中国科学: 信息科学 2021 年 第 51 卷 第 11 期: 1885–1899
SCIENTIA SINICA Informationis
论文

基于轻量级区块链的隐私保护传染病监测数据聚合

胡柏吉, 李元诚*, 房方, 商兴宇

华北电力大学控制与计算机工程学院, 北京 102206
* 通信作者. E-mail: ycli@ncepu.edu.cn

收稿日期: 2021–01–06; 修回日期: 2021–03–09; 接受日期: 2021–03–30; 网络出版日期: 2021–10–22

中央高校基本科研业务费专项 (批准号: 2020YJ003) 资助项目

摘要 随着 COVID-19 在全球肆虐, 传染病监测有利于阻止传染病传播. 而保护监测系统中病人以及数据提供者的隐私能免除他们对隐私信息泄露的顾虑, 从而提高系统的数据采集能力. 本文提出了一个基于区块链的传染病监测数据聚合方案 (lightweight-blockchain based privacy-preserving data aggregation scheme, LBPDA), 可以在不依赖可信第三方的情况下对数据进行聚合. 特别地, 为了保护数据聚合过程中数据隐私, 采用基于 Paillier 密码系统的加法同态性来聚合病例数据. 此外, 为了减少时间开销和存储开销, 对使用的 Hyperledger Fabric 联盟区块链平台进行了改进, 从而轻量化数据聚合过程. 我们对提出的方案进行了仿真, 并进行安全性和性能分析以验证提出方案的可行性和有效性. 结果显示, 提出的方案能满足政府部门在聚合病例数据用于传染病监测的同时保护病人和数据提供者的隐私的需求. 此外, 通过对比也证明本文对区块链的轻量化是有效的.

关键词 区块链, 传染病监测, 隐私保护数据聚合, Paillier 密码系统, Hyperledger Fabric

1 引言

为了通过仪表板和统计视图实时展示传染病传播, 快速检测新型传染病爆发, 有必要对来自医院、疾病预防与控制中心 (centers for disease prevention and control, CDC)、医学检验实验室、诊所等公共卫生机构的传染病相关数据进行高效、安全、实时的聚合. 数据聚合是疾病监测系统必不可少的一部分, 对特定综合症病人进行计数可以发现有意义的趋势, 而单个病例数据通常不足以检测传染病. 一些人口统计的和地理统计的信息有助于检测和定位在人口统计学以及地理上定义的传染病. 尽管单个病人水平数据不足以支撑监测系统, 仍然需要确保病人和数据提供者不能被从聚合数据中识别出来, 特别是当总体数据量很小的时候 [1]. 一方面传染病数据中可包括国籍、测序数据、数据提供者、致病属性、种类等元数据. 这其中一些属性比较敏感, 可以关联到病人和暴露隐私信息, 从而对其造成精

引用格式: 胡柏吉, 李元诚, 房方, 等. 基于轻量级区块链的隐私保护传染病监测数据聚合. 中国科学: 信息科学, 2021, 51: 1885–1899, doi: 10.1360/SSI-2021-0002
Hu B J, Li Y C, Fang F, et al. Lightweight-blockchain based privacy-preserving data aggregation for epidemic disease surveillance (in Chinese). Sci Sin Inform, 2021, 51: 1885–1899, doi: 10.1360/SSI-2021-0002

神、经济等方面的损失. 另一方面, 一些医生、诊所、医院等数据提供者不希望计数数据跟他们联系起来, 从而被别人知道已经受到传染病爆发的影响. 这样会导致数据提供者因为隐私、机密性方面存在的顾虑不愿提交传染病相关数据. 因此, 为了促进监测系统中病人和数据提供者的参与度, 从而提高系统数据采集能力, 需要保护病人和数据提供者的隐私.

尽管在疾病监测数据聚合的隐私保护问题上已有很多研究, 但仍有一些等待解决的问题. 第 1 个挑战是大多数据聚合方案中, 通常由一个可信第三方担任聚合节点 [2], 但其存在单点故障问题, 容易受到恶意攻击, 从而导致数据聚合服务不可用. 第 2 个挑战是大部分方案采用假名或者匿名的方式解决用户身份隐私问题, 一方面存在假名管理复杂和资源消耗大的问题, 另一方面通过数据的关联分析, 敌手仍然有大概率可以发现用户的真实身份和敏感信息之间的联系 [3]. 第 3 个挑战是, 目前传染病监测很大程度上依赖医疗保健网络聚合个人健康信息, 典型的医疗保健系统通常以分层的方式组织架构, 囊括不同行政级别的卫生行政部门、诊所、医院、家庭医生等, 为了迅速应对疾病的爆发, 系统需要支持实时的数据聚合, 支持跨组织的分布式数据源, 然而这会导致采集的数据量大而复杂, 对系统的时间和存储开销方面提出了较高的性能要求.

