基于区块链的大数据确权方案

王海龙 田有亮 尹 鑫
(贵州大学计算机科学与技术学院 贵阳 550025)
(贵州省公共大数据重点实验室(贵州大学) 贵阳 550025)
(贵州大学密码学与数据安全研究所 贵阳 550025)

摘 要 数据确权一直是大数据交易面临的挑战性之一。传统的确权手段采用提交权属证明和专家评审的模式,但是缺乏技术可信度,且存在潜在的篡改等不可控因素。为解决这些问题,迫切需要操作性强的确权方案。基于区块链技术和数字水印技术,提出了一种新的大数据确权方案。首先,引入审计中心和水印中心,以分离大数据完整性审计和水印生成的职责。其次,基于数据持有性证明技术和抽样技术,实现对大数据完整性的轻量级审计。再次,利用数字水印技术的特殊安全性质,实现对大数据起源的确认。最后,针对整个确权过程中涉及到的证据的完整性和持久性,利用区块链的原生特点实现确权结果与相关证据的强一致性。正确性和安全性分析结果表明,该方案能够为大数据的所有权界定提供新的技术思路和方法。

关键词 数据确权,区块链,数字水印技术,密码学
中图法分类号 TP309.7 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.003

Blockchain-based Big Data Right Confirmation Scheme

WANG Hai-long TIAN You-liang YIN Xin
(College of Computer Science & Technology, Guizhou University, Guiyang 550025, China)
(Guizhou Provincial Key Laboratory of Public Big Data(Guizhou University), Guiyang 550025, China)
(Institute of Cryptography & Data Security, Guizhou University, Guiyang 550025, China)

Abstract Data right confirmation has always been one of the most challenging problems in big data trading. Traditional means of right confirmation adopt the model of submitting the ownership evidence and expert reviewing, but they lack technological credibility and there are some uncontrollable factors such as potential tampering. In order to solve these problems, a strongly operational confirmation scheme is urgently needed. This paper put forward a new big data right confirmation scheme based on the technologies of blockchain and digital watermarking. Firstly, the auditing center and watermarking center are brought in to separate the duties of the integrity auditing of big data and watermarking generation. Secondly, using provable data possession technique and sampling technique, the lightweight audit of the integrity of big data is realized. Thirdly, the special security properties of digital watermarking technology are used to confirm the origin of big data. Finally, in the light of the integrity and persistence of evidence involved in the right confirmation, the native features of blockchain, such as the shared ledger, are used to implement strong consistency of the right confirmation result and relevant evidence. Correctness and security analysis results show that the proposed scheme can provide a new technical solution for the definition of ownership of big data.

Keywords Data right confirmation, Blockchain, Digital watermarking technology, Cryptography

1 引言

数据是继物质、能源之后的第三大基础性战略资源。2016年12月,国务院印发的《"十三五"国家信息化规划》中明确指出,优先开展数据资源共享开放行动。在推进数据资源开放共享的实践中,须解决的首要问题是[1]:数据作为一种重要资产,其流通和应用必然涉及数据的所有权问题。明晰数据的所有权,是大数据交易的前提和基础。数据的权属关系不清晰,一方面可能造成后续开发利用中产生权属纠纷问题;更严重的是,在数据归属模糊的情况下进行大数据分析关联,也难以界定权责归属,数据安全和个人隐私难以得到保障。这些问题严重制约着大数据的共享开放实践。因此,数

到稿日期:2017-10-29 返修日期:2018-01-08 本文受国家自然科学基金(61363068,61662009,61772008)资助。
王海龙(1993-),男,硕士生,CCF学生会员,主要研究方向为密码学与安全协议,E-mail:2277581700@qq.com;田有亮(1982-),男,博士,教授,CCF会员,主要研究方向为算法博弈论、密码学与安全协议,E-mail:youliangtian@163.com(通信作者);尹 鑫(1992-),女,硕士生,CCF学生会员,主要研究方向为密码学与安全协议,E-mail:csyxcryp@163.com。

据确权在大数据时代尤为关键，关系到大数据产业的创新活力及大数据交易市场的繁荣。

数据确权[1]一般是确定数据的权利人，即谁拥有对数据的所有权、占有权、使用权、受益权，以及对个人隐私权的保护责任等。本文研究数据确权时，主要聚焦于数据的所有权，即数据归属问题。具体地说，产生这批数据或者第一个收集这些数据的企业主体就是这批数据的所有者。通过其他任何方式(交易等)获得这批数据的企业或个人都只拥有使用权，而无所有权。目前，学术界对数据确权的研究成果相对较少。彭云[2]于 2016 年在大数据环境下研究了数据确权的核心问题。同年，涂燕辉[3]从法律的角度论述了数据确权的紧迫性和必要性。郭兵等人[4]于 2017 年以保护个人数据产权、知情权、隐私权和收益权为核心，提出了一种个人大数据资产管理与增值服务系统。王帅宇等人[5]于 2017 年公开了一种基于区块链技术的大数据确权方法及系统，但该方法未涉及对大数据源头的确认。因此，亟须从技术角度给出一种可靠且可操作性强的大数据确权方法。

