基于简单随机抽样的大数据可信性验证方法

任正伟[1],孙小雁[2],王丽娜[3],王 骞[3],徐明迪[1],张茂胜[2,3]
(1. 武汉数字工程研究所,湖北武汉 430205;
2. 玉林师范学院复杂系统优化与大数据处理广西高校重点实验室,广西玉林 537000;
3. 武汉大学计算机学院,湖北武汉 430072)

摘 要: 针对大数据的可信性验证问题,本文提出了一种大数据可信性验证方法以验证数据来源和数据内容的可信性. 本文首先通过验证数据属主的身份证书来实现数据来源可信性的验证,再在简单随机抽样和可聚合的广播签名方案的基础上,设计了一个交互式质询-应答协议,使得用户只需抽样少量数据就能以高置信率验证数据内容的可信性. 理论分析和实验结果表明,本文方法是安全的,且性能开销在合理范围内,能够实现大数据的可信性验证.

关键词: 大数据安全; 可信性验证; 认证标签; 安全协议
中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112 (2017)10-2484-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.10.024

A Trustworthiness Verification Approach for Big Data Based on Simple Random Sampling

REN Zheng-wei[1], SUN Xiao-yan[2], WANG Li-na[3], WANG Qian[3], XU Ming-di[1], ZHANG Mao-sheng[2,3]
(1. *Wuhan Digital and Engineering Institute, Wuhan, Hubei* 430205, *China*;
2. *Guangxi Colleges and Universities Key Laboratory of Complex System Optimization and Big Data Processing, Yulin Normal University, Yulin, Guangxi* 537000, *China*;
3. *School of Computer Science, Wuhan University, Wuhan, Hubei* 430072, *China*)

Abstract: In this paper, we propose an approach to verify the trustworthiness of big data including the trustworthiness of data source and data content. The trustworthiness of data source is achieved by verifying the identity of data owner based on public key certificate. We design an interactive challenge-response protocol to verify the trustworthiness of data content on the basis of simple random sampling test and aggregated signature-based broadcast encrypting scheme. The protocol can make the users who know the public parameters of the data owner validate the verification result with high probability by sampling only a small amount of data. Formal security analysis and experimental evaluations are also conducted, showing that the proposed scheme is practical to achieve the design goal for big data.

Key words: big data security; trustworthiness verification; verification tag; security protocol

1 引言

大数据的概念被提出以来,得到了学术界和工业界的广泛关注. 大数据包含了大量的原始信息,大数据分析能够帮助人们透过现象更准确地把握事物背后的规律;基于分析出的知识,人们可以更准确地对自然和社会现象进行预测[1].

然而,大数据在其发展过程中,也面临着一些安全与隐私保护问题[1~3],数据的可信性验证即是其中之一,即在真实可信的数据之上进行分析,可以得到真实可信的结果;但在不可信的数据之上进行分析,则会被数据欺骗,得出错误的结论. 特别是在一些场景明确的

收稿日期:2016-03-07;修回日期: 2017-05-05;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61373169, No. 61502438);国防预研项目(No. JCKY2016207A055, No. 6140242010103);复杂系统优化与大数据处理广西高校重点实验室基金(No. 2016CSOBDP0005);广西自然科学基金(No. 2014GXNSFBA118010, No. 2014GXNSFBA118268)

应用中,欺骗者可能会伪造或刻意制造数据,以诱导分析者得出对其有利的结论. 虚假的信息往往隐藏于大量信息中,使得分析者无法鉴别其真伪,从而做出错误的判断. 因此,在大数据分析中,应当能够对数据的可信性,包括来源可信性和内容可信性进行验证[1],使得数据使用者能够对数据的可信性做出评估,防止在不可信的数据之上进行分析而得出无意义或者错误的结果.

数据来源可信性要求数据使用者能够确信大数据是由可信源发布的,验证数据属主(Owner)的身份可实现该目标. 具体而言,对代表了 Owner 身份的公钥证书的有效性进行验证,即可验证 Owner 的身份. 数据内容可信性要求数据使用者能够确信其使用的数据不是伪造的或未被篡改,验证数据内容的完整性可实现该目标. 安全哈希函数和散列消息认证码都可用来验证数据内容的完整性,但基于安全哈希函数的方案易受到替换攻击,而基于散列消息认证码的方案需将 Owner 的私钥共享给用户,存在着 Owner 私钥泄漏的安全隐患;当用户较多时,Owner 还需实时在线处理用户的数据访问请求,而且对数据的验证粒度较粗,难以满足大数据场景下的数据服务需求.

