也无从判别输出地址的归属.如图 5 所示的例子,即使我们知道某个用户(如 A_1)参与了混币交易,我们仍然无从判断 A_4,A_5,A_6 中哪一个属于该用户,因为看上去 3 个地址是同质的.

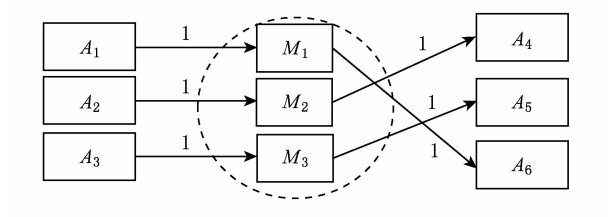

Fig. 8 A simplified mixing transaction

图 8 一个简化的混币交易

为了找到针对比特币的反洗钱措施,文献[69]通过真实参与 Bitcoin Fog,BitLaundry 以及 Blockchain.info 等提供的混币服务,进而利用区块链数据进行了大量实验,结果表明:基于 BitLaundry 的混币交易仍然能被追踪;而基于 Bitcoin Fog 和 Blockchain.info 提供的混币服务则不能再追踪.此外,基于发现的混币的特点,作者提出了许多相关的反洗钱措施.除通过混币服务洗钱以外,文献[34]概括了 3 种典型疑似洗钱的交易模式:汇聚(aggregations)、折叠(folding)和分割(splits).汇聚是指将许多相关比特币中的余额转入同一个地址;折叠则是将非法地址与其他正常地址混在一起做汇聚交易;分割则是将一个地址上的比特币转移到多个不同的地址上.

2) 诈骗.由于区块链同时具有参与者匿名、无国界限制、金融支付等天然属性,加上相关法律法规相对滞后,以比特币、以太坊等区块链为平台的各种非法行为(如黑市交易、庞氏骗局等)泛滥,这一方面为国家市场监管提出了巨大的挑战,另一方面也给参与区块链的投资者带来巨大的损失.2017 年 5 月,美国证券交易委员会(SEC)对涉嫌数字货币的庞氏骗局运营商提出诈骗指控,并罚款 7 400 万美元[72].2017 年 9 月 4 日,中国人民银行联合 7 家金融监管机构联合发布 ICO 禁令,其中一个重要原因是 ICO 可能涉及金融诈骗.这些事实表明目前基于区块链技术存在大量的非法行为.因而,研究如何通过区块链数据分析、识别潜在的骗局,是一个现实且紧迫的研究问题.

文献[70]通过聚合网络中各种骗局报告,搜集了 192 个基于比特币的诈骗案例.通过分析这些诈骗案例的运行特征,该文将它们划分为 4 种不同类型:庞氏骗局、挖矿骗局、诈骗钱包和虚假交易所等.通过获取的其中部分骗局涉及的地址信息,该文分析了骗局受害者的数量及造成的经济损失.

庞氏骗局是一种经典的骗局,其核心就是利用新投资者的资金回报老投资者,而骗局运营者则从中收取手续费等收入.智能合约的火热使得基于智能合约的庞氏骗局开始出现,由于基于区块链的智能合约具有可信、自动执行等特点,这种形式的骗局更具有欺骗性.文献[73]通过阅读各种开源合约的源代码以及搜索网络中的各种相关资料,确定了基于以太坊的 100 多个基于智能合约庞氏骗局,同时通过运用计算合约的字节码的基于标准化 Levenshtein 距离[74]的相似度,识别了部分隐藏源代码的庞氏骗局.在此基础上,该文将庞氏骗局划分为 4 种不同的类型,并深入分析了庞氏骗局的生存周期以及给参与者带来的收益和损失等问题.受此研究启发,我们提出利用机器学习和数据挖掘的手段识别区块链中存在的诈骗行为的研究问题,并以基于以太坊的庞氏骗局为突破口撰写了研究论文[75].在该文中,我们深入分析庞氏骗局合约的账户交易特征和字节码特征,并基于分析结果提取了相关特征,在此基础上通过训练分类器,获得了能以较高的准确度识别潜在的庞氏骗局的模型.值得指出的是,我们基于字节码建立的模型已经达到一定的准确度,这使得我们可以在智能合约刚上线还没有造成影响时预警庞氏骗局的可能性.通过将我们的模型应用到以太坊 20 多万非开源合约上,我们估计,基于以太坊有超过 400 个庞氏骗局在运行.未来,我们将会进一步提高模型的预测效果同时扩展骗局识别的种类.

