计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200017
58(2): 237-252, 2021

基于区块链的数据透明化:问题与挑战

孟小峰[1]　　刘立新[1,2]
[1](中国人民大学信息学院　北京　100872)
[2](内蒙古科技大学信息工程学院　内蒙古包头　014010)
(xfmeng@ruc.edu.cn)

Blockchain-Based Data Transparency: Issues and Challenges

Meng Xiaofeng[1] and Liu Lixin[1,2]
[1](*School of Information, Renmin University of China, Beijing* 100872)
[2](*School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou, Inner Mongolia* 014010)

Abstract　With the high-speed development of Internet of things, wearable devices and mobile communication technology, large-scale data continuously generate and converge to multiple data collectors, which influences people's life in many ways. Meanwhile, it also causes more and more severe privacy leaks. Traditional privacy aware mechanisms such as differential privacy, encryption and anonymization are not enough to deal with the serious situation. What is more, the data convergence leads to data monopoly which hinders the realization of the big data value seriously. Besides, tampered data, single point failure in data quality management and so on may cause untrustworthy data-driven decision-making. How to use big data correctly has become an important issue. For those reasons, we propose the data transparency, aiming to provide solution for the correct use of big data. Blockchain originated from digital currency has the characteristics of decentralization, transparency and immutability, and it provides an accountable and secure solution for data transparency. In this paper, we first propose the definition and research dimension of the data transparency from the perspective of big data life cycle, and we also analyze and summary the methods to realize data transparency. Then, we summary the research progress of blockchain-based data transparency. Finally, we analyze the challenges that may arise in the process of blockchain-based data transparency.

Key words　blockchain; accountability; privacy protection; data monopoly; data-driven decision-making

摘　要　物联网、穿戴设备和移动通信等技术的高速发展促使数据源源不断地产生并汇聚至多方数据收集者,由此带来更严峻的隐私泄露问题,然而传统的差分隐私、加密和匿名等隐私保护技术还不足以应对.更进一步,数据的自主汇聚导致数据垄断问题,严重影响了大数据价值实现.此外,大数据决策过程中,数据非真实产生、被篡改和质量管理过程中的单点失败等问题导致数据决策不可信.如何使这些

收稿日期:2020-01-06;**修回日期**:2020-06-28
基金项目:国家自然科学基金项目(91646203,61941121,61532010,91846204,61532016)
This work was supported by the National Natural Science Foundation of China (91646203, 61941121, 61532010, 91846204, 61532016).
通信作者:刘立新(99liulixin@163.com)

问题得到有效治理,使数据被正确和规范地使用是大数据发展面临的主要挑战.首先,提出数据透明化的概念和研究框架,旨在增加大数据价值实现过程的透明性,从而为上述问题提供解决方案.然后,指出数据透明化的实现需求与区块链的特性天然契合,并对目前基于区块链的数据透明化研究现状进行总结.最后,对基于区块链的数据透明化可能面临的挑战进行分析.

关键词 区块链;问责;隐私保护;数据垄断;数据驱动的决策

中图法分类号 TP391

随着大数据技术和人类生产生活的交汇融合,丰富的数据通过多种方式源源不断地被多方数据收集者收集,进而依据这些数据进行数据决策和提供服务.这种先予后取的数据收集模式已成为越来越多应用的必要条件.固然大规模数据收集为个人、企业和国家带来巨大的数据价值,但也带来隐私泄露和决策不可信等问题,表现为大规模数据收集(mass collection)、大规模数据监视(mass surveillance)和大规模数据操纵(mass manipulation)三个方面.

1) 大规模数据收集.大规模数据通过被动、主动和自动方式被收集,如医疗就医、购物、网站搜索、个人移动通信、出行和位置轨迹等数据.然而,作为数据生产者,我们不知道哪些数据被收集、被谁收集、数据被收集后会流向何处以及作何使用,导致隐私泄露追踪问责困难.

2) 大规模数据监视.大规模数据收集导致大规模数据监视,例如医疗就医和个人移动通信等数据被政府部门收集,购物、社交和出行等数据被各大公司掌握.个人在享受服务的同时也时刻处于被监视状态,个人隐私在深度和广度受到巨大冲击.

3) 大规模数据操纵.由于现有政策、技术和制度的不完善,数据战略合作和数据交易等过程中存在大量用户隐私与安全问题.在数据决策过程中,数据非真实产生、数据被篡改、数据质量管理过程中的单点失败等问题导致决策数据不可靠,由此导致数据决策结果不可信[1-2].然而,我们深受数据操纵影响却对此束手无策.

"Facebook-剑桥分析事件"是大规模数据收集、大规模数据监视和大规模数据操纵的典型案例.匿名和差分等传统隐私保护技术主要解决数据发布时的隐私泄露问题,致使其并不能很好地解决当下数据自主汇聚产生的隐私泄露问题.同时,数据决策应用于人类生产生活的方方面面,决策数据不可靠导致的决策不可信是影响大数据进一步发展和应用的重要因素[3].

进一步,数据自主汇聚还导致数据垄断现象出现.数据本身的易聚集特性、大公司覆盖各数字化领域的商业模式和庞大的用户规模等因素加剧数据聚集现象,各公司数据持有量出现差异[4].我们在2019年《中国隐私风险指数分析报告》中对3000万移动用户的权限数据(权限数据是指在移动场景下,某用户安装并使用一系列App,数据收集者通过App的权限体系获取该用户的个人隐私数据)收集情况进行分析,数据收集者获取权限数据的分布如图1所示[5].可以看出前10%的数据收集者获取大于99%的数据,数据垄断现象已悄然形成.数据垄断可能会阻碍市场竞争、使消费者福利受损、阻碍行业技术创新和带来更严重的个人隐私泄露风险等.现实世界财富获取的"二八定律"指20%的人占有80%的社会财富,这依赖于法律、税收等方式的调节.而在虚拟世界,如果将数据比作财富,还是一个没有得到有效调节和分配的领地.因此,急需建立相关技术手段和法律法规.

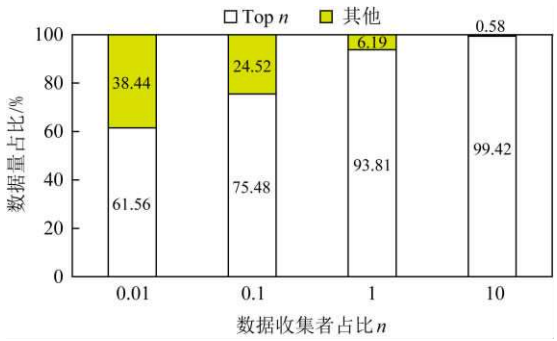

Fig. 1 Data acquisition distribution of the collectors

图1 数据收集者权限数据获取分布

如何使这些问题得到有效治理,使数据得到正确、合理和规范地使用是大数据发展面临的主要挑战.导致这些问题的主要原因是大数据价值实现过程中存在不透明性,数据获取和数据等共享流通过程的不透明性使隐私泄露问题问责困难和数据垄断问题缺乏解决依据,数据决策的不可审计性导致大数据驱动的决策不可信.工业界对大数据价值实现过程的透明性提出迫切需求.苹果CEO库克在2019年

《时代周刊》发表评论建议设立新框架增强企业处理用户数据的透明性,并建议建立数据清算和要求所有数据中介在清算所注册,从而使用户能够跟踪被捆绑并被销售的数据.Gartner 发布的 2020 年战略性技术研究趋势报告中也将"透明性与可追溯性"作为十大战略性技术趋势之一[6].

增加大数据价值实现的透明性,是促进大数据正确使用的重要举措和必经之路.据此,本文提出数据透明性的概念,指在大数据价值实现过程中,各个参与方都能获取与自身相关的全部数据信息.并将数据透明性分为数据获取透明性、数据共享透明性、数据云存储服务透明性、数据决策透明性和法律法规透明性 5 个部分,通过这 5 个部分实现数据透明化.数据透明化需要公开透明地记录数据的获取和共享流通等信息,以及去中心化地管理数据和执行数据质量管理.这些需求与区块链的特性天然契合,而且区块链的去中心和不可篡改特性使数据透明化具有更强的问责能力.

1 数据透明化概述

数据透明化旨在增加大数据价值实现过程的透明性.其研究内容涉及数据生命周期内各阶段,其实现途径主要包括法律法规和技术方法等方面.

1.1 定义与研究框架

文献[7]在 2017 年提出数据透明化概念,并建议从数据透明性策略、日志系统和算法透明性 3 个方面进行实现.但是对数据透明化的研究维度划分没有涵盖大数据生态中的主要透明性需求,也没有深入分析数据透明化与当前大数据生态中的隐私保护、决策可解释和数据垄断关系.

本文提出的数据透明化研究与文献[7]一脉相承,都是保证大数据在其生命周期内各个阶段的透明性.但本文对数据透明化研究的划分更为清晰和具象,进一步将数据透明化研究放在大数据生态范围进行考虑,并阐述数据透明化研究与数据隐私保护、决策可解释和数据垄断的内在关系.