大数据一般通过云计算平台存储和管理[3],因而大数据的内容可信性验证也是一种外包数据的完整性验证. 在云存储模式下,Ateniese 等人[4,5]首次提出了数据持有性证明(Provable Data Possession,PDP)的概念及其形式化定义,并给出了具体方案,用来验证存储在不可信的云服务器上的文件的完整性. 该方案对数据分块,然后基于 RSA 签名算法生成数据的认证标签,再在简单随机抽样的基础上,以概率的方式来验证远程数据的完整性. 概率分析表明该方案能以高置信率保证验证结果的有效性. Jules 等人[6]提出了云存储中数据可恢复性模型 PoR(Proofs of Retrievability),PoR 模型也可验证远程数据的完整性,但与 PDP 模型相比,更侧重于保障数据的可用性,通过冗余编码对受损数据进行修复. 在原始 PDP 和 PoR 模型的基础上,研究者们深入研究了远程数据完整性验证的通信、计算、存储等开销以及安全性等[7~9],并根据实际应用需求考虑了一些新的特性,如由第三方进行的公开验证[10~12],在公开验证过程中保护数据的隐私[12,13],支持数据的动态更新[14~16]等.

PDP 和 PoR 方案的应用场景与外包大数据的可信性验证既有相同之处,如数据都是外包存储于不(半)可信的云计算平台上,但也存在着不同,不能直接应用于外包大数据的可信性验证. 在 PDP 和 PoR 方案中,不需要考虑数据来源(也即数据属主 Owner)的可信性问题,第三方验证者受 Owner 的委托验证其数据的完整性并将结果报告给 Owner,不会实际下载使用数据. 但在大数据场景下,数据使用者首先要确信其使用的数据是由可信源发布的,其次要确信其使用的数据内容是可信的,然后才会实际使用数据. 因而,在大数据场景中,可能还会存在 time-of-check-time-of-use(TOCTOU)攻击,即云服务提供商在数据验证时使用的是真实数据,但验证之后却将不可信的数据给使用者使用;攻击者在数据的下载过程中也可能篡改数据. 因此,大数据的可信性验证还需要考虑数据的本地验证.

针对大数据场景下的数据可信性验证问题,本文借鉴 PDP 模型的思想,在简单随机抽样和改进的可聚合的广播签名方案 ASBB(Aggregated Signature-Based Broadcast)[17]的基础上,提出了一种大数据可信性验证方法. 该方法可在流程上对数据的来源和内容的可信性进行顺序验证,即用户首先验证 Owner 的身份证书,以判断数据是否由可信源发布. 然后,用户根据数字证书中的信息对数据内容的完整性进行验证. 具体而言,Owner 首先从签证机构(Certification Authority,CA)处获得能够代表其身份的数字证书. 鉴于在发布数据时,无法事先预知哪部分数据会被哪个用户使用,并且为了支持数据的细粒度操作,Owner 会对数据进行分块处理,再采用改进的 ASBB 算法为每个数据块生成认证标签,使得任一知道 Owner 的公开参数(可通过 Owner 的数字证书获取这些公开参数)的用户都能通过一个交互式质询-应答协议在只抽样少量数据块的情况下,以高置信率验证大数据内容的可信性. 为了防止 TOCTOU 攻击,本文方案还支持用户对下载到本地的数据的可信性再次进行验证.

2　方案设计

本节将介绍本文所设计方法的整体框架及各部分的功能模块与算法,然后给出方案的安全模型和所基于的困难性假设.

2.1　整体框架

本文提出的大数据可信性验证方法中,共有 4 个参与方:云服务提供商(Cloud Services Provider,CSP),数据属主 Owner,数据使用者 User,签证机构 CA,其整体框架如图 1 所示.

1)CSP. CSP 拥有存储和计算资源,提供数据存储、发布、检索等服务,Owner 将其数据外包存储于 CSP 处以供 User 访问和使用.

2)Owner. Owner 对数据进行分块,并为数据块生成认证标签,将数据、认证标签以及代表其身份的数字证书一起存储于 CSP 处.

3)User. User 通过 CSP 访问和使用 Owner 发布的数据. 在使用数据前,User 会对数据的来源和内容的可信性进行验证.

4) CA. CA 为可信实体,其主要功能是对 Owner 的身份进行认证,并为 Owner 颁发代表其身份的数字证书,证书中还包含了对数据内容完整性进行验证时所需的公开参数等信息.