4 趋势与挑战

区块链技术是一项新兴的技术,具有颠覆许多行业的可能性,尽管目前除了加密货币以及智能合约这个 2 个典型的应用场景外,尚缺乏有足够影响力的应用,但未来区块链技术将在广阔的领域发挥基础性作用.因而基于区块链的数据分析也将展现出其目标的多样性与技术的独特性.这节我们概述 3 个可能的未来研究方向及其面临的挑战.

4.1 趋 势

1) 网络特性与规律. 在 3.3 节的介绍中,我们已经指出利用复杂网络分析中成熟的各种技术分析交易网络是目前区块链数据分析中采用较多的方法. 然而,将交易数据简单地建构为一个复杂网络损失了大量的有用信息. 因而,未来在区块链数据分析的建模选择上需要考虑更多的信息,如交易的方向、数额甚至交易时间等,即通过交易数据构造有向网络(directed network)、加权网络(weighted network)、时间网络(temporal network)等,进而研究各种网络特性.

比特币网络是一个匿名、跨境、无中心的支付网络,目前该网络已经发展成一个巨大的财富帝国,尽管针对区块链数据分析已经取得了相对丰富的成果,但在诸多问题上仍然需要进一步深入研究. 一个典型的问题是:网络的生成机制. 这个网络是如何一步一步生成的? 是否有某种内在的机制决定了其生长? 它是否会继续壮大,还是会在某些因素的影响下最终消亡?这些问题的回答将帮助我们更好地理解加密货币网络的发展,但目前对这些问题的探讨尚非常初步. 此外,比特币作为一个加密货币系统,有其特殊的"货币"发行方式,这种发行方式对网络的形成与发展有什么影响、"货币"在系统中的流动又有什么规律?与现实金融系统相比其货币流动规律有什么特别的地方?现实与虚拟的 2 个世界是否有共通的地方等,回答这些问题不仅有助于我们理解虚拟货币的规律,也能为现实的一些金融政策提供参考.

2) 区块链监管与价值挖掘. 区块链技术的兴起和发展,给社会各行业和投资者带来前所未有的机会. 一方面,区块链技术带来了各种虚拟、匿名的世界. 匿名的世界用户的行为与现实世界可能存在极大不同,一个典型的案例是基于比特币的"丝绸之路"网络交易平台,明目张胆地将各种非法物品交易放置在了网络上. 此外,各种非法行为如庞氏骗局、钓鱼诈骗等利用区块链匿名的特性大行其道,相对滞后的法律措施和尚在发展的数据分析手段使得这些非法行为更是日益猖獗. 因而,区块链数据分析的一个重要任务是为区块链上的各种监管和追责提供丰富的技术手段.

另一方面,各种基于区块链技术的行业应用如雨后春笋般出现,尽管中国政府因为看到 ICO 包含的风险,已将其禁止,但全球范围内各种 ICO 仍然盛行,因而如何从大量的白皮书中挖掘出有价值的项目,同时识别出各种基于 ICO 的诈骗项目将是一个具有重要社会意义的课题.