目前，比较有代表性的确权方法是贵阳大数据交易所从管理角度提出的“提交权属证明＋专家评审”模式。在此确权模式下，首先，大数据的拥有者提交权属证明；其次，大数据交易所组织专家进行评审；最后，大数据交易所公布结果。专家在评审过程中有可能掺杂主观情感甚至偏见，破坏数据确权的公平性；且大数据交易所的内部人员出于利益关系可能存在恶意修改等行为。大数据交易所缺乏一种机制来永久保存评审材料和评审结果以备审计，当前交易所一般采用纸质文档和电子文档保存的方式，存在易丢失和易被篡改等问题。此外，鉴于大数据数据量大的特性，如何实现在不发送整批大数据的前提下高效、轻量地完成数据确权也是一大挑战。这些不可控因素表明，在技术上寻找一种解决方案迫在眉睫。

针对这些挑战，本文基于数字水印技术[6]和区块链技术[7]，提出了一种新的大数据确权方案，该方案具有确权的公平性、完整性和不可欺骗性。在初始化阶段，数据源供应商首先将大数据分块，并采用 BLS 短签名方案[8]对数据块取认证器，利用认证器的同态特性使数据源供应商不必发送原始数据；在确权请求、证据挑战和验证阶段，引入审计中心，数据源供应商和审计中心基于隐私保护数据持有性证明[9]和抽样技术[10]交互完成大数据的完整性审计，其中抽样技术确保了审计中心在挑战时抽取数据块的随机性；在水印生成和嵌入阶段，引入水印中心，由数据源供应商将能唯一标识自己身份信息的数据发送给水印中心，请求水印生成。水印中心将生成的水印发送给数据源供应商，由数据源供应商完成水印嵌入数据块的工作。在登记上链和查询阶段，基于区块链的内生优势(分布式、不可篡改、共享账本)实现确权结果和相关证据的链上高冗余保存，确保确权结果的完整性及不可篡改性。本文力图在技术上杜绝传统确权模式下大数据交易所篡改确权结果的完整性以及破坏确权的公平性，确保数据源供应商的利益，进一步为大数据交易市场的健康有序运作提供技术支撑。

本文第 2 节简要介绍双线性映射、BLS 短签名方案、同态认证器、区块链和数字水印技术；第 3 节提出一种大数据确权方案；第 4 节对大数据确权方案进行正确性分析、安全性分析和复杂度分析；最后总结全文。

2 准备知识

2.1 双线性映射

定义 1 设 G_1，G_2 和 G_t 是阶为 p 的乘法循环群，其中 p 为大素数。g_1，g_2 分别是 G_1，G_2 的生成元。若满足下列 3 条性质，则称映射 $e:G_1\times G_2\rightarrow G_t$ 为双线性映射[11]。

1)双线性性：设任意 $g_1\in G_1$，$g_2\in G_2$，$a,b\in Z_p$，有 $e({g_1}^a,{g_2}^b)=e(g_1,g_2)^{ab}$；

2)非退化性：对每一个 $g_1\in G_1\backslash\{1\}$，总存在 $g_2\in G_2$，使得 $e(g_1,g_2)\neq 1$；

3)有效可计算性：对任意的 $g_1\in G_1$，$g_2\in G_2$，存在有效的算法可以计算出 $e(g_1,g_2)$。

2.2 BLS 短签名方案

设 G 是一个阶为 p 的乘法循环群，其中 p 是一个大素数，g 是 G 中的一个生成元，群 G 上的 DDH(Decisional Diffie-Hellman)问题和 CDH(Computational Diffie-Hellman)问题定义如下。

DDH：设 $a,b,c\in Z_p^*$，$g,g^a,g^b,g^c\in G$，判定 $c\equiv ab\pmod p$ 是否成立；

CDH：设 $a,b\in Z_p^*$，$g,g^a,g^b\in G$，计算 g^{ab}。

在 G 中，如果 DDH 问题容易解决，但 CDH 问题在计算上不可行，则称 G 为 GDH(Gap Diffie-Hellman)群。

BLS 短签名方案是由 Boneh 等人[8]提出的一种短消息签名方案。BLS 短签名方案由 3 个算法组成：密钥生成算法 *KeyGen*，签名算法 *Sign*，签名验证算法 *Verify*。$H:\{0,1\}^*\rightarrow G\backslash\{1\}$ 是一个 hash 函数，其中 1 是 G 中的单位元。