实现数据透明化涉及到大数据生命周期内多方参与主体,各个参与主体有不同的数据透明性需求.目前,参与主体主要包括数据生产者(data contributors)、数据收集者(data collectors)、数据使用者(data consumers)、数据处理者(data processors)和数据监管者(data supervises) 5 个角色.其中,数据生产者是指产生数据的个人或机构;数据收集者是指收集数据的个人或机构,如服务提供者和科研工作者;数据使用者是指任何形式使用数据的个人或机构;数据处理者是指在授权的情况下代替数据使用者处理数据的个人或机构;数据监管者是指对数据生命周期各阶段的数据共享流通等情况进行监管的机构,主要包括政府部门、可信第三方组织等.各参与主体之间可能存在重合,例如当数据收集者自己使用数据并且具有处理能力时,数据收集者也充当数据处理者和数据使用者.

定义 1. 数据透明性.在大数据价值实现过程中,使所有参与主体均能有效获取与自身相关的全部数据信息.其中,数据信息包括原始数据、间接数据和决策数据.

数据透明化研究围绕各方参与主体的数据透明性需求展开,根据大数据生命周期和各方参与主体的透明性需求,将数据透明性分为数据获取透明性、数据共享透明性、数据云存储服务透明性、数据决策透明性和法律法规透明性 5 个部分.通过实现数据获取透明性和数据共享透明性来记录数据获取和共享流通等信息,在隐私泄露和数据滥用等事件发生后进行追踪溯源,并对违反规范的参与方进行问责;通过实现云储存服务透明性增加云存储服务的可信性;通过实现数据决策透明性对决策数据进行审计,从而促进大数据驱动的决策的可信性.数据透明化研究框架和各部分信息如图 2 所示.

1) 数据获取透明性.数据获取透明性指对数据收集内容、形式和使用目的等信息进行记录,数据生产者、数据收集者和数据监管者等能获知相关信息.目前,通过透明增强工具(transparency enhanced tools)、数据使用协议和可审计的访问控制等方式实现获取透明.

2) 数据共享透明性.依据数据共享方式,数据共享透明性可以分为支持溯源问责的数据共享、可验证分布式数据集共享和可验证的分布式机器学习.当发生数据访问和流通时,需要实现支持溯源问责的数据共享,对数据流向进行记录,数据生产者和数据监管者能够据此对数据共享情况和隐私泄露进行追踪问责,数据处理者和数据使用者能据此说明是合法使用数据.当由于传输代价和法律法规等因素限制,需要在不泄露原始数据情况下通过分布式数据集共享技术和分布式机器学习等方式进行数据共享,这时需要对数据提供者(包括数据生产者和数据收集者)提供的加密数据和参数等进行记录,数据使用者可对共享过程进行验证.

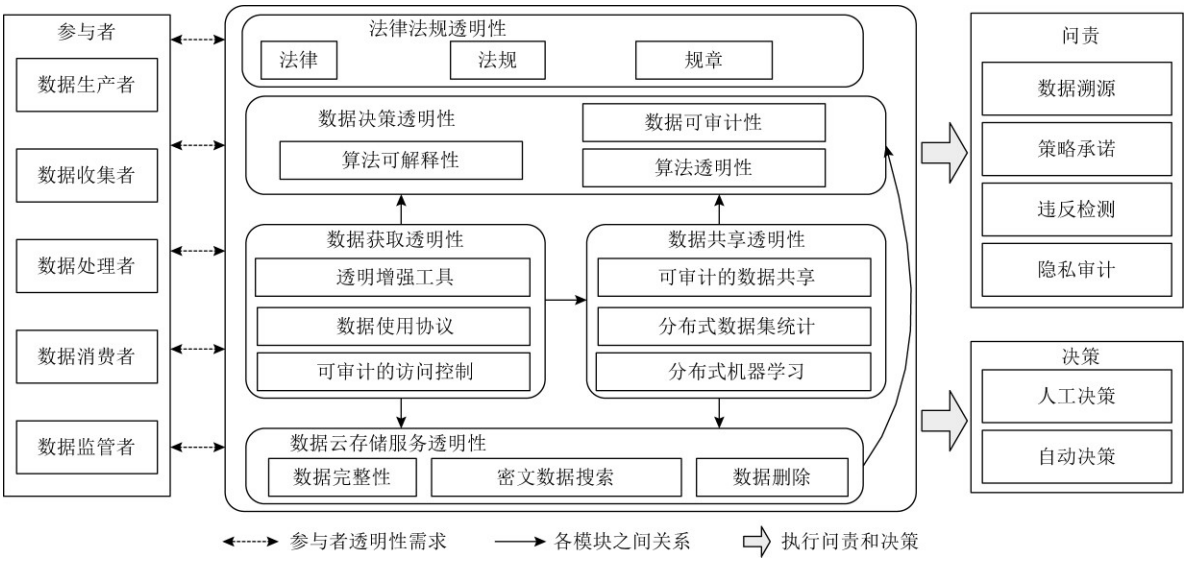

Fig. 2 Dimensions of data transparency

图 2 数据透明化研究框架

3) 数据云存储服务透明性.越来越多的企业和个人将数据存储到云服务器,享受云存储服务带来的便利.然而传统的数据完整性验证、可验证可搜索加密、确定性数据删除等云数据安全和隐私保护技术通常依赖于可信的第三方且实现过程存在不透明性,实现数据云存储服务透明性旨在增加其透明性.

4) 数据决策透明性.数据是决策的基础,所以数据使用者需要对决策数据进行审计和追踪溯源.除此之外,数据决策透明性的实现还需要算法可解释和算法透明的支持①.算法可解释性主要是指机器学习算法的可解释性,即合理解释特定机器学习算法做决策原理以及判断算法是否存在不公平现象.算法透明是指选择合适方式公开决策算法.

5) 法律法规透明性.法律法规是技术之外重要的数据透明化实现手段.世界各国家和组织出台法律法规将知情同意作为个人隐私数据获取、共享、使用和存储等过程的基本要求.知情同意是指数据收集者在收集个人数据之时,应当充分告知有关个人数据被收集、处理和利用的情况,并征得主体明确的同意.例如,欧盟实施的《一般数据保护条例》将透明性作为数据主体的基本权利.

通过上述 5 个部分的数据透明性实现可以将各方参与主体所需要的数据信息作为溯源数据记录下来.之后,可以依据这些溯源数据实施追踪问责和对数据决策进行验证.通常情况下,在问责过程中,需要策略承诺(policy compliance)、违反检测(violation detection)和隐私审计(privacy audit)等支持;在决策过程中,对数据决策验证后,还需要综合考虑数据自动决策和人工决策去获取更加全面的决策结果.

1.2 实现途径

数据透明化需要从法律法规和技术 2 种主要途径进行考虑.法律法规具有威慑和事后惩罚的作用,技术上实现数据透明性能够事先预防和为事后提供依据.

法律法规中数据透明性要求的实现建立在法律法规约束、第三方信用背书和道德自律的基础上.然而,第三方信用背书仅从形式上告知用户数据获取内容、数据共享情况和如何使用用户数据等情况[8].而由于数据获取、数据共享和数据使用等过程对外不可见,其契约履行情况也无从考证.

技术上实现数据透明性为各个参与主体获取与自身相关的数据信息提供技术支持.数据获取透明性和数据共享透明性的实现需要可信的“账本”记录数据获取和共享流通等信息;数据云存储服务透明性和数据决策透明性需要去中心方式执行验证、管理数据和执行质量管理等.数据透明化的这些需求与区块链[9-10]的不可篡改、可追踪、去中心和公开透明的特性相契合.

① 目前,关于算法透明和算法可解释有 2 种理解.一类认为两者是不同的,透明是指算法源代码或者实现原理等内容的公开(public),而可解释则是指向用户解释算法是如何做决策,侧重解释(interpretation)和理解(understanding);另一类认为两者相同,指解释和理解.

总体而言,法律法规和技术2种途径之间既存在互相支持关系也存在互补关系.本文主要探讨技术途径实现数据透明性.

2 数据获取透明性和数据共享透明性

大数据获取形式多样且共享流通错综复杂,对于直接发生数据流通的场景,需要实现支持溯源问责的数据获取和共享,当发生隐私泄露时,数据生产者和数据监管者能够进行溯源问责;对于需要在分布式数据集上实现数据共享和机器学习的场景,除了需要考虑安全和隐私,还需要考虑透明性和可验证性,数据使用者能对其过程进行验证.

2.1 支持溯源问责的数据获取和数据共享

现有关于支持溯源问责的数据获取研究还相对有限,数据未知收集、数据过度收集和用户缺乏控制权等问题有待解决.皮尤研究中心一份关于美国隐私状况的报告指出:91%的受访者认为他们对个人数据被收集和使用已经失去控制,61%的受访者对不了解数据收集者如何使用个人数据感到沮丧[11].文献[12]提出基于区块链管理移动应用程序的权限,通过权限透明管理实现数据的获取透明和支持溯源问责.当用户安装App时,将权限列表存入区块链,数据经加密后存储在分布式散列表(distributed Hash table, DHT),用户发送交易实现权限授予、更新与回收.