3) 区块链数据分析+行业需求. 未来,随着技术的不断进步,诸多行业将采用区块链技术作为业务支撑技术,区块链技术将成为公司、行业基础设施,不同行业又通过跨链操作成为一个"区块链网". 区块链数据将成为公司、行业乃至整个社会数据的重要存在形式. 得益于区块链数据完整、可信的特征,未来区块链数据分析将体现更大的价值,发挥更大的作用. 在产业应用上,适应不同行业要求、针对不同应用问题的区块链数据分析技术将成为数据分析乃至形成"智能业务"的重要技术. 比如对存在大量审计需求的行业(如银行),可以通过区块链技术实现相关数据的存储,而区块链数据的可信、不可篡改的特性将使得基于区块链数据分析出的信息具有真实、准确的特征,因而一些传统需要人工审计的内容,也许可以交给实现了某种区块链数据分析技术的智能合约自动执行,从而形成"智能审计",提高行业效率.

在服务方式上,得益于行业联盟区块链中统一的数据格式,数据分析服务将可以以智能合约的形式存在. 公司和企业将可以通过购买相应的数据分析智能合约,基于行业统一的数据格式,通过给定本公司相应数据,即可以得到相应的分析结果.

因而,未来的区块链数据分析,如何结合行业特点,针对应用需求,通过智能合约的形式提供服务将成为研究重点.

4.2 挑 战

由于区块链技术具有广阔的应用前景,区块链数据分析也展现出了巨大的研究前景和应用价值. 然而,区块链数据具有的独特特点,使得区块链数据分析充满挑战.

1) 基于区块链数据的网络分析与传统的网络分析有着明显的不同. 在典型的网络研究中(如社交网络),节点和连边的含义是相对明确的,但在区块链数据中则不然. 比如在比特币交易网络中,通常我们希望了解交易背后的用户行为,但在构造网络时,只能选择 2 类节点:地址和实体. 而由于比特币网络允许一个用户拥有多个地址,且通常情况下,很难识别出一个用户拥有的所有地址. 因此,以地址和实体作为网络中的节点,事实上很难反映出真实的用户网络状态,而只是真实情况的一个近似. 虽然也能取

得一定的研究结论,但很难评估研究结果与真实情况之间的差距.在采用账户机制的网络中,节点的意义相对要明确很多,但其连边却复杂不少.以以太坊为例,节点可能是普通地址与合约地址,但普通地址与合约地址,以及合约地址之间的关系却非常复杂.由于合约既可以作为一个普通地址接受以太币,也可以作为一个功能集合(函数集合)对外提供某种服务,因而一个看似普通的交易背后,可以是转移支付,也可以是不同的函数调用,而由于系统设计的特点,交易还可能失败.因而,在以太坊网络的分析中,节点的连边信息更加丰富,可以构造的网络也更多样.比如基于同样的交易数据至少可以构造出以太币转移网络、合约创建网络以及合约调用网络等.总之,由于其节点和连边所代表意义不甚明确,区块链网络数据分析在方法的采用和结果的解读上都面临着新的挑战.

2) 区块链的去中心化和用户匿名特征让基于区块链数据分析的监管和价值挖掘充满挑战.在中心化的环境中,由于监管者可以获得所有的相关数据以及控制所有的节点,因而监管面临的问题只是手段需要升级.比如在银行交易记录中识别洗钱行为,一旦通过技术手段明确了有非法行为的账户,银行可以轻易地禁止相关账户的交易,避免损失和影响的进一步扩大,同时可以通过账户持有人信息找到相应的非法人员.但在区块链的世界里,我们无法禁止交易的发生,也很难找到非法行为背后的始作俑者.这一现状,将是区块链监管面临的巨大挑战.此外,由于监管的缺失,区块链项目在不需要满足任何条件的情况下就可以在市场上募集资金,这是区块链世界里普通投资者面临的巨大威胁.庆幸的是,与中心化情形下,数据通常不公开不同,区块链上数据都是公开的,这种特性为广大研究人员提供了前所未有的研究机会,也必将促进监管与价值挖掘手段的快速和不断升级.

3) 如果区块链成为一种"底层设施",数据实现了全行业流通,数据分析人员将面临全新的挑战.①数据的全行业流通必然导致数据意义的多样化,为了通过数据分析得出有价值的结论,需要分析人员更加透彻地理解数据背后的实际意义;②由于数据的可信和规范化,许多传统情况下需要人工完成的工作,可能会被用智能合约实现的 AI 取代.因而,未来的数据分析人员,不仅需要深刻地理解全方面的数据意义,同时需要了解智能合约的相关知识.