KeyGen：选择随机数 $x\leftarrow_R Z_p^*$，接着计算 $v\equiv g^x\pmod p$，其中 x 为签名私钥，v 为签名公钥。

Sign：给定私钥 x 和消息 m($m\in\{0,1\}^*$)，计算 $h\leftarrow H(m)$ 和 $\sigma\leftarrow h^x$，签名为 $\sigma\in G\backslash\{1\}$。

Verify：给定公钥 v、消息 m 和签名 σ，如果 $e(g,\sigma)=e(v,h)$，那么验证者接受该签名，否则，验证者拒绝该签名。

2.3 同态认证器

同态认证器(Homomorphic Authenticator)，也称同态可验证标签(Homomorphic Verifiable Tags)，最早由 Ateniese 等[12]提出，是构造云上数据公开审计方案的基本工具。同态认证器除了需要满足不可伪造性，还需要满足以下性质[13]。

令 (pk,sk) 为签名者的公私钥对，σ_1，σ_2 分别为数据块 $m_1,m_2\in Z_p$ 的签名。

Blockless 可验证性(Blockless Verifiability)：给定 σ_1,σ_2，两个随机数 $\alpha_1,\alpha_2\in Z_p$ 和一个数据块 $m'=\alpha_1 m_1+\alpha_2 m_2\in Z_p$，验证者能够在不知道 m_1 和 m_2 的情况下验证 m' 的正确性。

不可延展性(Non-malleability)：给定 σ_1,σ_2，两个随机数 $\alpha_1,\alpha_2\in Z_p$ 和一个数据块 $m'=\alpha_1 m_1+\alpha_2 m_2\in Z_p$，没有私钥 sk 的用户不可以通过线性组合 σ_1 和 σ_2 生成一个关于数据块 m' 的有效签名 σ'。

2.4 区块链

简单来讲，区块链是利用密码学和分布式系统将系统内

的有效交易打包到一个只可附加型账本。区块链一般满足共享账本、不可篡改和分布式等基本特征。根据节点参与共识是否有准入机制，区块链一般被分为许可链（Permissioned Blockchain）和非许可链（Permissionless Blockchain）。本文采用的 Hyperledger 项目下的 Fabric[14] 区块链是许可链。此外，选择 Fabric 还基于如下判断：1）Hyperledger 在 2017 年 7 月份发布了 Fabric v1.0 版本，其可以实现概念验证，是最接近商用需求（私密性、可审计等）的区块链实现；2）本文中的区块链权属登记商业网络属于商业应用，本质上有相似的功能和非功能需求。

2.5　数字水印技术

数字水印是指将标识信息嵌入到数据载体内部，以达到版本保护、保密通信、文件真伪鉴别和产品标识等目的。嵌入的信息不影响数据载体的使用，并且不易被提取或修改。一旦发生所有权纠纷，可以将水印提取出来进行检测，从而证明版权的归属。一个完整的水印系统包括水印的生成、嵌入、检测和提取。

3　大数据确权方案

3.1　系统模型

本方案包括 4 个主体单元：数据源供应商 P，审计中心 T，水印中心 C，Fabric 区块链权属登记商业网络 B。图 1 是本方案中使用的模型。

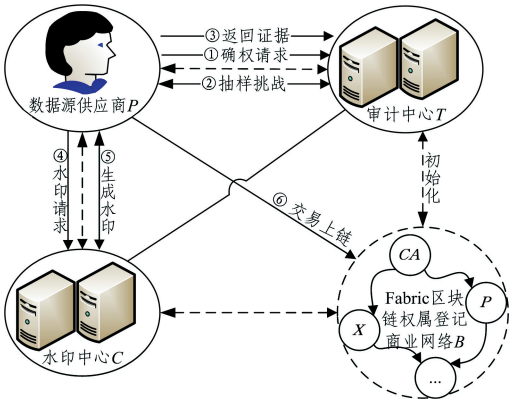

图 1　系统模型

Fig. 1　System model

本方案中各主体单元首先向证书机构 CA 申请公钥数字证书（同时完成认证接入）；其次，数据源供应商 P 对大数据进行分块，运用 BLS 短签名方案对数据块分别取认证器，并把数据块数、数据块认证器、大数据标识符 ID 的签名等与大数据确权相关的确权请求信息发送给审计中心 T；再次，审计中心 T 利用数据源供应商 P 的公钥验证签名的正确性，若验证通过，则审计中心 T 会向数据源供应商 P 发送证据挑战请求；最后，审计中心 T 收到证据后，利用双线性对的双线性质来验证等式是否成立，若成立，则数据源供应商 P 向水印中心 C 申请生成水印并完成水印嵌入。审计中心 T 和水印中心 C 将挑战证据、相关确权信息、水印以及相关元信息以 JSON 的格式封装成一笔交易并发送给数据源供应商 P，待数据源供应商 P 签名以后再将其发送到 Fabric 区块链权属登记商业网络 B。网络中的共识节点在鉴定交易中签名的合法性后，按照共识算法的要求最终将权属信息写入 Fabric 区块链。