现有研究大多基于区块链实现支持溯源问责的数据共享.数据被收集后,由数据收集者存储并通过访问控制等方式与其他第三方进行数据共享.然而大多数访问控制遵循OAuth开放网络标准实现访问授权,由数据收集者作为处理访问控制逻辑的授权引擎,这导致数据共享不支持审计溯源问责.通过访问控制与区块链结合实现数据共享透明可以支持溯源问责,已经应用在物联网[13-15]、医疗[16-17]、社交网络[18]和边缘计算等场景[19-20].

基于区块链实现的访问控制可以概括为"数据获取层—存储层—区块链层—共享层"4层.在数据获取层,数据收集者获取数据生产者产生的数据,需要实现数据获取透明.在存储层,采用传统数据库管理系统、云存储和分布式存储系统等方式存储数据[21],同时为保证数据安全通常需要将数据加密后存储.在区块链层,与传统访问控制模型自主访问控制(discretionary access control, DAC)、强制访问控制(mandatory access control, MAC)、基于属性的访问控制(attribute based access control, ABAC)和基于角色的访问控制(role-based access control, RBAC)等相结合,由区块链执行访问控制,使任何数据访问情况都通过交易被记录在区块链.在共享层,实现数据共享并对共享关系进行保护.

图3为基于区块链实现社交网络数据共享.①~③为服务提供者向用户申请获取数据,④⑤为访问请求与授权,⑥~⑨为区块链执行访问控制.

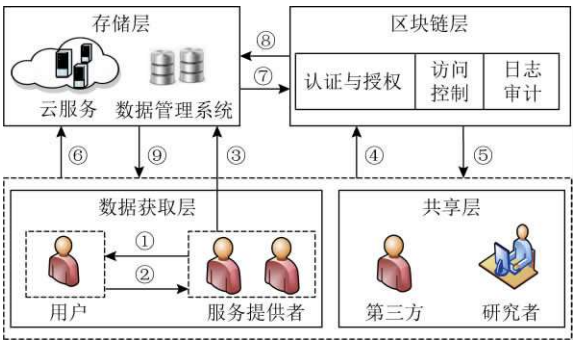

Fig. 3 Data sharing based on blockchain for access control

图3 基于区块链实现社交网络数据共享透明性

基于区块链实现访问控制可分为基于交易和基于智能合约2种方式.基于交易方式是使用区块链的交易对访问控制的策略/权限进行管理.大多方法基于比特币的安全性,应用OP_RETURN指令在比特币上存储策略/权限.由于比特币脚本不适合实现复杂的业务逻辑,所以常结合DAC模型实现访问控制[12].在物联网数据共享场景中,考虑到底层区块链的可扩展性,区块链层之上增加虚链层来提高系统可扩展性[14,22].针对物联网设备计算和存储能力受限,目前有2种解决方法:一种方法是采用RBAC模型的扩展模型OrBAC,引入比特币钱包执行访问控制代理,并通过授权令牌形式管理权限[13];另一种方法是在区块链之下添加边缘设备层,由边缘设备管理设备的身份验证、创建交易、收集和发送数据至存储层[15].

基于智能合约方式是将访问控制策略编写为智能合约,由智能合约自动执行,当前研究尝试与DAC或ABAC等访问控制模型相结合.DAC模型基于身份进行授权,与智能合约结合实现不同身份的用户权限判断透明.文献[16]将策略存储在以太坊智能合约实现分布式医疗数据库共享.但是随着策略规模增加,以太坊智能合约运行成本会增大,且其权限管理不够灵活;考虑到分布式数据库可能存在的安全问题,文献[17]基于Fabric并采用对称

密码加密医疗数据并将其存储在符合法律法规要求的云存储;文献[18]依据 Fabric 实现社交网络数据共享透明;文献[19]实现不同利益相关者边缘设备上数据共享透明,并提出符合边缘设备应用的共识机制、交易类型和区块来适应边缘设备计算和存储能力.DAC 模型与区块链相结合能支持问责,但区块链公开透明性也会泄露共享关系和身份隐私,一定程度上仅依据假名并不能保护用户隐私.ABAC 模型通过属性对实体及约束进行描述,按照访问者权限条件设置属性和权限的关系,将区块链与 ABAC 模型相结合能实现细粒度的、支持身份隐私保护和透明的共享.文献[23]基于区块链解决属性签名时密钥管理问题实现医疗数据共享;文献[24]基于 EbCoin 区块链实现 ABAC 访问控制;文献[25]基于属性签名和密文策略属性加密实现物联网数据共享;文献[26]设计密文策略属性加密实现数据共享.采用 ABAC 模型,策略不会随用户数量呈指数增长,但需权衡问责与隐私保护.

上述方法依据区块链实现访问控制直接进行数据共享流通,可能会带来隐私泄露和数据滥用等问题,例如数据收集者承诺使用数据用于科研,而实际却是用于广告推荐.由区块链执行访问控制,与同态加密、安全多方计算相结合[27-28]实现可控的间接数据共享,可避免上述因数据共享流通带来的问题.此外,数据共享过程中,还可借助区块链实现无需第三方的公平支付,激励数据共享[29-32].

表 1 为基于区块链的数据获取和共享透明方法对比.在区块链层,基于交易的方式多采用比特币;在共享层,大部分方法都不支持共享关系保护[33-37].同时,大部分方法都是实现访问控制,只有文献[32]提出通用的数据共享协议,具有普适性和更广泛应用场景.此外,这些方法都基于现有区块链实现,没有考虑现有区块链可扩展性对实现数据获取与共享透明的影响[38-40].

Table 1 Access Control Methods Based on Blockchain

表 1 基于区块链的数据获取和共享透明性实现方法对比

文献	应用场景	区块链	方式	访问控制类型	存储	加密方式	共享关系保护
[12]	移动应用	Bitcoin	交易	DAC	DHT	对称加密	×
[14]	物联网	Bitcoin like	交易	DAC	普适性存储	对称加密	×
[16]	医疗	Ethereum	智能合约	DAC	分布式数据库	×	×
[17]	医疗	Fabric	智能合约	DAC	云存储	对称加密	×
[18]	社交网络	Fabric	智能合约	DAC	普适性存储	对称加密	×
[19]	边缘计算	Designed One	智能合约	DAC	普适性存储	×	×
[24]	普适性方法	Ebcoin	智能合约	ABAC	分布式数据库	属性加密	√
[32]	普适性方法	Cothority	智能合约	DAC	普适性存储	对称加密	√

注:"√"表示支持;"×"表示不支持.

综上所述,区块链和传统访问控制模式结合从技术上实现数据获取性和数据共享透明性,使数据生产者能控制自己的数据.但是还存在 5 个待解决问题:1)数据获取透明性相关研究仍然有限;2)大量访问控制请求带来区块链存储和可扩展性需求,区块链系统的效率成为亟待解决的重要问题;3)将策略和权限存储在区块链,很容易被攻击者找到漏洞,同时会泄漏共享关系,因此需要有效的方法对其进行保护;4)区块链交易确认时间会影响权限更新的及时性;5)大部分研究只给出理念和系统设计,并未提供具体技术实现方法.

2.2 可验证的分布式数据集共享

在医学研究、公共安全和商业合作等很多领域,限于一些安全和隐私因素并不能直接传输原始数据,需要在分布式数据集上执行统计分析实现数据共享.分布式数据集共享方法如图 4 所示.早期的 PeerDB[41]和混合 P2P 系统[42]等传统分布式数据管理和共享系统并没有考虑隐私和安全.考虑安全和隐私的方法可以分为中心化和去中心化 2 类.中心化方法基于可信的第三方、诚实且好奇的第三方、可信的硬件实现,该类方法的通信代价较低但可能存在单点失败[43-48].去中心化方法主要有秘密共享、安全多方计算和多计算节点等方式.基于秘密共享方式是数据提供者将隐私数据存储在多个服务端并通过秘密共享方式解密数据[49-50],致使数据提供者失去数据控制权.安全多方计算在不泄露数据情况下执行计算,但目前一些安全多方计算的编译库并不支持多于三方参与[51-55].多计算节点方式采用多个

计算节点解决单点失败问题,同时保证数据提供者仍然能控制自己的数据且适用于大规模数据提供者场景.但在实际应用中,数据提供者可能是不可信的,计算节点也可能被攻击或恶意违背执行协议从而导致结果错误,因此需要对数据提供者和计算节点进行验证,增强分布式数据集共享的可验证性.

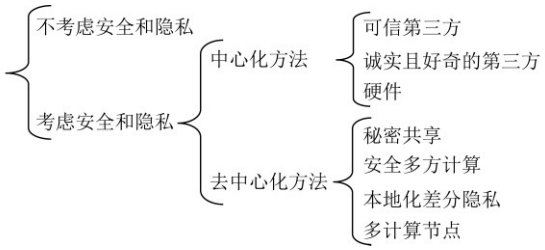

Fig. 4 Ways of distributed data sharing

图 4 分布式数据集共享方法

为实现分布式数据集共享的可验证性,采用区块链或公告牌(bulletin board)公共存储验证信息,并通过零知识证明对数据提供者的输入数据和计算节点计算过程进行验证[56-58].此外,多计算节点共享方式还需考虑数据机密性、数据提供者和数据之间不可连接性、查询结果机密性和计算结果的鲁棒性等安全和隐私问题.文献[57]假设数据提供者是诚实且好奇的,且至少存在一个计算节点是诚实的,但没有将验证信息公开.文献[58]假设数据提供者和计算节点都是恶意的,将区块链作为验证层,实现分布式数据共享.文献[59]假设数据提供者是恶意的,基于公告牌实现去中心的、可验证的在线信誉评价系统.