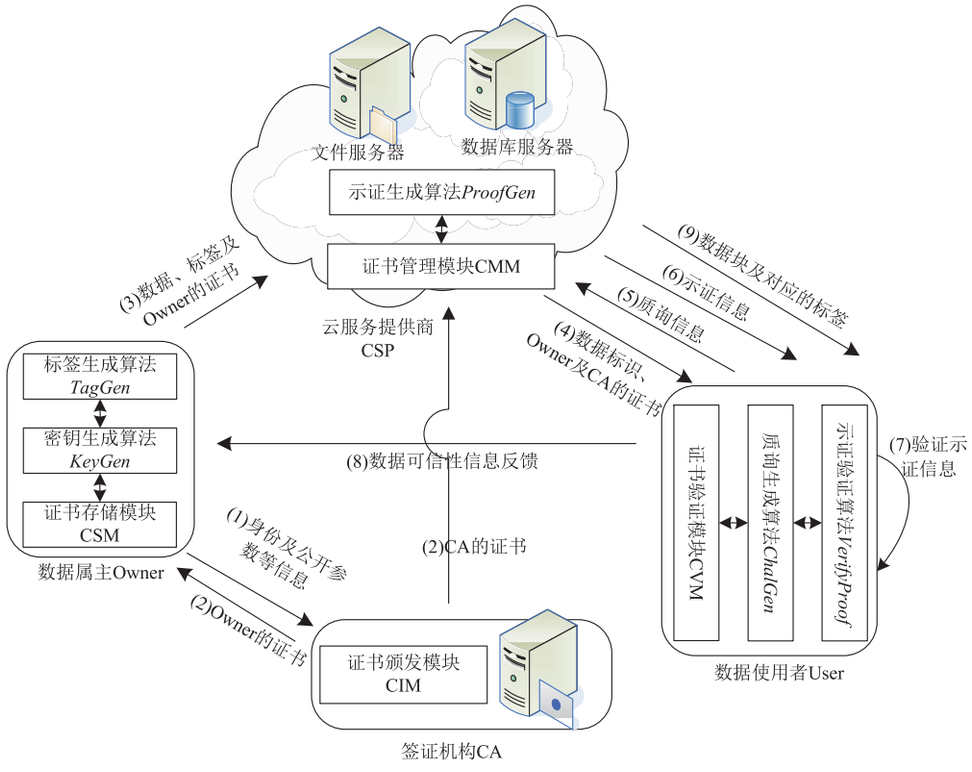

图1 基于简单随机抽样的大数据可信性验证框架

图 1 所示的框架中,证书存储、证书颁发、证书管理、证书验证等模块用于验证 Owner 的身份也即大数据来源的可信性;密钥生成 *KeyGen*、标签生成 *TagGen*、质询生成 *ChalGen*、示证生成 *ProofGen*、示证验证 *VerifyProof* 等算法用于验证大数据内容的可信性.

2.2 安全模型与困难性假设

1) 安全模型. 给定一对交互式实体(P,V),P 为(无界的)概率算法,V 为确定性的多项式时间算法. 定义$P(x)$表示示证者 P 拥有秘密 x,$\langle P,V\rangle(x)$表示示证者 P 和验证者 V 在协议执行过程中共享秘密 x,$\langle P(F,\sigma),V\rangle(pk)$为一个公开示证,其中,示证者 P 以文件 F、认证标签集合 σ、公钥 pk 为输入. 则本文方案的安全模型可定义[18]为:

正确性:对于任意的 $\sigma\in TagGen(sk,F)$,有

$$\Pr[\langle P(F,\sigma),V\rangle(pk)=1]\geqslant 1-1/p_1(\kappa) \quad (1)$$

完备性:对于任意的 $\sigma^*\notin TagGen(sk,F)$ 和任意一交互实体 P^*,有

$$\Pr[\langle P^*(F,\sigma^*),V\rangle(pk)=1]\leqslant 1/p_2(\kappa) \quad (2)$$

其中,$p_1(\cdot)$和 $p_2(\cdot)$为多项式,$\kappa\in\mathbb{N}$为安全参数,$1/p_1(\kappa)$称为正确率,$1/p_2(\kappa)$称为完备率. 完备性表明用错误的签名来欺骗验证者的成功概率是可忽略的,可被认为是数据签名不可伪造的严格定义. 当完备性成立时,抽样数据块一定是可信的.