3.2　方案的构造

3.2.1　参数

方案中用到的相关参数及其意义如下。

1）审计中心 T：负责大数据完整性审计事宜实施的专业机构。

2）数据源供应商 P：发起大数据确权请求的实体单元，一般为政府部门或互联网企业。

3）水印中心 C：负责为数据源供应商 P 生成水印，实现大数据起源的确认。

4）Fabric 区块链权属登记商业网络 B：负责将审计中心 T 和水印中心 C 联合构造的数据（交易）登记上链。该网络包括注册中心 CA、数据源供应商 P、节点 X 等负责共识的主体单元，其中 CA 负责对其他主体单元进行身份核实并签发公钥证书。

5）设 G_1，G_2 和 G_t 是素数阶为 p 的乘法循环群，且它们之间存在双线性映射 $e:G_1\times G_2\rightarrow G_t$，$g$ 是 G_2 的生成元，密码哈希函数 $H:\{0,1\}^*\rightarrow G_1$，密码哈希函数 $h:G_t\rightarrow Z_p$ 将 G_t 中的元素均匀地映射到 Z_p。

3.2.2　方案构造的具体过程

本文构造的方案具体包括 4 部分：初始化；确权请求、证据挑战和验证；水印生成和嵌入；登记上链和查询。

1）初始化

①数据源供应商 P、审计中心 T、水印中心 C 和其他负责共识的主体单元在注册中心 CA 注册；注册中心 CA 对各主体的身份信息进行审核。审核通过后，注册中心 CA 为各主体单元签发证书，该证书用于识别和认证网络中的主体。同时，初始化一个 Fabric 区块链权属登记商业网络 B。

②数据源供应商 P 将待确权大数据 D 分成 n 个数据块 $d_1,\cdots,d_n\in Z_p$，p 是一个大素数，即大数据 $D=\{d_i\}(i\in[1,n])$。数据块是数据确权时的基本单位。数据源供应商 P 选择一个随机的签名密钥对 (spk,ssk)，$x\leftarrow_R Z_p$，$u\leftarrow G_1$，并且计算公钥 $v\leftarrow g^x$。数据源供应商 P 将参数 $pk=(spk,v,g,u,n,e(u,v))$ 公开，对参数 $sk=(x,ssk)$ 保密。

③数据源供应商 P 为每个数据块 d_i 计算认证器 $\sigma_i\leftarrow(H(W_i)\cdot u^{d_i})^x\in G_1$，其中 $W_i=name\parallel i$，$name$ 是数据源供应商 P 均匀、随机地从 Z_p 中选择的作为待确权大数据 D 的标识符 ID。W_i 是大数据标识符 ID 和数据块索引 i 的连接，将 $\psi=\{\sigma_i\}_{1\leqslant i\leqslant n}$ 记为数据块认证器集合。

④为了保证大数据标识符 ID 的完整性，数据源供应商 P 计算 $tag=name\parallel Sign_{ssk}(name)$，并将其作为大数据 D 的标签，其中 $Sign_{ssk}(name)$ 是在私钥 ssk 下对 $name$ 的签名。

2）确权请求、证据挑战和验证

①数据源供应商 P 将验证数据 $(\{\sigma_i\}_{1\leqslant i\leqslant n},tag,v,n,spk)$ 发送给审计中心 T。审计中心 T 收到验证数据后，数据源供应商 P 再对大数据 D 的任何添加、删除都可以在证据挑战阶段被发现，从而有效防止数据源供应商 P 对大数据 D 进行增、删、改操作。

②审计中心 T 通过公钥 spk 验证签名 $Sign_{ssk}(name)$，若

验证成功,则恢复出大数据的 ID,即 $name$;若验证不通过,则终止确权。

审计中心 T 首先从大数据 D 的分块索引集合[1,n]中随机挑选 c 个块索引$\{s_1,\cdots,s_c\}$。具体抽取可用式(1)计算[10]:

$$s_k=((j^k(tag))\bmod n)+1 \tag{1}$$

$$j^k(tag)=\begin{cases}j^k(tag), & k=1\\ j(j^{k-1}(tag)), & k=2,\cdots,c\end{cases} \tag{2}$$

其中,j 为单向哈希函数。

接着,审计中心 T 对每个块索引 $i\in\{s_1,\cdots,s_c\}$ 选取一个相应的随机数 $v_i\leftarrow_R Z_{p/2}$,最后将它们组成挑战请求 $chal=\{(i,v_i)\}_{s_1\leqslant i\leqslant s_c}$,并将 $chal$ 发送给数据源供应商 P。