图 5 将区块链作为验证层,实现多计算节点的分布式数据集共享.数据提供者和计算节点基于零知识证明和密码学承诺(commitment)把证明存入区块链.区块链作为验证层,执行验证并记录共享过程.

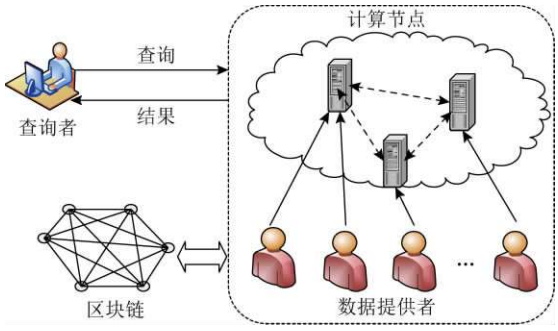

Fig. 5 Verified distributed data sharing system

图 5 可验证分布式数据集共享

综上所述,通过区块链或公告牌作为验证层可增加分布式数据共享的透明性和支持可验证.但是还存在 2 个待解决问题:1)现有方法采用零知识证明和密码学承诺的方法对数据提供者和计算节点进行验证,然而零知识证明生成证明和验证过程都存在较大计算开销;2)现有方法大都依据范围承诺对数据提供者的输入进行验证,适用范围有限.

2.3 可验证的分布式机器学习

分布式机器学习通过数据并行(data parallelism)或模型并行(model parallelism)实现,能间接实现数据共享.目前,分布式机器学习常采用中心化方式,即 1 个主节点(master)和多个参与节点(parties)共同完成机器学习任务.主节点单点失败[60-63]和参与节点投毒攻击(poisoning attack)[64-66]等原因会影响机器学习结果.所以,存储分布式机器学习过程中重要参数信息是必要的,识别哪些节点贡献了哪些参数以及该参数对整个模型的影响[67-70].

基于区块链可以实现可验证的分布式机器学习,由区块链记录和传递重要参数,同时参数传递过程中采用差分隐私、秘密共享和同态加密等技术对参数进行保护.实现方式也分为中心化和去中心化 2 种:中心化方式是指保持固定的主节点和参与节点,由区块链存储机器学习过程中产生的参数[71-72],但仍然存在单点失败;去中心化方式依据区块链共识算法产生主节点,通过区块链交易交换并存储参数信息[73-75].

图 6 为去中心化的分布式机器学习模型.其中,①为各个数据提供者依据本地数据获得本地梯度信息(gradient descent, GD),并通过区块链交易上传至区块链.②为区块链网络中各个矿工交叉验证.③为矿工通过共识算法生成并更新全局梯度信息.

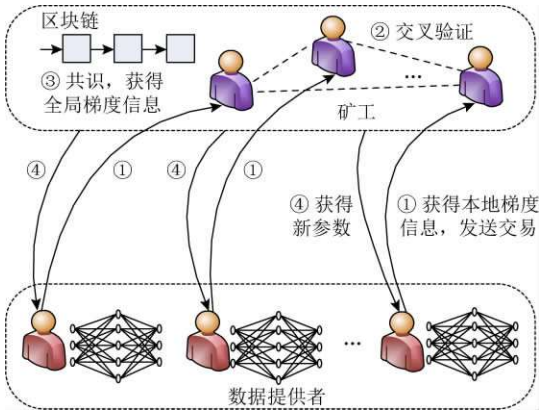

Fig. 6 Machine learning based on blockchain

图 6 基于区块链的机器学习

④为区块链网络再将更新后的全局梯度信息发送给各个数据提供者.重复迭代执行①～④,直至满足要求.

表 2 为基于区块链实现可验证分布式机器学习方法.大多数方法都采用去中心化方式实现,采用差分隐私对数据提供者的本地梯度信息进行保护.而且,文献[58]同时支持分布式数据集统计分析和分布式机器学习.

Table 2 Comparison of Distributed Machine Learning Methods

表 2 可验证分布式机器学习方法对比

文献	验证方式	区块链	隐私保护	目标	支持公平
[58]	去中心化验证	Consortium Blockchain	差分隐私	统计分析,机器学习	×
[66]	去中心化验证	Ethereum	差分隐私	机器学习	×
[71-72]	中心化验证	Designed One	×	机器学习	×
[73]	去中心化验证	Designed One	×	机器学习	×
[74]	去中心化验证	Corda	差分隐私	深度学习	√
[75]	去中心化验证	Consortium Blockchain	×	机器学习	√

注:"√"表示支持;"×"表示不支持.

综上所述,依据区块链实现分布式机器学习,可以增加透明性和支持可验证.但是还存在 3 个待解决问题:1)零知识证明生成证明和验证过程都存在较大计算开销;2)区块链延迟性对分布式机器学习产生影响[76];3)虽然可以通过经济激励和区块链实现公平,如何合理激励数据提供者并解决激励带来的新问题.目前分布式机器学习方法大多假设数据提供者有足够的数据且愿意参加,事实上对数据提供者奖励应该与其数据量多少和数据质量等因素成正比,但这也会促使数据提供者为获得奖励而虚报数据量等问题.

3 数据云存储服务透明性

越来越多的数据拥有者 (data owner, DO)将数据存储至云端,享受云服务提供商(cloud service provider, CSP)提供的云存储服务.由于 DO 和 CSP 之间不存在完全信任,数据完整性验证、可搜索加密和确定性数据删除等是保障云存储数据安全和隐私的重要技术.现有方法大多基于 CSP 是不完全可信、DO 是诚实可信的假设条件,进而引入可信的第三方审计(third party audit, TPA)并支持 DO 实施验证.然而,这些假设条件在实际部署和实施时是有限制的,而且大多数方法实现仍然缺乏透明性.事实上,TPA 也可能会发生错误或合谋,DO 也可能进行欺诈[77],所以需要增加 CSP,TPA,DO 之间交互的透明性和可信性.应用区块链可以在不依赖可信第三方的情况下实现服务透明.此外,依据区块链还可以实现不依赖可信第三方的数据云存储服务公平.

1) 数据完整性验证.数据完整性验证方法有数据持有证明(provable data possession, PDP)和数据可恢复证明(proof of retrievability, POR).PDP 可以快速验证数据是否被云端正确地持有.POR 不仅能够识别数据是否已丢失或损坏,还能对丢失或损坏的数据进行修复.对数据完整性进行验证时,通常依赖 TPA 执行验证,由于验证过程缺乏透明性,DO 只能相信 TPA 返回的验证结果.虽然已有研究通过支持 DO 复审[78-79]、多 TPA 验证[80]、可信硬件[81]解决验证过中 TPA 的不可信和验证过程不透明问题,但是这些方法需要引入其他可信方.区块链与传统完整性验证方法相结合能够增加透明性和可信性,有去中心化验证和中心化验证 2 种方式.去中心化验证是指区块链网络代替 TPA 执行验证.文献[82]结合 PDP 和以太坊实现数据完整性验证,但是并没有考虑如何减少 GAS 开销,文献[83]也采用 PDP 和以太坊实现完整性验证,并实现不依赖第三方的服务公平;文献[84]采用联盟链验证,并设计符合应用场景的共识机制.中心化验证指仍由 TPA 执行数据完整性验证,但将完整性验证挑战信息存入区块链用于日后复审.文献[85]利用区块中 nonce 字段构建完整性验证时的挑战信息,由 DO 对 TPA 验证结果进行复审.这种方式能支持批量处理,提高验证效率,但要求 DO 具有一定的计算能力执行复审.

2) 可搜索加密.可搜索加密技术,根据实现功能不同可以分为单关键词搜索、连接关键词搜索和复杂逻辑结构搜索;根据构造算法不同可以分为对称可搜索加密(symmetric searchable encryption, SSE)和非对称可搜索加密(asymmetric searchable

encryption, ASE)[86].可搜索加密结果完整性验证方法大多都假设可信的 TPA 执行公共验证,缺乏透明性.区块链与传统可搜索加密方法相结合能够增加透明性和可信性,可分为去中心化搜索和中心化搜索 2 种方式.去中心化搜索时,由区块链网络中各节点通过执行智能合约代替 CSP 执行搜索,共识过程保证搜索结果是正确的,不需要数据拥有者对搜索结果进行验证[87].中心化搜索指仍然由 CSP 执行搜索,在给 DO 返回搜索结果的同时将验证信息存入区块链[88].此外,除了传统中心云存储,结合区块链还可以实现 Storj 和 Filecoin 等去中心云存储关键字搜索结果完整性验证[89-90].