2) 困难性假设. 本文方案安全性模型的成立依赖于离散对数(Discrete Logarithm, DL)问题和 Computational Diffie-Hellman(CDH)问题. 假定$(G,\cdot)$为一个循环群,其阶为素数 q,g 为 G 的一个生成元,即 $G=\langle g\rangle$,则 DL 问题和 CDH 问题可分别描述为:

1) DL 问题. 给定 $g,g^t\in G$,计算 $t\in\mathbb{Z}_q^*$.

2) CDH 问题. 给定 $g,g^\alpha,h\in G,\alpha$ 未知,计算 $h^\alpha\in G$.

3 方案描述

本文对大数据来源的可信验证是直接采用普通的数字证书实现的,因而本节将省略其详细的实现过程,而着重描述基于交互式质询-应答协议的大数据内容的可信性验证过程,以及数据的访问流程.

3.1 算法及协议描述

假定 G 和 G_T 是两个具有相同素数阶 p 的乘法循环群,e 为一个可计算的双线性映射 $e:G\times G\rightarrow G_T$,即对于 G 的生成元 g 有 $e(g,g)\neq 1$,并且对于所有的 $u,v\in\mathbb{Z}_p$,有 $e(g^u,g^v)=e(g,g)^{uv}$[19]. 令 $h(\cdot):\{0,1\}^*\rightarrow G$ 为安

全哈希函数,可将字符串与 G 中的元素一一映射. 本文方案的交互式质询-应答协议中各算法的详细描述如下:

1)密钥生成算法 *KeyGen*:Owner 随机选择两个元素 $r\in\mathbb{Z}_p^*$,$X\in G\backslash\{1\}$,计算 $R=g^{-r}$,$A_1=e(X,g)$,$A=e(X,g^r)$. 公钥为 $pk=(R,A)$,私钥为 $sk=(r,X)$,公开参数为(g,h,p,G,G_T,e,R,A_1,A).

2)标签生成算法 *TagGen*:给定一个数据文件 F,Owner 为其生成一个标识符 $fid_F\in\mathbb{Z}_p$,并将其等分为 n 个数据块,也即 $F=(m_1,m_2,\cdots,m_n)$,并且 $m_i\in\mathbb{Z}_p^*$ $(1\leqslant i\leqslant n)$. 对于每个数据块 m_i,Owner 为其计算认证标签 $\sigma_i=(X^{m_i}H_i)^r$,其中,$H_i=h(fid_F\|i)$,i 为 m_i 的索引. Owner 将所有数据块的认证标签集合表示为 $\phi=\{\sigma_i\}_{1\leqslant i\leqslant n}$.

3)质询生成算法 *ChalGen*:在每轮验证中,User 首先随机选取元素 $t\in\mathbb{Z}_p^*$ 和 $m\in G_T$,并计算 $c_1=g^t$,$c_3=A^t$,$c_2=R^t$. 然后 User 从集合$\{1,\cdots,n\}$中随机选取 c 个元素 $I=\{s_1,\cdots,s_c\}$,不失一般性,假定 $s_1\leqslant\cdots\leqslant s_c$(可通过一个伪随机排列算法实现). 对于 I 中的每个 s_i,User 选取一个随机数 $v_i\in\mathbb{Z}_p^*$,计算 $H_i=h(fid_F\|i)$ 和 $\omega=m\cdot e(\prod_{i\in I}H_i^{v_i},c_2)$,则质询为 $chal=\{(i,v_i)_{i\in I},c_1,c_3,\omega\}$.

4)示证生成算法 *ProofGen*:CSP 首先根据数据标识 fid_F 及质询 *chal* 找到指定的文件及抽样数据块的索引号,计算抽样数据块的线性组合值 $\mu=\sum_{i\in I}v_im_i$ 和认证标签聚合值 $\sigma=\prod_{i\in I}\sigma_i^{v_i}$;然后计算 $B=c_3^{\mu}$ 和 $m^*=\omega\cdot e(\sigma,c_1)/B$;最后将 $Prf=\{m^*\}$ 作为示证.

5)示证验证算法 *VerifyProof*:User 验证等式(3)是否成立. 若成立,输出 1,表示所抽样的数据块是可信的. 否则,输出 0,表示数据不可信.

$$m\overset{?}{=}m^* \tag{3}$$

3.2　数据访问流程

在本文方案中,数据访问流程可分为初始化设置、数据预处理、数据可信性验证和数据访问等 4 个阶段.