5 总 结

区块链技术是目前广受研究人员关注的一项新兴技术.本文总结目前区块链数据分析问题的研究进展.1)从数据分析的角度将区块链技术划分为三横一纵结构,紧接着介绍了使得区块链数据具有"防篡改"特性的核心技术;2)介绍了目前区块链中 2 种典型的数据类型;3)总结了目前区块链数据分析的七大问题与研究进展,并展望了未来区块链数据分析的研究发展方向和面临的挑战.

参 考 文 献

[1] Swan M. Blockchain: Blueprint for a New Economy [M]. Sebastopol, CA: O'Reilly Media Inc, 2015

[2] Zheng Zibin, Xie Shaoan, Dai Hongning, et al. An overview of blockchain technology: Architecture, consensus, and future trends [C] //Proc of the 5th IEEE Int Congress on Big Data. Piscataway, NJ: IEEE, 2017: 557-564

[3] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system [OL]. (2008-10-31) [2018-01-31]. https://bitcoin.org/bitcoin.pdf

[4] Buterin V. A next-generation smart contract and decentralized application platform [OL]. [2018-02-08]. https://cryptorating.eu/whitepapers/Ethereum/Ethereum_white_paper.pdf

[5] Burniske C, White A. Bitcoin: Ringing the bell for a new asset class [EB/OL]. [2018-01-30]. https://research.ark-invest.com/hubfs/1_Download_Files_ARK-Invest/White_Papers/Bitcoin-Ringing-The-Bell-For-A-New-Asset-Class.pdf

[6] Bonneau J, Miller A, Clark J, et al. SoK: Research perspectives and challenges for Bitcoin and cryptocurrencies [C] //Proc of the 36th IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2015: 104-121

[7] Du Mingxiao, Ma Xiaofeng, Zhang Zhe, et al. A review on consensus algorithm of blockchain [C] //Proc of 2017 IEEE Int Conf on Systems, Man, and Cybernetics (SMC). Piscataway, NJ: IEEE, 2017: 2567-2572

[8] Conti M, Lal C, Ruj S. A survey on security and privacy issues of Bitcoin [EB/OL]. [2018-02-08]. https://arxiv.org/pdf/1706.00916

[9] Atzei N, Bartoletti M, Cimoli T. A survey of attacks on Ethereum smart contracts (SoK) [C] //Proc of the 6th Int Conf on Principles of Security and Trust. Berlin: Springer, 2017: 164-186

[10] Hamida E B, Brousmiche K L, Levard H, et al. Blockchain for enterprise: Overview, opportunities and challenges [EB/OL]. [2018-02-08]. https://hal.archives-ouvertes.fr/hal-01591859/

[11] Zhao J L, Fan Shaokun, Yan Jiaqi. Overview of business innovations and research opportunities in blockchain and introduction to the special issue [EB/OL]. [2018-02-08]. https://link.springer.com/article/10.1186/s40854-016-0049-2

[12] Yli-Huumo J, Ko D, Choi S, et al. Where is current research on blockchain technology?—A systematic review [EB/OL]. [2018-02-08]. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163477

[13] Yuan Yong, Wang Feiyue. Blockchain: The state of the art and future trends [J]. 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[J]. 自动化学报, 2016, 42(4): 481-494)

[14] Zhu Liehuang, Cao Feng, Shen Meng, et al. Survey on privacy preserving techniques for blockchain technology [J]. 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[J]. 计算机研究与发展, 2017, 54(10): 2170-2186)

[15] Szabo N. Smart contracts: Building blocks for digital markets [EB/OL]. [2018-01-30]. http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart_contracts_2.html

[16] Antonopoulos A M. Mastering Bitcoin: Unlocking Digital Cryptocurrencies [M]. Sebastopol, CA: O'Reilly Media, Inc, 2014