3) 确定性数据删除.确定性数据删除方法有覆盖写删除(deletion by overwriting)和密码学删除(deletion by cryptography).当进行确定性数据删除时,DO 发出删除请求之后,CSP 执行删除操作并返回 1 位的"成功"或"失败"作为响应.DO 无法根据此响应来确定云端数据是否已经被删除,删除过程亦缺乏透明性.已有研究依赖于用户能访问存储介质[91]、沙漏模型[92]等假设条件,或者基于可信硬件[93]和可信第三方[94]实现可验证确定性数据删除,但仍缺乏透明性.由区块链记录删除证明可以增加数据确定性删除的透明性.在执行数据删除时仍由 CSP 执行删除,基于信任但可验证原则(trust-but-verify),将 DO 的删除请求和 CSP 的删除证明存入区块链.任何人都可以依据区块链执行验证操作,增加删除透明性,防止 DO 和 CSP 双方都可能存在的恶意行为.文献[95]采用覆盖写方法,假设 DO 和 CSP 之间已通过身份验证实现问责,并引入时间服务器为删除证明提供时间戳服务.文献[96]在此之上使用基于属性签名代替比特币中采用椭圆曲线数字签名增加隐私性和安全性,并将交易内容加密防止窃听攻击.

表 3 为基于区块链增加云存储服务透明性和公平性研究总结.数据完整性验证多采用去中化方法,仅支持 DO 验证.可搜索加密技术都考虑了公平性.确定性数据删除都采用去中心化方式和公有链实现,支持 DO 和 CSP 双方验证.

Table 3 Comparison of Cloud Storage Services Transparency

表 3 数据云存储服务透明性实现方法对比

文献	目标	验证方式	区块链	公平	DO 验证	CSP 验证
[82]	数据完整性验证	去中心化验证	Ethereum	×	√	×
[83]	数据完整性验证	去中心化验证	Bitcoin	√	√	×
[84]	数据完整性验证	去中心化验证	Consortium Blockchain	×	√	×
[85]	数据完整性验证	中心化验证	Ethereum	×	√	×
[87]	可搜索加密	去中心化验证	Ethereum	√	√	×
[88]	可搜索加密	中心化验证	Public Blockchain	√	√	√
[95]	数据删除	去中心化验证	Public Blockchain	×	√	√
[96]	数据删除	去中心化验证	Public Blockchain	×	√	√

注:"√"表示支持;"×"表示不支持.

综上所述,应用区块链可增加数据云存储服务的透明性和公平性.但是还存在 3 个待解决问题:1)数据完整性验证,采用中心化验证方式仍然需要 DO 执行复审,增加 DO 的计算负担;采用去中心化方式,由于以太坊智能合约的执行需要消耗燃料(gas),燃料需要通过以太币进行购买,所以需要尽力优化实现代码等方式减少代价消耗.2)可搜索加密技术,一般来说密文和索引都有可能造成不同程度的信息泄露,需要设计更安全的模型使陷门和索引都不泄露关键词信息;采用中心化搜索方式需要实现可验证的多关键词和复杂逻辑结构搜索.此外,可搜索加密技术也存在以太坊计算燃料消耗问题.3)确定性数据删除,现有方法均基于覆盖写删除方法,依赖于 DO 事后发现数据仍然存在的假设,需要设计可验证的即时数据删除方法.此外,如果数据拥有者和云服务提供者要求将区块链上信息也同时删除,此时应该考虑链上数据删除技术[97-99].

4 数据决策透明性

在基于"数据—信息—知识—智慧"模型的数据决策过程中[100],首先需要收集数据,并对其加工处理之后形成对决策有价值的信息,进一步对信息使用归纳、演绎方法得到知识,最后利用这些知识并

经由探讨得出最终决策.然而,在大数据环境下,此模型的有效性受到冲击.数据被篡改、数据质量管理过程中的单点失败等问题会导致决策数据不可靠;训练数据偏见、算法设计偏见和算法错误都可能导致决策算法不可靠.为此,数据决策透明性需要实现决策数据可审计、算法可解释[101-105]和算法透明.

区块链作为去中心化的分布式数据库,为决策数据可审计提供支持.通过获取透明、共享透明和服务透明,在对数据进行追踪溯源的同时也为数据使用者对决策数据进行审计有促进作用.此外,基于区块链的去中心化存储模式,数据使用者可以验证数据是否被篡改和对数据进行追踪,在金融保险[106]、医疗[107-110]和供应链[111-114]等数据完整性要求较高领域有重要意义.区块链作为分布式数据库,区块链的可扩展性[115]、安全[116]]和隐私[117]等问题是影响其应用的重要因素.此外,考虑到区块链存储限制,通常采用"链上"存储元数据与"链下"存储数据相结合的方式,并进一步在这些可信数据上执行查询分析.大部分区块链查询系统仅提供区块、交易和账户等信息的简单查询,并未提供复杂查询功能.实际应用中还需要实现范围查询和 Top-k 查询等复杂查询[118]、数据查询完整性验证[119]、密文查询[120]和细粒度在线查询溯源[121]等.

多源数据的格式、标准不统一等问题也会影响数据质量,进而影响数据决策.然而传统数据质量管理和质量控制方法通常依赖可信第三方执行,存在缺乏透明性、单点失败和时间资源消耗较大的问题.依靠智能合约自动执行可以制定统一数据格式、规则来提高数据质量管控的透明度[122-123].

综上所述,基于区块链可以促进决策数据可审计,进而有助于决策可解释.但是还存在 3 个待解决问题:1)大数据来源广泛,虽然采用区块链存储和管理数据可以实现数据追踪问责,但是如何保证数据在存入区块链之前的真实可信是挑战问题;2)支持区块链上复杂数据查询、查询隐私保护和密文数据查询等;3)如何保证"链下"存储数据的安全性.

5 挑战问题

基于区块链的数据透明化旨在增加大数据价值实现过程的透明性,记录数据获取、数据共享和数据使用等信息.进而依据这些信息实现具有不可篡改性质的溯源问责和数据在其生命周期内的可审计,为隐私保护和数据决策可审计提供支持.数据透明性、溯源问责和数据可审计的实现主要面临 5 个挑战问题:

1) 符合数据透明化需求的区块链架构问题.基于区块链的数据透明化具有更强的问责能力,但现有区块链的技术和系统无法被直接应用于数据透明化.例如,实现数据获取透明性和数据共享透明性对区块链的可扩展性提出较高要求;实施溯源需要涉及多区块链之间的互操作性;实施问责与现有公有链的监管困难相冲突.为此,需要设计符合数据透明化需求的高可扩展性、隐私与监管并重、轻量级的区块链,而非完全依赖于现有的、开源的区块链平台.

2) 具有用户控制权的数据获取透明性问题.目前的数据获取过程缺乏透明性和用户控制权,导致隐私泄露问题严峻.然而,目前关于数据获取透明性的研究仍然相对有限,亟需一种全新的数据获取架构以及政策和法律法规的支持,实现数据获取透明性.此外,用户(即数据生产者)在数据获取过程中缺少控制权,用户或者同意数据收集者制定的数据协议而付出所有数据收集者要求的数据,或者不同意但会导致不能享受服务.在数据获取透明性实现过程中,如何将控制权还给用户,由用户决定数据内容、目的和形式,并根据用户同意的数据提供服务是挑战问题.由此,用户所获得的服务与自身数据隐私损失之间的平衡也至关重要.

3) 保证数据使用协议的数据共享透明性问题.服务提供者(service provider)作为数据收集者收集用户数据并为用户提供服务,但服务提供者是否依据数据使用协议执行数据共享是不透明的.为此如何实现强制执行数据使用协议进行共享并对数据共享情况进行透明记录是一个挑战问题.此外,数据开放共享平台是重要的数据共享流通方式,可以促进不同领域资源相融合,使数据发挥更大价值.例如,政府公共部门数据共享开放可以促进更智能高效的服务和据此为公共问题提出有效的方案.但是数据开放共享平台的数据可能会涉及众多个人隐私,原则上数据开放共享需要征求个人同意,但实现难度较大且可能会导致数据出现偏差.为此,在数据提供者和数据使用者能够保护数据义务的前提下,可以考虑个人同意让位于集体公共利益,在不经过个人同意情况下实现共享.那么,针对这种"捆绑"数据共享流通情况,如何在隐私保护前提下实现数据共享透明性,并使用户能追踪与他们数据有关的共享流通信息也是一个挑战问题.

4) 具有不可篡改性质的溯源问责问题.通过获取透明和共享透明可以获得溯源数据,依据这些溯源数据可以实现溯源问责.溯源问责的前提是溯源数据的完备性,然而如何使所有的数据获取和共享事件都被记录是一个挑战问题.除技术手段,还需要政策、法律法规等多方面的支持.例如,可采用激励等非技术手段,将记录数据获取和共享信息与企业信誉相关联,主动记录数据获取和共享信息的企业获得较高的信誉,增加用户的信任,利于其业务发展.进一步,在大规模数据收集和数据共享流通错综复杂背景下,如何实现跨平台和跨领域的溯源问责是仍未解决的挑战问题.同时,由于溯源数据描述数据获取和共享流通整个脉络,在数据溯源过程中也可能会泄露其他隐私信息,所以溯源过程的隐私保护也至关重要.进一步,如何根据策略承诺和溯源数据自动进行违反检测也是一个挑战问题.