(1)初始化设置阶段

Owner 运行 *KeyGen* 算法,生成公私钥和公开参数,之后将其身份及公开参数等信息发送给 CA. CA 根据 Owner 提供的信息为 Owner 生成身份数字证书 $\mathrm{Cert}_{\mathrm{Owner}}$,并将该证书发送给 Owner,$\mathrm{Cert}_{\mathrm{Owner}}$ 中包含了 Owner 的身份和公开参数等信息. CA 还需为自身生成一个证书 $\mathrm{Cert}_{\mathrm{CA}}$,其中包含了 CA 的身份和公钥等信息,并将 $\mathrm{Cert}_{\mathrm{CA}}$ 上传给 CSP.

(2)数据预处理阶段

Owner 运行 *TagGen* 算法为每个数据块生成认证标签,将数据 F 及其标识符 fid_F、认证标签集合 ϕ、身份证书 $\mathrm{Cert}_{\mathrm{Owner}}$ 上传给 CSP.

(3)数据可信性验证阶段

数据可信性验证包括数据来源的可信性验证和内容的可信性验证.

数据来源的可信性验证主要是对 Owner 的身份进行验证,可分为三个步骤:

1)User 从 CSP 处获取 $\mathrm{Cert}_{\mathrm{Owner}}$ 和 $\mathrm{Cert}_{\mathrm{CA}}$;

2)User 验证 $\mathrm{Cert}_{\mathrm{Owner}}$ 的有效性,若验证未通过,则停止验证过程;验证通过则继续执行步骤 3);

3)User 验证 $\mathrm{Cert}_{\mathrm{CA}}$ 的有效性,若验证未通过,则停止验证过程;验证通过则继续执行数据内容可信性的验证过程.

数据内容的可信性验证过程为:User 运行 *ChalGen* 算法生成质询 *chal*,指定当前要抽样的数据块,并将 *chal* 发送给 CSP. CSP 收到 *chal* 后,运行 *ProofGen* 算法生成对应的示证 *Prf*,并将 *Prf* 返回给 User,User 收到 *Prf* 后运行 *VerifyProof* 以判断当前抽样的数据块是否可信.

(4)数据访问阶段

对数据的来源和内容可信性进行验证后,User 就可访问并使用数据了. User 对数据的使用可以是在线的,也可以是离线的,即 User 可直接在 CSP 上对数据进行分析,也可以将数据下载到本地进行分析. 本文主要考虑 User 对数据的离线使用过程,具体过程如下:

1)User 根据 Owner 的身份、数据文件标识符 fid_F 以及数据可信性验证的结果,向 CSP 发起数据访问请求,请求中包含了可信数据块的索引号.

2)CSP 收到 User 的数据访问请求后,根据请求中的 Owner 的身份、数据文件标识符 fid_F 以及数据块索引号,查找到相应的数据块集合及其对应的认证标签集合,返回给 User.

3)为防止 TOCTOU 攻击,User 可在本地再次验证接收到的数据的可信性. 将 User 接收的数据块索引号表示为集合 $I_1=\{i_1,\cdots,i_c\}$,则 User 可计算 $\sigma_{\mathrm{off}}=\prod_{i\in I_1}\sigma_i$,并验证等式(4)是否成立,成立则表明数据是可信的,就可使用和分析数据.

$$e(\sigma_{off},g)\cdot e(\prod_{i\in I_1}H_i,R)\overset{?}{=}A^{\sum_{i\in I_1}m_i} \tag{4}$$

4　安全性分析

4.1　正确性

对等式(3)右边展开并进行等价替换和运算,即可验证本文质询-应答协议的正确性.

$$\begin{aligned}m^* &= \omega\cdot e(\sigma,c_1)/B\\ &= m\cdot e(\prod_{i\in I}H_i^{v_i},c_2)\cdot e(\prod_{i\in I}\sigma_i^{v_i},g^t)/B\\ &= m\cdot e(\prod_{i\in I}H_i^{v_i},g^{-rt})\cdot e(\prod_{i\in I}(X^{m_i}H_i)^{rv_i},g^t)/B\end{aligned}$$

$$= m \cdot e\left(\prod_{i\in I} X^{rv_i m_i}, g^t\right)/B$$
$$= m \cdot e(X, g^r)^{t\sum_{i\in I} v_i m_i}/B$$
$$= m \cdot A^{t\mu}/B\left(A = e(X, g^r), \mu = \sum_{i\in I} v_i m_i\right)$$
$$= m(c_3 = A^t, B = c_3^{\mu})$$