数据源供应商 P 接收到请求 $chal$ 后,首先计算 $\mu=r+\gamma\sum_{i=s_1}^{s_c}v_id_i \bmod p$,其中 $r\leftarrow_R Z_p$,$R=e(u,v)^r\in G_t$,$\gamma=h(R)\in Z_p$。同时,数据源供应商 P 计算一个聚合认证器 $\sigma=\prod_{i=s_1}^{s_c}\sigma_i^{v_i}$。然后,数据源供应商 P 将$\{\sigma,R,\mu\}$作为证据返还给审计中心 T。

审计中心 T 接收到证据$\{\sigma,R,\mu\}$后,计算 $\gamma=h(R)$,然后通过式(3)来判断接收到的证据的正确性:

$$R\cdot e(\sigma^\gamma,g)=e((\prod_{i=s_1}^{s_c}H(W_i)^{v_i})^\gamma\cdot u^\mu,v) \tag{3}$$

3)水印的生成和嵌入

审计中心 T 验证完式(3)后,不管成功与否,都将把结果 $o\in$ {success,failure}返回给数据源供应商 P。只有在结果 o 为 success 的情况下,审计中心 T 才将结果 o 以及自己对申请大数据确权主体信息的签名 $Sign_{skT}(P)$发送给水印中心 C。数据源供应商 P 接收到 success 后,可以向水印中心 C 发出申请水印生成请求,数据源供应商 P 将能唯一标识自己的身份信息 $info$(例如企业标识或签名)的签名 $Sign_{sk}(info)$发送给水印中心 C,水印中心 C 用数据源供应商 P 的公钥 spk 对签名信息验证成功后,利用水印生成算法生成水印 $w_i=F(info,k)$,$i=1,\cdots,q_n(q\in(0,1))$,其中 F 为不可逆的水印生成算法。接着,水印中心 C 将这些水印 w_i 发送给数据源供应商 P,数据源供应商 P 将水印 w_i 嵌入数据块 d_i 中,即 $d_i'=A(d_i,F(info,k))$,其中 A 表示水印嵌入算法(编码算法),d_i,$info$,k 分别表示原始数据块、原始水印信息以及密钥。由于大数据的特殊性,数据源供应商 P 在嵌入水印的过程中随机抽取规模为 qn 的数据块嵌入水印信息即可,具体的抽取比例 q 可以根据数据块的规模 n 而定。相反,在式(3)验证不成功的情况下,如果数据源供应商 P 能够提供进一步的有力确权证据,那么数据源供应商 P 和审计中心 T 将反复进行确权请求和挑战两个阶段,直到验证通过。

4)登记上链和查询

为了防止数据源供应商 P 私下添加不合法或者过期的水印到数据块中,水印中心 C 会将生成的水印 w_i 签名后发送到区块链权属登记商业网络 B,审计中心 T 会将整个大数据完整性审计过程中涉及的确权证据$\{\sigma,R,\mu\}$、验证数据$(\{\sigma_i\}_{1\leqslant i\leqslant n},tag)$、挑战请求 $chal=\{(i,v_i)\}_{s_1\leqslant i\leqslant s_c}$、参数 pk 等信息利用 BLS 短消息方案签名后发送到区块链权属登记商业网络 B。审计中心 T 和水印中心 C 会联合构造好一笔包含上述水印和完整性审计数据的交易 TripleTx,并将构造好的交易 TripleTx 发送给数据源供应商 P;数据源供应商 P 在查看交易内容后,用自己的私钥签名。换言之,当 3 个主体均完成签名后,这笔交易上链时才是有效的。最后,审计中心 T 或水印中心 C 拿到带有数据源供应商 P 签名的交易 TripleTx 后,将 TripleTx 广播出去。区块链权属登记商业网络 B 中的共识节点对交易 TripleTx 中签名的有效性进行验证,按照相应的共识算法完成共识后将其写入到链上。数据源供应商 P 可以通过 web/app 等方式查询登记在区块链权属登记商业网络 B 上的确权结果。图 2 为 TripleTx 的交易结构,交易中带有尖括号的签名有待数据源供应商 P 填入。

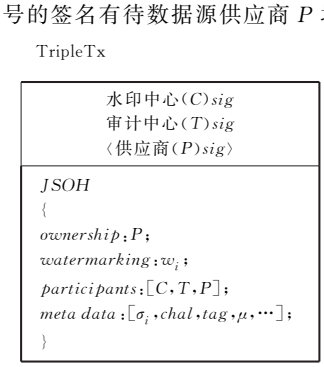

图 2 TripleTx 的交易结构

Fig. 2 Transaction structure of TripleTx

至此,整个大数据的确权方案完成。

4 方案分析

4.1 正确性分析

引理 1 首先证明等式(3)的正确性。

证明:因为 $R=e(u,v)^r$,$\sigma=\prod_{i=s_1}^{s_c}\sigma_i^{v_i}$,$v\leftarrow g^x$,$\sigma_i\leftarrow(H(W_i)\cdot u^{d_i})^x$,所以有式(4):