5) 保证数据在其数据周期内的可审计问题.在数据生命周期内,数据是否真实产生和处理、数据在共享流通过程中是否被篡改等问题都会影响数据决策结果.虽然基于区块链进行去中心化存储和管理数据会使数据使用者能够对数据完整性进行验证和追溯,但并不能防止数据在存入区块链之前数据被伪造和篡改等问题.此外,为保证决策结果可解释性,应该保证数据的准确性.然而数据隐私保护技术会在某种程度上扰动数据,必然会造成数据准确性降低,并影响决策数据的可解释性.如何平衡数据隐私保护和决策数据可审计是一个挑战问题.

6 总　　结

如何保证数据得到正确、合理和规范的使用已经成为大数据生态中亟待解决的根本问题,建立数据透明化的治理体系是有效途径和重要举措.本文提出数据透明化研究框架,并总结和分析该框架下的基于区块链的数据透明化研究现状,最后提出主要面临的挑战问题.此外,作为一个跨学科问题,数据透明化将数据获取和共享流通置于新的范式之下,如何确保用户具备足够的法律法规素养来理解和应对这种变化,也是需要学界和全社会共同去探索的课题.于此同时,我们更要遵从“管理数据、理解数据、敬畏数据”的理念,从而促进大数据生态良性发展.

参 考 文 献

[1] Aber K. How computer science risks to lose its innocence and should attempt to take responsibility [C] //Proc of the 37th Int Conf on Distributed Computing Systems. Piscataway, NJ: IEEE, 2017: 1873-1875

[2] Julia S, Serge A, Gerome M. Dataresponsibly: Fairness, neutrality, and transparency in data analysis [C] //Proc of the 19th Int Conf on Extending Database Technology. Berlin: Springer, 2016: 718-719

[3] O'Neil C. Weapons of Math Destruction [M]. New York: Penguin Random House, 2016: 20-98

[4] Meng Xiaofeng, Zhu Minjie, Liu Lixin. Research on data monopoly and its governance modes [J]. Journal of Information Security Research, 2019, 5(9): 789-797 (in Chinese)
(孟小峰, 朱敏杰, 刘立新. 数据垄断与其治理模式研究[J]. 信息安全研究, 2019, 5(9): 789-797)

[5] Meng Xiaofeng, Zhu Minjie, Liu Junxu, et al. China's privacy risk index [EB/OL]. 2019 [2020-03-30]. http://idke.ruc.edu.cn/reports/report2018_cn.htm

[6] High P. Gartner announces top 10 strategic technology trends for 2020 [EB/OL]. (2019-10-21) [2020-03-30]. https://www.forbes.com/sites/peterhigh/2019/10/21/breaking-gartner-announces-top-10-strategic-technology-trends-for-2020/#21c3ea044074

[7] Elisa B. Big data security and privacy and transparency [C] //Proc of the 37th Int Conf on Distributed Computing Systems. Piscataway, NJ: IEEE, 2017: 1180-1183

[8] Janice M, Veugen T, Wijbenga J. Transparency enhancing tools (TETs): An overview [C] //Proc of the Workshop on Socio-Technical Aspects in Security and Trust. Piscataway, NJ: IEEE, 2013:18-25

[9] Meng Xiaofeng, Zhang Xiaojian. Big data privacy management [J]. Journal of Computer Research and Development, 2015, 52(2): 265-281 (in Chinese)
(孟小峰, 张啸剑. 大数据隐私管理[J]. 计算机研究与发展, 2015, 52(2): 265-281)

[10] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system [EB/OL]. (2018-06-02) [2020-03-30]. https://bitcoin.org/bitcoin.pdf

[11] Pew Research Center. Engineering privacy by design: Are engineers ready to live up to the challenge? [EB/OL]. (2014-11-12) [2020-03-30]. https://www.tandfonline.com/doi/full/10.1080/01972243.2019.1583296

[12] Zyskind G, Nathan O, Pentland A. Decentralizing privacy: Using blockchain to protect personal data [C] //Proc of IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2015: 180-184

[13] Ouaddah A, Abou E, Ait O, et al. FairAccess: A new blockchain-based access control framework for the Internet of things [J]. Security and Communication Networks, 2016, 9(18): 5943-5964

[14] Hossein S, Lukas B, Simon D. Droplet: Decentralized authorization for IoT data streams [J]. arXiv preprint, arXiv: 1806.02057, 2018

[15] Li Ruinian, Song Tianyi, Mei Bo, et al. Blockchain for large-scale Internet of things data storage and protection [J]. IEEE Transactions on Services Computing, 2018, 12(5): 762-771

[16] Azaria A, Ekblaw A, Vieira T, et al. MedRec: Using blockchain for medical data access and permission management [C] //Proc of the Int Conf on Open & Big Data. Piscataway, NJ: IEEE, 2016: 25-30

[17] Dubovitskaya A, Xu Zhigang, Ryu S, et al. Secure and trustable electronic medical records sharing using blockchain [J]. American Medical Informatics Association, 2017, 11(6): 650-659

[18] Truong N, Sun Kai, Lee G, et al. GDPR-compliant personal data management: A blockchain-based solution [J]. arXiv preprint, arXiv:1904.03038, 2019

[19] Xu Chenhan, Wang Kun, Li Peng, et al. Making big data open in edges: A resource-efficient blockchain-based approach [J]. IEEE Transactions on Parallel and Distributed Systems, 2018, 30(4): 870-882

[20] Yang Ruizhe, Yu F R, Si Pengbo, et al. Integrated blockchain and edge computing systems: A survey, some research issues and challenges [J]. IEEE Communications Surveys & Tutorials, 2019, 21(2): 1508-1532

[21] Benet J. IPFS-content addressed, versioned, P2P file system [J]. arXiv preprint, arXiv:1407.3561, 2014

[22] Muneeb A, Jude C, Nelson R, et al. Blockstack: A global naming and storage system secured by blockchains [C] //Proc of the USENIX Annual Technical Conf. Berkeley, CA: USENIX Association, 2016: 181-194

[23] Guo Rui, Shi Huixian, Zhao Qinlan, et al. Secure attribute-based signature scheme with multiple authorities for blockchain in electronic health records systems [J]. IEEE Access, 2018, 6: 11676-11686

[24] Liu Aodi, Du Xuehui, Wang Na, et al. A blockchain-based access control mechanism for big data [J]. Journal of Software, 2019, 30(9): 2636-2654 (in Chinese)
(刘敖迪, 杜学绘, 王娜, 等. 基于区块链的大数据访问控制机制[J].软件学报, 2019, 30(9): 2636-2654)

[25] Zhang Yunru, He Debiao, Kim-Kwang R, et al. BaDS: Blockchain-based architecture for data sharing with ABS and CP-ABE in IoT [J]. Wireless Communications and Mobile Computing, 2018, 2783658: 1-9

[26] Yuan Chao, Xu Mixue, Si Xueming, et al. Blockchain with accountable CP-ABE: How to effectively protect the electronic documents [C] //Proc of the 23rd Int Conf on Parallel and Distributed Systems. Piscataway, NJ: IEEE, 2017: 800-803

[27] Zhou Lijing, Wang Licheng, Sun Yiru. BeeKeeper: A blockchain-based IoT system with secure storage and homomorphic computation [J]. IEEE Access, 2018, 6: 43472-43488

[28] Xu Wenyu, Wu Lei, Yan Yunxue. Privacy-preserving scheme of electronic health records based on blockchain and homomorphic encryption [J]. Journal of Computer Research and Development, 2018, 55(10): 2233-2243 (in Chinese)
(徐文玉, 吴磊, 阎允雪. 基于区块链和同态加密的电子健康记录隐私保护方案[J]. 计算机研究与发展, 2018, 55(10): 2233-2243)

[29] Andrychowicz M, Dziembowski S, Malinowski D, et al. Secure multiparty computations on bitcoin [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2014: 443-458

[30] Zhou Jiayu, Tang Fengyi, Zhu He. Distributed data vending on blockchain [J]. arXiv preprint, arXiv:1803.05871, 2018

[31] Dziembowski S, Eckey L, Faust S. Fairswap: How to fairly exchange digital goods [C] //Proc of the ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 967-984

[32] Eleftherios K, Enis C, Sandra D, et al. Hidden in plain sight: Storing and managing secrets on a public ledger [EB/OL]. IACR Cryptology ePrintArchive. 2019 [2020-03-30]. https://pdfs.semanticscholar.org/9c3d/3e64ffead7ea93cd21ab7ab89fc9afcd690c.pdf

[33] Chen Wei, David S, Laura B, et al. TD-CHAIN: A system to enhance transparency in data flows [C] //Proc of the Information Systems Security and Privacy. Boston: AIS, 2017: 1-2

[34] Zhang Rui, Xue Rui, Liu Ling. security and privacy on blockchain [J]. ACM Computing Surveys, 2019, 52(3): 51: 1-51:34

[35] Luu L, Chu D, Olickel H, et al. Makingsmart contracts smarter [C] //Proc of the 22nd ACM SIGSAC Int Conf Computer and Communications Security. New York: ACM, 2016: 254-269

[36] Kosba A, Miller A, Shi E, et al. Hawk: The blockchain model of cryptography and privacy-preserving smart contracts [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2016: 839-858