等式(4)的正确性验证为:

$$e(\sigma_{off}, g) \cdot e\left(\prod_{i\in I_1} H_i, R\right)$$
$$= e\left(\prod_{i\in I_1} (X^{m_i} H_i)^r, g\right) \cdot e\left(\prod_{i\in I_1} H_i, g^{-r}\right)$$
$$= e\left(X^{\sum_{i\in I_1} m_i}, g^r\right)$$
$$= e(X, g^r)^{\sum_{i\in I_1} m_i}$$
$$= A^{\sum_{i\in I_1} m_i}$$

4.2 完备性

本文将可聚合的广播签名方案 ASBB[17] 签名算法 $\sigma = Xh(s)^r$ 改为 $\sigma = X^m h(s)^r$,验证等式则由 $e(\sigma, g)e(h(s), R) = A_1$ 变换为 $e(\sigma, g)e(h(s), R) = A^m$. 因此,本文方案中的签名算法的安全性可归约为 ASBB 方案中签名算法的安全性,也即在本文签名算法中,如果敌手在已知公钥(R, A)的情况下,能够伪造合法的签名对(m, s, σ),那么该敌手在 ASBB 方案中,就能够在已知公钥$(R^{m^{-1}}, A)$的情况下,伪造合法的签名对$(s, \sigma^{m^{-1}})$. 但是,文献[17]已经证明了在 ASBB 方案中,这种概率是可忽略的. 因此,本文的签名算法是安全的.

在签名算法是安全的情况下,等式(4)利用聚合特性[19]验证了下载到本地的离线数据的可信性,其完备性也可以保证[17]. 下面将通过定理 1 证明本文的挑战-应答协议是完备的.

定理 1 如果生成数据认证标签的签名方案是不可伪造的,并且 DL 问题和 CDH 问题在双线性映射中是难解的,那么在随机预言模式下,本文的数据内容可信性验证协议是完备的.

证明 本文的交互式质询-应答协议的完备性证明过程与文献[7]的证明过程类似,由于篇幅有限,本文省略其详细的证明过程.

5 实验及分析

本小节主要对本文所提方案中的交互式质询-应答协议的性能进行评估分析.

5.1 实验环境

本文实验环境配置为 Intel Core(TM)2 Quad 2.67GHz,2GB 内存,操作系统为 Ubuntu 14.04,内核版本为 3.13,编程语言为 C,PBC 库版本号为 0.5.7,OpenSSL 库版本号为 0.9.8,椭圆曲线为 MNT 曲线,有限域的大小为 159 比特,安全系数的长度为 80 比特,也即$|v_i| = 80, |p| = 160$.

5.2 实验结果及分析

本文选取 5 个不同大小的文件测试 *TagGen* 算法的运行时间. 文件大小分别为 128MB,256MB,512MB,1024MB 和 2048MB,每个数据块的大小设置为 4KB. 对于每个文件,运行 *TagGen* 算法 20 次,最后计算平均值. 实验结果如图 2 所示,从图中可知,生成文件的认证标签需要花费较多的时间. 但在本文方案中,对于每个文件,*TagGen* 算法只需运行一次,而且可采用并行化方式计算数据块的认证标签,生成的认证标签可在之后的验证过程中重复使用,无需再运行 *TagGen* 算法重新生成. 因此,该开销是可以接受的.