$$\begin{aligned}R\cdot e(\sigma^\gamma,g)&=e(u,v)^r\cdot e((\prod_{i=s_1}^{s_c}(H(W_i)\cdot u^{d_i})^{x\cdot v_i})^\gamma,g)\\&=e(u^r,v)\cdot e((\prod_{i=s_1}^{s_c}(H(W_i)^{v_i}\cdot u^{d_iv_i})^\gamma,g)^x\\&=e((\prod_{i=s_1}^{s_c}H(W_i)^{v_i})^\gamma\cdot u^{r+\gamma\sum_{i=s_1}^{s_c}d_i\cdot v_i},v)\end{aligned} \tag{4}$$

又由于 $\mu=r+\gamma\sum_{i=s_1}^{s_c}v_id_i$,因此式(4)可以进一步化简为 $e((\prod_{i=s_1}^{s_c}H(W_i)^{v_i})^\gamma\cdot u^\mu,v)$,则 $R\cdot e(\sigma^\gamma,g)=e((\prod_{i=s_1}^{s_c}H(W_i)^{v_i})^\gamma\cdot u^\mu,v)$。证毕。

4.2 安全性分析

本节将从公平性、完整性、不可欺骗性 3 方面来分析大数据确权方案。

定理 1 该大数据确权方案满足确权公平性。

证明:首先,在确权方案的初始化阶段,数据源供应商 P、审计中心 T 以及区块链权属登记商业网络 B 均向 CA 申请公钥证书,参与主体的身份得到了确认,交互行为处在相对可靠的环境中,同时基于 CA 可以实现行为不可抵赖性。其次,方案中采用“水印中心 C+审计中心 T”替代传统方法中大数据交易所负责确权实施的模式,消除了大数据交易所内部人

员和专家评审过程中带来的篡改和主观威胁等，同时在证据挑战阶段，式(3)的可公开验证性可以进一步增强审计中心 T 在确权过程中的公平性。本方案将确权中的完整性审计和水印分发进行分离，审计中心 T 负责大数据完整性的审计工作，水印中心 C 负责大数据水印的生成，数据源供应商 P 负责将水印嵌入到数据块中。数据上链也需要水印中心 C、审计中心 T 和数据源供应商 P 三方签名后才能有效，缺少任何一方，均完成不了该过程。

定理 2　该大数据确权方案满足确权数据的完整性。

证明：确权数据的完整性一方面是指已经完成确权之后的大数据的完整性。在确权方案的初始化阶段，数据源供应商 P 首先对 D 进行分块，将分块数据记为 d_i，其中 $1\leqslant i\leqslant n$，并且采用 BLS 短签名方案对每个数据块 d_i 取认证器 σ_i。然后，供应商 P 将认证器集合和大数据 ID 的标签 $\{\psi=\{\sigma_i\}_{1\leqslant i\leqslant n},tag\}$ 一起发送给审计中心 T，审计中心 T 收到后，若供应商 P 试图对大数据 D 本身进行修改或者对分块数据 d_i 进行修改，则只有供应商 P 能够重新找到一个 $d_i'\neq d_i$ 使得 $(H(W_i)\cdot u^{d_i})^x=(H(W_i)\cdot u^{d_i'})^x\in G_1$ 才成功，这显然是不可能的，从而保证了大数据的完整性。另一方面，是指确权结果和相关证据的完整性。在登记上链阶段，区块链中区块和区块之间通过利用密码哈希函数顺序勾连，攻击者篡改任意一个区块中的数据均会对后续的区块产生作用。因此，区块链权属登记商业网络 B 中的任意一个节点(包括共识节点)对区块中的任意一笔交易信息(即权属信息)的本地化篡改均不能奏效，除非超过 1/3 的共识节点集体“作弊”。但这种概率几乎是可以忽略的，尤其是整个确权生态趋于健全时。因此，链上数据(确权结果和相关证据)的完整性也得到了有效保证。综上，所提确权方案满足完整性。

定理 3　该大数据确权方案具有不可欺骗性。

证明：初始化阶段，数据源供应商 P 首先对 D 进行分块，将分块数据记为 d_i，其中 $1\leqslant i\leqslant n$，并且采用 BLS 短签名方案对每个数据块 d_i 取认证器 σ_i。在确权请求、证据挑战和验证阶段，数据源供应商 P 将认证器集合 ψ 发给审计中心 T。数据源供应商 P 用自己的私钥 x 对数据块 d_i 签名，具有不可欺骗性。在证据挑战时，利用 BLS 签名机制的聚合性，P 将挑战请求 $chal=\{(i,v_i)\}_{s_1\leqslant i\leqslant s_c}$ 聚合成 σ 发送给审计中心 T。审计中心 T 会验证数据源供应商 P 的签名，只有验证成功，数据源供应商 P 才会向水印中心 C 请求水印生成且在后续参与交易上链。水印是由水印中心 C 生成的，水印中心 C 会将水印记录在链上。后期的数据使用者在交易这批数据时可以从大数据中提取出水印，进而与链上的水印进行比对，水印一致时提取的水印才是合法的。在登记上链和查询阶段，审计中心 T 和水印中心 C 将确权结果和相关证据以一笔交易的形式发给区块链权属登记商业网络 B，其中的共识节点会验证审计中心 T 和水印中心 C 发起的交易的真实性，只有通过验证的交易才会写入链上。因此，整个确权过程中的参与确权请求、证据挑战和水印生成的主体行为均具备不可欺骗性。