[37] Cheng R, Zhang Fan, Kos J, et al. Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contract execution [J]. arXiv preprint, arXiv: 1804.05141, 2018

[38] Dai Mingjun, Zhang Shengli, Wang Hui, et al. A low storage room requirement framework for distributed ledger in blockchain [J]. IEEE Access, 2018, 6: 22970-22975

[39] Eyal I, Gencer A, Sirer E. Bitcoin-NG: A scalable blockchain protocol [C] //Proc of the 13th Symp on Networked Systems Design and Implementation. Berkeley CA: USENIX Association, 2016: 45-59

[40] Ren Zhijie, Cong Kelong, Pouwelse J. Implicit consensus: Blockchain with unbounded throughput [J]. arXiv preprint, arXiv: 1705.11046, 2017

[41] Ng W, Ooi B, Tan K, et al. PeerDB: A P2P-based system for distributed data sharing [C] //Proc of the 19th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2003: 633-644

[42] Yang Min, Yang Yuanyuan. An efficient hybrid peer-to-peer system for distributed data sharing [J]. IEEE Transactions on Computers, 2010, 59(9): 1158-1171

[43] Ohrimenko O, Schuster F, Fournet C, et al. Oblivious multi-party machine learning on trusted processors [C] // Proc of the 25th USENIX Security Symp. Berkeley CA: USENIX Association, 2016: 619-636

[44] Schuster F, Costa M, Fournet C, et al. VC3: Trustworthy data analytics in the cloud using SGX [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2015: 38-54

[45] Jagadeesh K, Wu D, Birgmeier J, et al. Deriving genomic diagnoses without revealing patient genomes [J]. Science, 2017, 357(6352): 692-695

[46] Luca M. Efficient private statistics with succinct sketches [J]. arXiv preprint, arXiv:1508.06110, 2016

[47] Raisaro J, Troncoso J, Misbach M, et al. Medco: Enabling secure and privacy-preserving exploration of distributed clinical and genomic data [J]. Transactions on Computational Biology and Bioinformatics, 2019, 16(4): 1328-1341

[48] Bater J, Elliott G, Eggen C, et al. SMCQL: Secure querying for federated databases [J]. Proceedings of the VLDB Endowment, 2016, 10(6): 673-684

[49] Bogdanov D, Laur S, Willemson J. Sharemind: A framework for fast privacy-preserving computations [C] //Proc of the European Symp on Research in Computer Security. Berlin: Springer, 2008: 192-206

[50] Fabian B, Ermakova T, Junghanns P. Collaborative and secure sharing of healthcare data in multi-clouds [J]. Information Systems, 2015, 48: 132-150

[51] Mahdi Z, Mahnush M, Jared S. Millions of millionaires: Multiparty computation in large networks [EB/OL]. IACR Cryptology ePrint Archive. 2014 [2020-03-30]. https://eprint.iacr.org/2014/149

[52] Bellare M, Hoang V T, Keelveedhi S, et al. Efficient garbling from a fixed-key blockcipher [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2013: 478-492

[53] Chang Liu, Wang Xiaoshaun, Nayak K, et al. ObliVM: A programming framework for secure computation [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2015: 359-376

[54] Nayak K, Wang Xiaoshaun, Ioannidis S, et al. GraphSC: Parallel secure computation made easy [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2015: 377-394

[55] Songhori E M, Hussain S U, Sadeghi A R, et al. TinyGarble: Highly compressed and scalable sequential garbled circuits [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2015: 411-428

[56] Corrigan-Gibbs H, Boneh D. Prio: Private, robust, and scalable computation of aggregate statistics [C] //Proc of the 14th USENIX Symp on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2017: 259-282

[57] Froelicher D, Egger P, João S, et al. UnLynx: A decentralized system for privacy-conscious data sharing [C] // Proc of the Privacy Enhancing Technologies. Piscataway, NJ: IEEE, 2017: 232-250

[58] Froelicher D, Troncoso-Pastoriza J, João S, et al. Drynx: Decentralized, secure, verifiable system for statistical queries and machine learning on distributed datasets [J]. arXiv preprint, arXiv:1902.03785, 2019

[59] Ajmal A M, Samiran B, Feng H. PrivBox: Verifiable decentralized reputation system for the on-line marketplaces [J]. Future Generation Computer Systems, 2018, 89(12): 315-329

[60] Song Congzheng, Ristenpart T, Shmatikov V. Machine learning models that remember too much [C] //Proc of the 30th ACM SIGSAC Int Conf on Computer and Communications Security. New York: ACM, 2017: 587-601

[61] Melis L, Song Congzheng, Cristofaro E, et al. Inference attacks against collaborative learning [J]. arXiv preprint, arXiv: 1805.04049, 2018

[62] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the gan: Information leakage from collaborative deep learning [C] //Proc of the 30th ACM SIGSAC Int Conf on Computer and Communications Security. New York: ACM, 2017: 603-618

[63] Aono Y, Hayashi T, Wang Lihua, et al. Privacy-preserving deep learning via additively homomorphic encryption [J]. Transactions on Information Forensics and Security, 2018, 13(5): 1333-1345

[64] Shokri R, Shmatikov V. Privacy-preserving deep learning [C] //Proc of the 22nd ACM SIGSAC Int Conf on Computer and Communications Security. New York: ACM, 2015: 909-910

[65] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for privacy-preserving machine learning [C] // Proc of the 24th ACM SIGSAC Int Conf on Computer and Communications Security. New York: ACM, 2017: 1175-1191

[66] Chen Xuhui, Ji Jinlong, Luo Changqing, et al. When machine learning meets blockchain: A decentralized, privacy-preserving and secure design [C] //Proc of the Int Conf on Big Data. Piscataway, NJ: IEEE, 2018: 1178-1187

[67] Eugene B, Andreas V, Hua Y, et al. How to backdoor federated learning [J]. arXiv preprint, arXiv: 1807.00459, 2018

[68] Anusha L, Osman C, Tara J, et al. Peer-to-peer federated learning on graphs [J]. arXiv preprint, arXiv:1901.11173, 2019

[69] Jatinder S. Decision provenance: Capturing data flow for accountable systems [J]. arXiv preprint, arXiv:1804.05741, 2018

[70] Verma D, Calo S, Cirincione G. Distributed AI and security issues in federated environments [C] //Proc of the Int Conf on Distributed Computing and Networking, Piscataway, NJ: IEEE, 2018: 4:1-4:6

[71] BoreN, Raman R, Markus I, et al. Promoting distributed trust in machine learning and computational simulation via a blockchain network [J]. arXiv preprint, arXiv:1810.11126, 2019

[72] Ravi K, Roman V, Michael H, et al. Trusted multi-party computation and verifiable simulations: A scalable blockchain approach [J]. arXiv preprint, arXiv:1809.08438, 2018

[73] Tsung T, Lucila O. ModelChain: Decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks [J]. arXiv preprint, arXiv:1802.01746, 2018

[74] Weng Jiasi, Weng Jian, Li Ming, et al. Deepchain: Auditable and privacy-preserving deep learning with blockchain-based incentive [EB/OL]. Cryptology ePrint Archive, 2018 [2020-03-30]. https://eprint.iacr.org/2018/679

[75] Kuo T, Rodney A, Lucila O, et al. Fair compute loads enabled by blockchain: Sharing models by alternating client and server roles [J]. Journal of the American Medical Informatics Association, 2019, 26(5): 392-403

[76] Kim H, Park J. On-device federated learning via blockchain and its latency analysis [J]. arXiv preprint, arXiv: 1808.03949, 2018

[77] Wang Cong, Ren Kui, Lou Wenjing, et al. Toward publicly auditable secure cloud data storage services [J]. IEEE Network, 2010, 24(4): 19-24

[78] Xu Jia. Auditing the auditor: Secure delegation of auditing operation over cloud storage [EB/OL]. IACR Cryptology ePrint Archive. 2011 [2020-03-30]. https://eprint.iacr.org/2011/304

[79] Wu Yilun, Lin Xinye, Lu Xicheng, et al. A secure light-weight public auditing scheme in cloud computing with potentially malicious third party auditor [J]. Transactions Information System, 2016, 99(10): 2638-2642

[80] Huang Kun, Xian Ming, Fu Shaojing, et al. Securing the cloud storage audit service: Defending against frame and collude attacks of third-party auditor [J]. IET Communications, 2014, 8(12): 2106-2113

[81] Xiao Da, Yang Lüyin, Sun Bin, et al. Provable data possession system for realistic cloud storage environments [J]. Journal of Software, 2016, 27(9): 2400-2413 (in Chinese)
(肖达, 杨绿茵, 孙斌, 等. 面向真实云存储环境的数据持有性证明系统[J]. 软件学报, 2016, 27(9): 2400-2413)

[82] Hao Kun, Xin Junchang, Wang Zhiqiong, et al. Decentralized data integrity verification model in untrusted environment [C] //Proc of the Int Conf Asia-Pacific Web. Berlin: Springer, 2018: 410-424

[83] Zhang Yinghui, Deng R, Liu Ximeng, et al. Blockchain based efficient and robust fair payment for outsourcing services in cloud computing [J]. Information Sciences, 2018, 462: 262-277