[17] Churyumov A. Byteball: A decentralized system for storage and transfer of value [EB/OL]. [2018-02-08]. https://byteball.org/Byteball.pdf

[18] Ethereum. Solidity documents [EB/OL]. [2018-02-08]. http://solidity.readthedocs.io/en/develop/

[19] Bartoletti M, Carta S, Cimoli T, et al. Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact [J]. arXiv preprint, arXiv: 1703.03779, 2017

[20] Wood G. Ethereum: A secure decentralized generalized transaction ledger [OL]. [2018-02-08]. http://gavwood.com/Paper.pdf

[21] Luu L, Chu D H, Olickel H, et al. Making smart contracts smarter [C] //Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 254-269

[22] Chen Ting, Li Xiaoqi, Luo Xiapu, et al. Under-optimized smart contracts devour your money [C] //Proc of the 24th IEEE Int Conf on Software Analysis, Evolution and Reengineering (SANER). Piscataway, NJ: IEEE, 2017: 442-446

[23] Zheng Peilin, Zheng Zibin, Luo Xiapu, et al. A detailed and real-time performance monitoring framework for blockchain systems [C] //Proc of the 40th Int Conf on Software Engineering: Software Engineering in Practice Track. Piscataway, NJ: IEEE, 2018: 134-143

[24] Dinh T T A, Wang Ji, Chen Gang, et al. Blockbench: A framework for analyzing private blockchains [C]//Proc of the 2017 ACM Int Conf on Management of Data. New York: ACM, 2017: 1085-1100

[25] Ron D, Shamir A. Quantitative analysis of the full Bitcoin transaction graph [C] //Proc of the 17th Int Conf on Financial Cryptography and Data Security. Berlin: Springer, 2013: 6-24

[26] Pinna A, Tonelli R, Orrú M, et al. A petri nets model for blockchain analysis [J]. arXiv preprint, arXiv: 1709.07790, 2017

[27] Gregory M. CoinJoin: Bitcoin privacy for the real world [EB/OL]. (2013-08-22) [2018-01-31]. https://bitcointalk.org/index.php?topic=279249

[28] Bonneau J, Narayanan A, Miller A, et al. Mixcoin: Anonymity for Bitcoin with accountable mixes [C] //Proc of the 18th Int Conf on Financial Cryptography and Data Security. Berlin: Springer, 2014: 486-504

[29] Valenta L, Rowan B. Blindcoin: Blinded, accountable mixes for Bitcoin [C] //Proc of the 19th Int Conf on Financial Cryptography and Data Security. Berlin: Springer, 2015: 112-126

[30] Harrigan M, Fretter C. The unreasonable effectiveness of address clustering [C] //Proc of the 2016 Int Conf on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress. Piscataway, NJ: IEEE, 2016: 368-373

[31] Androulaki E, Karame G O, Roeschlin M, et al. Evaluating user privacy in Bitcoin [C] //Proc of the 17th Int Conf on Financial Cryptography and Data Security. Berlin: Springer, 2013: 34-51

[32] Athey S, Parashkevov I, Sarukkai V, et al. Bitcoin pricing, adoption, and usage: Theory and evidence [EB/OL]. [2018-02-08]. https://ssrn.com/abstract=2826674

[33] Remy C, Rym B, Matthieu L. Tracking Bitcoin users activity using community detection on a network of weak signals [C] //Proc of the 6th Int Workshop on Complex Networks and Their Applications. Berlin: Springer, 2017: 166-177

[34] Meiklejohn S, Pomarole M, Jordan G, et al. A fistful of Bitcoins: Characterizing payments among men with no names [C] //Proc of the 6th Int Conf on Internet Measurement. New York: ACM, 2013: 127-140

[35] Fleder M, Kester M S, Pillai S. Bitcoin transaction graph analysis [EB/OL]. [2018-02-08]. https://arxiv.org/abs/1502.01657

[36] Reid F, Harrigan M. An Analysis of Anonymity in the Bitcoin System: Security and Privacy in Social Networks [M]. Berlin: Springer, 2013: 197-223