4.3　复杂度分析

下面给出本方案在大数据完整性审计、水印生成和确权相关证据组成的交易上链三阶段的复杂度分析。本方案的通信复杂度由确权过程中(审计、水印生成和交易上链)的通信轮数表示，如表 1 所列。

表 1　复杂度分析

Table 1　Complexity analysis

	计算复杂度	通信复杂度
完整性审计	$O(n)+O(c)$	$O(1)$
水印生成	$O(qn)$	$O(1)$
交易上链	$O(1)$	$O(1)$

审计阶段的计算复杂度由两部分组成，分别是数据源供应商的 $O(n)$ 和审计中心的 $O(c)$，通信复杂度为 $O(1)$，其中 n 为大数据分块数目，c 为随机抽取的数据块数。水印生成阶段的计算复杂度主要是水印中心产生水印的 $O(qn)$，通信复杂度为 $O(1)$，其中 q 为给数据块嵌入水印的抽取比例。交易上链阶段的计算复杂度主要由审计中心 T、水印中心 C 和数据源供应商 P 计算签名组成，共计 $O(1)$，通信复杂度为 $O(1)$。

结束语　本文引入审计中心和水印中心，以分离大数据的完整性审计和证明数据所有权的水印生成和嵌入，进而替代传统确权中大数据交易所直接实施确权的组织和评审以及由此引发的确权不公平和不可信局面；其次，鉴于大数据交易所负责确权结果的保存给确权结果的完整性带来了篡改威胁和不确定性风险，本文引入区块链，利用其分布式、高度冗余等特点将确权结果的保存从传统的大数据交易所一家转入整个大数据交易生态圈，实现了确权结果的始终一致性。在本文工作的基础上，下一步将基于区块链研究大数据交易权的流转，力图实现大数据交易生命周期中的可追责性。

参 考 文 献

[1] DU Z H. Research on Data Confirmation Right in Big Data Application[J]. Mobile Communications, 2015(13): 12-16. (in Chinese)
杜振华. 大数据应用中数据确权问题探究[J]. 移动通信, 2015(13): 12-16.

[2] PENG Y. Research on Authenticating Data Rights in Big Data environment[J]. Modern Science &Technology of Telecommunications, 2016, 46(5): 17-20. (in Chinese)
彭云. 大数据环境下数据确权问题研究[J]. 现代电信科技, 2016, 46(5): 17-20.

[3] TU Y H. Study on the Legal Rights of Big Data [J]. Journal of Foshan University(Social Science Edition), 2016, 34(5): 83-87. (in Chinese)
涂燕辉. 大数据的法律确权研究[J]. 佛山科学技术学院学报(社会科学版), 2016, 34(5): 83-87.

[4] GUO B, LI Q, DUAN X L, et al. Personal Data Bank: A New Model of Personal Big Data Asset Management and Value-Added Services Based on Bank Architecture[J]. Chinese Journal of Computers, 2017, 40(1): 126-143. (in Chinese)
郭兵，李强，段旭良，等. 个人数据银行——一种基于银行架构的个人大数据资产管理与增值服务的新模式[J]. 计算机学报, 2017, 40(1): 126-143.

(下转第 24 页)

的优势体现在以下 3 个方面:1)提高了系统的容错性能,相比于 PBFT,本文算法可以容忍小于一半的节点为拜占庭节点,可以达到 Hyperledger 中 XFT 共识算法相同的容错能力;2)具有较好的可扩展性,相比于传统的拜占庭共识算法,本文算法能够保证在系统规模增大时,节点间仍能高效率地传递信息;3)系统中的提案节点随着区块链长度的变化而转移,使系统具有均衡动态负载的性能。系统中所有节点都处于对等的地位,从而解决了单点故障问题。本文工作对区块链技术的共识算法在容错性和可扩展性两个方面的研究提供了一定的参考。

参 考 文 献

[1] YUAN Y, WANG F Y. 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.

[2] NAKAMOTO S. Bitcoin: A peer-to-peer electronic cash system [OL]. https://www.cs.bgu.ac.il/~crp161/wiki.files/Bitcoin-Paper.pdf.