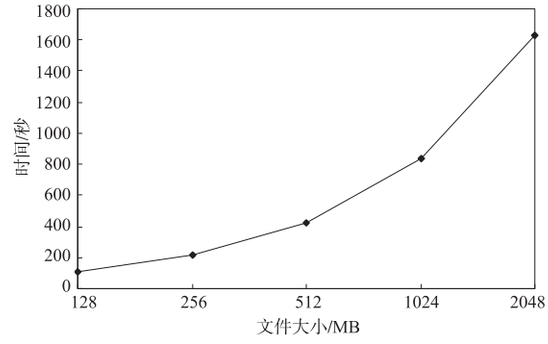

图2 标签生成时间

本文的交互式质询-应答协议是基于简单随机抽样设计的,其概率分析与文献[4]相同,即当有 1% 的数据与其认证标签不匹配时,抽样块数分别为 300 和 460 能分别达到 95% 和 99% 的检测概率. 因此,本文实验中选取的数据块数也为 300 和 460. 表 1 给出了本文方案与 PDP 方案中两篇代表性文献[7]和[12]的协议运行时间对比. 其中,文献[7]中的参数 s 被设置为 1,以便在同一基准下做比较. 从表 1 中可知,本文方案整体的时间开销与文献[7]和文献[12]的时间开销相当. 但本文方案在数据内容的示证验证阶段的计算和时间开销远低于文献[7]和文献[12]. 该结果使得本文方案在实际应用中更具有灵活性和可扩展性,即在图 1 所示的通用框架中,User 可采取三种验证方式:1)独立验证. User 单独与 CSP 交互完成整个验证过程;2)委托验证. User 委托一个第三方验证者(Third Party Auditor,TPA)(如同 PDP 和 PoR 模型中的公开验证方案)替其与 CSP 交互,执行验证过程,并将验证结果返回给 User;3)共同验证. User、TPA 和 CSP 三者互相交互完成数据的可信性验证,此时,可由 TPA 完成数字证书的验证(数字证书的验证也可由 User 完成)和执行 *ChalGen* 算法,并将随机元素 m 发送给 User,User 接收来自 CSP 的示证 *Prf* 后,只需执行最后的 *VerifyProof* 算法,即可完成验证过程. 在 User 不考虑数据的本地验证和不想仅依赖于

TPA 返回的结果来判断数据的可信性时,可选择第三种验证方式. 此时,User 无需具备复杂的密码算法运算能力(主要是双线性配对运算),也无需获取或计算更多的其他参数,就可参与到验证过程中,以很小的计算和时间开销得到验证结果;而且还可进一步减小计算和通信开销,即 TPA 发送随机元素 m 的哈希值给 User、CSP 发送示证 *Prf* 的哈希值给 User,User 比较两者的哈希值就可完成验证过程.

表 1　本文方案与文献[7]和文献[12]的协议时间对比

	本文方案		文献[7]		文献[12]	
抽样块数(c)	300	460	300	460	300	460
ChalGen 时间(毫秒)	447.575	696.541	238.179	345.029	231.799	352.257
ProofGen 时间(毫秒)	223.481	347.165	225.892	352.293	234.018	349.738
VerifyProof 时间(毫秒)	0.001	0.001	236.295	352.514	266.724	368.085

下面分析本文的质询-应答协议的存储开销和通信开销. 本文方案为每个数据块都生成了认证标签,在椭圆曲线选定后,每个认证标签的大小为常量. 在本文实验中,每个认证标签的大小约为 20B,与采用 BLS 签名方案[20]生成的认证标签在长度上是相当的,且安全强度与 BLS 签名方案一致;并且在相同安全强度下,本文认证标签的长度小于 RSA 签名方案[4]生成的认证标签的长度. 当每个数据块的大小为 4KB 时,消耗的存储空间比约为 0.49%,在合理的开销范围内. 本文方案的协议通信开销主要包括质询 *chal* 和示证 *Prf*,其中,示证 *Prf* 为常量,约为 20B;质询 *chal* 中包括常量和变量,变量部分与数据大小和抽样数据块数正相关. 由于抽样 460 块数据就可达到很高的置信率[4],假定可用 9B 来表示索引号(此时,数据总量将达到 ZB 级别),那么在抽样值为 460 时,通信开销为 8740B,总的通信开销约为 8.6KB,相对于总体数据量和验证效果而言,在可接受的范围内. 与文献[12]相比,本文方案多了 20B 的通信开销,但简化了最后的验证过程,使得用户可根据需求选取不同的验证方式.

6　结束语

针对大数据场景下的数据可信性验证问题,本文提出了一种大数据可信性验证方法,能够对数据的来源和内容的可信性进行验证. 本文通过验证数据属主的数字身份证书实现对大数据来源的可信性验证. 本文基于简单随机抽样和可聚合的广播签名方案设计了一个交互式的质询-应答协议,使得验证者只需抽样少量数据块,就可以高置信率验证大数据内容的可信性. 本文通过数字证书在流程上实现了数据来源和数据内容的可信性的顺序验证,并支持在本地对数据内容再次进行可信性验证,而且还可根据实际需求选取不同的验证方式. 本文对所设计方案的安全性进行了形式化的分析,结合实验及理论分析评估了方案的主要性能开销,并与其他工作进行了对比和分析. 理论分析和实验结果分析表明,本文方法的性能开销在合理的可接受范围内,能达到大数据可信性验证的目的.