[37] Di Luzio A, Mei A, Stefa J. Consensus robustness and transaction de-anonymization in the Ripple currency exchange system [C] //Proc of the 37th IEEE Int Conf on Distributed Computing Systems (ICDCS). Piscataway, NJ: IEEE, 2017: 140-150

[38] Krombholz K, Judmayer A, Gusenbauer M, et al. The other side of the coin: User experiences with Bitcoin security and privacy [C] //Proc of Int Conf on Financial Cryptography and Data Security. Berlin: Springer, 2016: 555-580

[39] Kondor D, Pósfai M, Csabai I, et al. Do the rich get richer? An empirical analysis of the Bitcoin transaction network [EB/OL]. [2018-02-08]. http://journals. plos. org/plosone/article?id=10. 1371/journal. pone. 0086197

[40] Badev A, Chen M. Bitcoin: Technical background and data analysis [EB/OL]. [2018-02-08]. https://ssrn. com/abstract=2544331

[41] Lischke M, Fabian B. Analyzing the Bitcoin network: The first four years [EB/OL]. [2018-02-08]. https://doi. org/10. 3390/fi8010007

[42] Tasca P, Hayes A, Liu S. The evolution of the Bitcoin economy: Extracting and analyzing the network of payment relationships [J]. The Journal of Risk Finance, 2018, 19 (2): 94-126

[43] Christin N. Traveling the silk road: A measurement analysis of a large anonymous online marketplace [C] //Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 213-224

[44] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks [EB/OL]. [2018-02-08]. https://arxiv. org/pdf/0803. 0476. pdf

[45] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32

[46] Maesa D D F, Marino A, Ricci L. Uncovering the Bitcoin blockchain: An analysis of the full users graph [C] //Proc of the 3rd IEEE Int Conf on Data Science and Advanced Analytics (DSAA). Piscataway, NJ: IEEE, 2016: 537-546

[47] Maesa D D F, Marino A, Ricci L. An analysis of the Bitcoin users graph: Inferring unusual behaviours [C] //Proc of the 5th Int Workshop on Complex Networks and their Applications. Berlin: Springer, 2016: 749-760

[48] Maesa D D F, Marino A, Ricci L. Data-driven analysis of Bitcoin properties: Exploiting the users graph [J]. IntInternational Journal of Data Science and Analytics, 2018, 6(1): 63-80

[49] Akcora C G, Gel Y R, Kantarcioglu M. Blockchain: A graph primer [J]. arXiv preprint, arXiv:1708. 08749, 2017

[50] Di Battista G, Di Donato V, Patrignani M, et al. Bitconeview: Visualization of flows in the Bitcoin transaction graph [C] //Proc of the 12th IEEE Symp on Visualization for Cyber Security (VizSec). Piscataway, NJ: IEEE, 2015: 1-8

[51] Haslhofer B, Karl R, Filtz E. O Bitcoin Where art thou? Insight into large-scale transaction graphs [EB/OL]. [2018-02-08]. http://ceur-ws. org/Vol-1695/paper20. pdf

[52] McGinn D, Birch D, Akroyd D, et al. Visualizing dynamic Bitcoin transaction patterns [J]. Big Data, 2016, 4(2): 109-119

[53] Bartoletti M, Lande S, Pompianu L, et al. A general framework for blockchain analytics [C] //Proc of the 1st Workshop on Scalable and Resilient Infrastructures for Distributed Ledgers. New York: ACM, 2017: 43-48

[54] Gervais A, Karame G, Capkun S, et al. Is Bitcoin a decentralized currency? [J]. IEEE Security & Privacy, 2014, 12(3): 54-60

[55] Yermack D. Handbook of Digital Currency [M]. San Diego: Academic Press, 2015: 31-43

[56] Dwyer G P. The economics of Bitcoin and similar private digital currencies [J]. Journal of Financial Stability, 2015, 17: 81-91

[57] Grinberg R. Bitcoin: An innovative alternative digital currency [EB/OL]. [2018-02-08]. https://heinonline. org/HOL/LandingPage? handle = hein. journals/hascietlj4&div = 6&id=&page=