[84] Tian Junfeng, Li Tianle. Data integrity verification based on model cloud federation of TPA [J]. Journal on Communications, 2018, 39(8): 113-124 (in Chinese)
(田俊峰, 李天乐. 基于TPA云联盟的数据完整性验证模型[J]. 通信学报, 2018, 39(8): 113-124)

[85] Xue Jingting, Xu Chunxiang, Zhao Jining, et al. Identity-based public auditing for cloud storage systems against malicious auditors via blockchain [J]. SCIENCE CHINA: Information Sciences, 2019,62(3): 32104-32110

[86] Dong Xiaolei, Zhou Jun, Cao Zhenfu. Research advances on secure searchable encryption [J]. Journal of Computer Research and Development, 2017, 54(10): 2107-2120 (in Chinese)
(董晓蕾, 周俊, 曹珍富. 可搜索加密研究进展[J]. 计算机研究与发展, 2017, 54(10): 2107-2120)

[87] Hu Shengshan, Cai Chengjun, Wang Qiang, et al. Searching an encrypted cloud meets blockchain: A decentralized, reliable and fair realization [C] //Proc of the IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2018: 792-800

[88] Zhang Yinghui, Robert H, Shu Jianggang et al. TKSE: Trustworthy keyword search over encrypted data with two-side verifiability via blockchain [J]. IEEE Access, 2018, 6: 31077-31087

[89] Cai Chengjun, Yuan Xingliang, Wang Cong. Towards trustworthy and private keyword search in encrypted decentralized storage [C] //Proc of the Int Conf on Communications. Piscataway, NJ: IEEE, 2017: 1-7

[90] Cai Chengjun, Yuan Xingliang, Wang Cong. Hardening distributed and encrypted keyword search via blockchain [C] //Proc of the IEEE Symp on Privacy-Aware Computing. Piscataway, NJ: IEEE, 2017: 119-128

[91] Paul M, Saxena A. Proof of erasability for ensuring comprehensive data deletion in cloud computing [C] //Recent Trends in Network Security and Applications. Berlin: Springer, 2010: 340-348

[92] Gilberg O. Cloud security without trust [D]. Norway: Norwegian University of Science and Technology, 2014

[93] Hao Feng, Dylan C, Avelino F. Deleting secret data with public verifiability [J]. IEEE Transactions on Dependable and Secure Computing, 2016, 13(6): 617-629

[94] Liang Xue, Yu Yong, Li Yannan, et al. Efficient attribute-based encryption with attribute revocation for assured data deletion [J]. Information Sciences, 2019, 479: 640-650

[95] Yang Changsong, Chen Xiaofeng, Yang Xiang. Blockchain-based publicly verifiable data deletion scheme for cloud storage [J]. Journal of Network & Computer Applications, 2018, 103: 185-193

[96] Liu Yining, Zhou Yuanjian, Lan Rushi, et al. Blockchain-based verification scheme for deletion operation in cloud [J]. Journal of Computer Research and Development, 2018, 55(10): 2199-2207 (in Chinese)
(刘忆宁, 周元健, 蓝如师, 等. 基于区块链的云数据删除验证协议[J]. 计算机研究与发展, 2018, 55(10): 2199-2207)

[97] Ateniese G, Magri B, Venturi D, et al. Redactable blockchain-or-rewriting history in bitcoin and friends [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 111-126

[98] Dominic D, Bernardo M, Sri A. Redactable blockchain in the permissionless setting [C] //Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 124-138

[99] Derler D, Samelin K, Slamanig D, et al. Fine-grained and controlled rewriting in blockchains: Chameleon-hashing gone attribute-based [C] //Proc of the 26th Annual Network and Distributed System Security. San Diego, CA: The Internet Society, 2019: 24-27

[100] Zeleny M. Human Systems Management [M]. Manhattan: John Wiley, 2015: 128-302

[101] Riccardo G, Anna M, Salvatore R, et al. A survey of methods for explaining black box models [J]. ACM Computing Surveys, 2018, 51(5): 1-42

[102] Ji Shouling, Li Jinfeng, Du Tianyu, et al. Survey on techniques, applications and security of machine learning interpretability [J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096 (in chinese)
(纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096)

[103] Dosilovic F, Brcic M, Hlupic N. Explainable artificial intelligence: A survey [C] //Proc of the 41st Int Convention on Information and Communication Technology, Electronics and Microelectronics. Piscataway, NJ: IEEE, 2018: 210-215

[104] Zhang Quanshi, Song Chunzhu. Visual interpretability for deep learning: A survey [J]. arXiv preprint, arXiv: 1802.00614, 2014

[105] Silva V, André F, Handschuh S. On the semantic interpretability of artificial intelligence models [J]. arXiv preprint, arXiv: 1907.04105, 2019

[106] Vo H. Blockchain-based data management and analytics for micro-insurance applications [C] //Proc of the ACM Int Conf on Information and Knowledge Management. New York: ACM, 2017: 2539-2542

[107] Vo H. Research directions in blockchain data management and analytics [C] //Proc of Int Conf on Extending Database Technology. Berlin: Springer, 2018: 445-448

[108] Vo H. Blockchain-powered big data analytics platform [C] //Proc of the Int Conf on Big Data Analytics. Berlin: Springer, 2018: 15-32

[109] Shae Z, Tsai J P. On the design of a blockchain platform for clinical trial and precision medicine [C] //Proc of the Int Conf on Distributed Computing Systems. Piscataway, NJ: IEEE, 2017: 1972-1980

[110] Tsai J. Transform blockchain into distributed parallel computing architecture for precision medicine [C] //Proc of the Int Conf on Distributed Computing Systems. Piscataway, NJ, IEEE, 2018: 1290-1299

[111] Xu Xiwei, Lu Qinghua, Liu Yue, et al. Designing blockchain-based applications a case study for imported product traceability [J]. Future Generation Computer Systems, 2019, 92: 399-406

[112] Swan M. Blockchain: Blueprint for a New Economy [M]. Sebastopol, CA: O'Reilly Media Inc, 2015: 1-18

[113] Vasco L, Luis A. An overview of blockchain integration with robotics and artificial intelligence [J]. arXiv preprint, arXiv: 1810.00329, 2018

[114] Salah K, Rehman M H U, Nizamuddin N, et al. Blockchain for AI: Review and open research challenges [J]. IEEE Access, 2019, 7: 10127-10149

[115] Pan Chen, Liu Zhiqing, Liu Zhen, et al. Research on scalability of blockchain technology: Problems and methods [J]. Journal of Computer Research and Development, 2018, 55(10): 2099-2110 (in Chinese)
(潘晨, 刘志强, 刘振, 等. 区块链可扩展性研究: 问题与方法[J]. 计算机研究与发展, 2018, 55(10): 2099-2110)

[116] Han Xuan, Yu Yong, Wang Feiyue. Security problems on blockchain: The state of the art and future trends [J]. Acta Automatica Sinica, 2019, 45(1): 208-227 (in Chinese)
(韩璇, 袁勇, 王飞跃. 区块链安全问题: 研究现状与展望[J]. 自动化学报, 2019, 45(1): 208-227)

[117] Zhu Liehuang, Gao Feng, Shen Meng, et al. Survey on privacy preserving techniques for blockchain technology [J]. Journal of Computer Research and Development, 2017, 54(10): 2170-2185 (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): 2170-2185)

[118] Li Yang, Zheng Kai, Yan Ying, et al. EtherQL: A query layer for blockchain system [C] //Proc of the Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2017: 556-567

[119] Xu Cheng, Zhang Ce, Xu Jianliang. vChain: Enabling verifiable Boolean range queries over blockchain databases [J]. arXiv preprint, arXiv:1812.02386, 2018

[120] Zhang Ce, Xu Cheng, Xu Jianliang, et al. GEM^2-Tree: A gas-efficient structure for authenticated range queries in blockchain [C] //Proc of the 35th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2019: 842-853

[121] Ruan Pingcheng, Chen Gang, Dinh T, et al. Fine-grained, secure and efficient data provenance on blockchain systems [C] //Proc of the Very Large Database. New York: ACM, 2019: 975-988

[122] Wu Chao, Zhou Liyi, Xie Chulin, et al. Data quality transaction on different distributed ledger technologies [C] //Proc of the Int Conf on Big Scientific Data Management. Piscataway, NJ: IEEE, 2018: 301-318

[123] Liang Danwei, An Jian, Cheng Jindong, et al. The quality control in crowdsensing based on twice consensuses of blockchain [C] //Proc of the Int Symp on Pervasive and Ubiquitous Computing and Wearable Computers. New York: ACM, 2018: 630-635

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. Fellow of CCF. His main research interests include cloud data management, Web data management, flash-based databases and privacy protection, etc.

孟小峰,1964 年生.中国人民大学教授,博士生导师,CCF 会士.主要研究方向为云数据管理、Web 数据管理、闪存数据库和隐私保护等.

Liu Lixin, born in 1983. PhD candidate at Renmin University of China. Student member of CCF. Her main research interests include privacy protection and blockchain.

刘立新,1983 年生.中国人民大学博士生,CCF 学生会员.主要研究方向为隐私保护和区块链技术等.