[3] LAMPORT L, SHOSTAK R, PEASE M. The Byzantine Generals Problem [J]. Acm Transactions on Programming Languages & Systems, 2016, 4(3): 382-401.

[4] 邹均. 区块链技术指南[M]. 北京:机械工业出版社, 2016: 109-128.

[5] MICALI S. ALGORAND: The Efficient and Democratic Ledger [OL]. http://pdfs.semanticscholar.org/0dc0/55052cda7179cd74d43e07479565121ef733.pdf.

[6] Larimer D. Transactions as proof-of-stake [OL]. [2017-07-05]. https://bravenewcoin.com/assets/Uploads/TransactionsAsProofOfStake10.pdf.

[7] CASTRO M. Practical byzantine fault tolerance and proactive recovery [J]. Acm Transactions on Computer Systems, 1999, 20(4): 398-461.

[8] ONGARO D, OUSTERHOUT J. In search of an understandable consensus algorithm [C]// Usenix Conference on Usenix Technical Conference. USENIX Association, 2014: 305-320.

[9] LEI C J, LIN Y P, LI J G, et al. Research on Byzantine Fault Tolerance Under Volunteer Cloud Environment [J]. Computer Engineering, 2016, 42(5): 1-7. (in Chinese)
雷长剑,林亚平,李晋国,等. 志愿云环境下的拜占庭容错研究[J]. 计算机工程, 2016, 42(5): 1-7.

[10] LI J. Distributed Gossip algorithms for Quantized Consensus [D]. Harbin: Harbin Institute of Technology, 2013. (in Chinese)
李婧. 基于量化共识的分布式 Gossip 算法研究[D]. 哈尔滨:哈尔滨工业大学, 2013.

[11] ANDRÉ A, DEMERS A, HOPCROFTT J E. Correctness of a gossip based membership protocol [C]// Twenty-Fourth ACM Symposium on Principles of Distributed Computing. ACM, 2005: 292-301.

[12] GUREVICH M, KEIDAR I. Correctness of gossip-based membership under message loss [C]// ACM Symposium on Principles of Distributed Computing. ACM, 2009: 151-160.

[13] GANSESH A J, KERMARREC A M, MASSOULI, et al. Peer-to-Peer Membership Management for Gossip-Based Protocols [J]. IEEE Transactions on Computers, 2003, 52(2): 139-149.

[14] 黄步添,王云霄,王从礼,等. 一种应用于区块链的拜占庭容错共识方法:中国,CN106445711A [P]. 2017-02-22.

[15] 张铮文. 一种用于区块链的拜占庭容错算法[OL]. [2017-07-03]. http://www.onchain.com/paper/66c6773b.pdf.

[16] KERMARREC A M, MASSOULIE L, GANESH A J. Probabilistic reliable dissemination in large-scale systems [J]. IEEE Transactions on Parallel & Distributed Systems, 2003, 14(3): 248-258.

(上接第 19 页)

[5] 王帅宇,李晨. 一种基于区块链技术的大数据确权方法及系统:中国,CN106815728A[P]. 2017-06-09.

[6] PETITCOLAS F A P, ANDERSON R J, KUHN M G. Information Hiding-a Survey[J]. Proceedings of the IEEE, 1999, 87(7): 1062-1078.

[7] NAKAMOTO S. Bitcoin: A Peer-to-Peer Electronic Cash system [OL]. https://bitcoin.org/bitcoin.pdf.

[8] BONEH D, LYNN B, SHACHAM H. Short Signatures from the Weil pairing[C]// International Conference on the Theory and Application of Cryptology and Information Security. Springer Berlin Heidelberg, 2001: 514-532.

[9] WANG C, CHOW S S M, WANG Q, et al. Privacy-Preserving Public Auditing for Secure Cloud Storage[J]. IEEE Transactions on Computers, 2013, 62(2): 362-375.

[10] DU W, JIA J, MANGAL M, et al. Uncheatable Grid Computing [C]// International Conference on Distributed Computing Systems, 2004. IEEE, 2004: 4-11.

[11] ZHANG F G. From Bilinear Pairings to Multilinear Maps[J]. Journal of Cryptologic Research, 2016, 3(3): 211-228. (in Chinese)
张方国. 从双线性对到多线性映射[J]. 密码学报, 2016, 3(3): 211-228.

[12] ATENIESE G, BURNS R, CURTMOLA R, et al. Provable Data Possession at Untrusted Stores[C]// ACM Conference on Computer and Communications Security. ACM, 2007: 598-609.

[13] WANG B, LI B, LI H. Oruta: Privacy-Preserving Public Auditing for Shared Data in the Cloud[C]// IEEE Fifth International Conference on Cloud Computing. IEEE Computer Society, 2012: 295-302.

[14] fabric[OL]. (2017-10-20). https://github.com/hyperledger/fabric.