[58] Kristoufek L. What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis [EB/OL]. [2018-02-08]. http://journals. plos. org/plosone/article? id = 10. 1371/journal. pone. 0123923

[59] Feder A, Gandal N, Hamrick J T, et al. The impact of DDoS and other security shocks on Bitcoin currency exchanges: Evidence from Mt. Gox [J]. Journal of Cybersecurity, 2018, 3(2): 137-144

[60] Atzei N, Bartoletti M, Cimoli T. A survey of attacks on Ethereum smart contracts (SoK) [C]//Proc of the 6th Int Conf on Principles of Security and Trust. Berlin: Springer, 2017: 164-186

[61] Cheung A, Roca E, Su J J. Crypto-currency bubbles: An application of the Phillips-Shi-Yu (2013) methodology on Mt. Gox Bitcoin prices [J]. Applied Economics, 2015, 47 (23): 2348-2358

[62] Osterrieder J, Strika M, Lorenz J. Bitcoin and cryptocurrencies—Not for the faint-hearted [J]. International Finance and Banking, 2017, 4(1): 56-94

[63] Bouoiyour J, Selmi R. What does Bitcoin look like? [J]. Annals of Economics & Finance, 2015, 16(2): 449-492

[64] Kondor D, Csabai I, Szüle J, et al. Inferring the interplay between network structure and market effects in Bitcoin [J]. New Journal of Physics, 2014, 16(12): 125003

[65] Cherubini U, Luciano E, Vecchiato W. Copula Methods in Finance [M]. New York: John Wiley & Sons, 2004

[66] Gandal N, Hamrick J T, Moore T, et al. Price manipulation in the Bitcoin ecosystem [J]. Journal of Monetary Economics, 2018, 95: 86-96

[67] Maesa D D F, Marino A, Ricci L. Detecting artificial behaviours in the Bitcoin users graph [J]. Online Social Networks and Media, 2017, 3: 63-74

[68] Yelowitz A, Wilson M. Characteristics of Bitcoin users: An analysis of Google search data [J]. Applied Economics Letters, 2015, 22(13): 1030-1036

[69] Moser M, Bohme R, Breuker D. An inquiry into money laundering tools in the Bitcoin ecosystem [C] //Proc of eCrime Researchers Summit (eCRS). Piscataway, NJ: IEEE, 2013: 1-14

[70] Vasek M, Moore T. There's no free lunch, even using Bitcoin: Tracking the popularity and profits of virtual currency scams [C] //Proc of the 19th Int Conf on Financial Cryptography and Data Security. Berlin: Springer, 2015: 44-61

[71] US Attorney's Office. Manhattan US attorney announces charges against liberty reserve [EB/OL]. (2013-05-28) [2018-02-08]. https://www.justice.gov/usao-sdny/pr/manhattan-us-attorney-announces-charges-against-liberty-reserve-one-world-s-largest

[72] Keirns G. "Gemcoin" Ponzi scheme operator hit with MYM74 million judgment [OL]. (2017-05-15) [2018-02-08]. https://www.coindesk.com/gemcoin-ponzi-scheme-operator-hit-74-million-judgment/

[73] Bartoletti M, Carta S, Cimoli T, et al. Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact [EB/OL]. [2018-02-08]. https://arxiv.org/pdf/1703.03779.pdf

[74] Li Yujian, Liu Bo. A normalized Levenshtein distance metric [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1091-1095

[75] Chen Weili, Zheng Zibin, Cui Jiahui, et al. Detecting Ponzi Schemes on Ethereum: Towards healthier blockchain technology [C] //Proc of the 27th Int World Wide Web Conf (WWW '18). New York: ACM, 2018: 1409-1418

Chen Weili, born in 1984. PhD candidate. His main research interests include blockchain, data mining, and financial security.

Zheng Zibin, born in 1982. PhD, professor. Member of CCF. His main research interests include services computing, software engineering, and blockchain.