参考文献

[1] 冯登国,张敏,李昊. 大数据安全与隐私保护[J]. 计算机学报,2014,37(1):246 – 258.

[2] 魏凯敏,翁健,任奎. 大数据安全保护技术综述[J]. 网络与信息安全学报,2016,2(4):1 – 11.

[3] 方滨兴,贾焰,李爱平,等. 大数据隐私保护技术综述[J]. 大数据,2016,2(1):1 – 18.

[4] G Ateniese,R Burns,R Curtmola,et al. Provable data possession at untrusted stores[A]. Proceedings of 14th ACM Conference on Computer and Communications Security[C]. ACM,2007. 598 – 609.

[5] G Ateniese,R Burns,R Curtmola,et al. Remote data checking using provable data possession[J]. ACM Transactions on Information and System Security,2011,14(1):12 – 34.

[6] A Juels,B Kaliski. PORs: Proofs of retrievability for large files[A]. Proceedings of 14th ACM Conference on Computer and Communications Security[C]. ACM,2007. 584 – 597.

[7] H Shacham,B Waters. Compact proofs of rretrievability[J]. Journal of Cryptology,2013,26(3):442 – 483.

[8] 李晖,孙文海,李凤华,等. 公共云存储服务数据安全及隐私保护技术综述[J]. 计算机研究与发展,2014,51(7):1397 – 1409.

[9] 谭霜,贾焰,韩伟红. 云存储中的数据完整性证明研究及进展[J]. 计算机学报,2015,38(1):164 – 177.

[10] F Armknecht,J Bohli,G Karame,et al. Outsourced Proofs of Retrievability[A]. Proceedings of 21st ACM Conference on Computer and Communications Security[C]. ACM,2014. 831 – 843.

[11] C Guan,K Ren,F Zhang,et al. Symmetric-key based proofs of retrievability supporting public verification[A]. Proceedings of 20th European Symposium on Research in Computer Security[C]. Springer,2015. 203 – 223.

[12] C Wang,Q Wang,K Ren,et al. Privacy-preserving public auditing for data storage security in cloud computing[A].

Proceedings of 2010 IEEE International Conference on Computer Communication[C]. IEEE,2010. 1 – 9.

[13] Z Ren,L Wang,Q Wu,et al. Data dynamics enabled privacy-preserving public batch auditing in cloud storage [J]. Chinese Journal of Electronics,2014,23(2): 297 – 301.

[14] M Etemad,A Küpçü. Transparent,distributed,and replicated dynamic provable data possession[A]. Proceedings of 11th International Conference on Applied Cryptography and Network Security[C]. Springer,2013. 1 – 18.

[15] C Erway,A. Küpçü,C Papamanthou,et al. Dynamic provable data possession[J]. ACM Transactions on Information and System Security,2015,17(4):15.

[16] Z Ren,L Wang,Q Wang,et al. Dynamic proofs of retrievability for coded cloud storage systems[J]. IEEE Transactions on Services Computing,2015,DOI: 10. 1109/TSC. 2015. 2481880.

[17] Q Wu,Y Mu,W Susilo,et al. Asymmetric group key agreement[A]. Proceedings of 28th Annual International Conference on Theory and Applications of Cryptography Techniques[C]. Springer,2009. 153 – 170.

[18] Y Zhu,H Wang,Z Hu,et al. Zero-knowledge proofs of retrievability[J]. Science China: Information Sciences,2011,54(8):1608 – 1617.

[19] D Boneh,C Gentry,B Lynn,et al. Aggregate and verifiably encrypted signatures from bilinear maps[A]. Proceedings of 22nd International Conference on Theory and Applications of Cryptographic Techniques[C]. Springer,2003. 416 – 432.

[20] D Boneh,B Lynn,H Shacham. Short signatures from the weil pairing[J]. Journal of Cryptology,2004,17(4):297 – 319.

作者简介

任正伟 男,1986 年 4 月出生,湖北武汉人. 2014 年于武汉大学获工学博士学位. 现为武汉数字工程研究所工程师,主要从事应用密码学、数据安全、可信计算等方面的研究工作.

E-mail: zhengwei_ren@163. com

孙小雁(通讯作者) 女,1981 年 8 月出生,广西玉林人. 2010 年于桂林电子科技大学获硕士学位. 现为玉林师范学院副教授,主要从事信息安全方面的研究工作.

E-mail: jgxysxy@126. com