为了解决上述挑战, 我们提出以下解决方案. 首先, 通过区块链的分布式共识算法选出主导节点临时充当聚合者来聚合传染病数据. 这种方式推选出的聚合者节点具有较高的可信性, 且只要大部分区块链节点未受到攻击, 系统都能正常提供数据聚合服务. 此外, 传染病数据保存在区块链上, 区块链的不可篡改性和可追溯性能保证数据完整性和认证. 相比于其他实时监测系统, 基于区块链的监测系统具有更好的可扩展性、安全性和互操作性 [4]. 其次, 我们同时考虑身份隐私和数据隐私, 其中身份隐私采用身份混淆 (identity mixer) 机制实现身份的匿名性以及不可追踪性, 数据隐私采用 Paillier 同态加密算法 [5]. 身份混淆机制是 Fabric 中提供的特别功能, 通过使用零知识证明来保持身份匿名和不可链接性, 可以防止敌手对用户认证信息的关联分析 [6]. 同态加密是一种无需对加密数据进行解密就可以执行计算的方法, 可以实现数据聚合的同时实现隐私保护. 最后, 考虑到目前大部分区块链项目使用 ECDSA (elliptic curve digital signature algorithm) 作为其签名算法, 无法进行签名聚合或密钥聚合, 只能挨个对签名进行验证. 在交易数量较大的时候, 逐个验证签名会耗费大量的时间. 相比之下, BLS [7] 短签名算法具有可批量签名验证特性和较短的签名大小 (签名大小是 ECDSA 的一半). 在区块链中采用 BLS 作为签名算法, 并对区块和交易的存储结构进行调整, 可以减少聚合过程中的时间和存储开销.

区块链可以被分为两种类型: 公有链和许可链. 公有链允许任何人加入网络, 并分享和获取可用数据. 它是完全去中心化的, 交易信息公开, 不利于交易数据隐私保护. 私有链是许可链的极端情况, 由一个实体控制, 实际上已经退化成一个中心化的系统 [8]. 联盟链是许可链的常见实现形式, 它由多个组织共同维护, 是部分去中心化的 [9]. 相比公有链, 联盟链只允许联盟内部的透明管理, 只有联盟里的机构及其用户才有权限访问数据, 在数据隐私保护上具有优势. 因此本文基于 Hyperledger Fabric [10] 这一具有代表性的联盟链来设计分布式的、保护隐私的传染病监测数据聚合方案.

本文的贡献如下: 设计了基于联盟链 Fabric 的传染病监测数据聚合架构, 无需可信任第三方, 以促进传染病监测数据聚合的安全性和可靠性; 为了消除数据提供者和病人对私密信息泄露的顾虑, 引入身份混淆机制和 Paillier 密码系统来保护数据聚合过程中的身份和数据隐私, 以提高传染病监测系统的数据采集能力; 改进区块链中签名的验证机制以及交易和区块存储结构, 从而减少系统的时间开销和存储开销. 本文其余部分组织如下: 在第 2 节总结了相关工作并在第 3 节介绍了一些预备知识; 第 4 节详细描述了提出的 LBPDA (lightweight-blockchain based privacy-preserving data aggregation scheme) 方案; 第 5 和 6 节分别对提出的系统进行安全和性能评估. 最后在第 7 节总结了当前工作并对未来工作进行展望.

2 相关工作

目前已有许多基于同态加密的方案用于数据聚合中的隐私保护. Lu 等 [11] 提出一个基于 Paillier 密码系统的同态加密方案来保护多维数据聚合的隐私, 并证明智能电网运营中心在选择明文攻击下是语义安全的. He 等 [12] 基于 BGN (Boneh-Goh-Nissim) 密码系统 [13] 提出了一种可抵抗内部攻击者的隐私保护数据聚合方案, 相比于基于双线性对的方案具有更高的计算效率. Li 等 [14] 提出一种基于 Paillier 的保护隐私的多子集数据聚合方案, 在满足智能电网控制中心对数据细粒度要求的同时保护了用户隐私. Zhang 等 [15] 提出一个轻量级的、可验证的用于智能物联网网络边缘计算数据聚合隐私保护方案, 采用 Paillier 密码系统用于同态加密, 并证明该方案在选择明文攻击下是不可伪造的. Emam 等 [16] 以保护病人和数据提供者的隐私为背景, 提出一种多方安全计算协议用于疾病监测的数据聚合. 但这些方案大部分用于智能电网、物联网中的数据聚合场景, 没有解决集中式数据聚合方案依赖可信第三方聚合者而存在的单点故障, 容易受到网络攻击等问题.

近年来, 区块链由于其去中心化、不可篡改等特性, 开始被引入电子医疗记录、电子健康记录、个人健康记录和疾病监测等公共医疗卫生的隐私保护、数据共享等. Xu 等 [17] 提出一种基于区块链的健康数据隐私保护方案用于智能医疗系统. Wang 等 [18] 将联盟区块链技术与卫生保健系统结合以实现全面的医疗数据共享、医疗记录审查和护理审计. Dagher 等 [19] 提出一种基于区块链的访问控制方案, 用于病人、数据提供者和第三方进行可互操作和有效的健康档案访问并保护病人敏感信息的隐私. Bellod Cisneros 等 [2] 将区块链引入公共卫生监测用于分布式共享基因组数据, 以解决卫生部门的互操作性问题. Li 等 [20] 提出一个基于区块链的医疗环境数据聚合方案, 来保护病人的隐私和提供个性化医疗服务.

随着区块链的发展, 也开始用于数据聚合的隐私保护. Guan 等 [3] 提出使用私有区块链来聚合数据, 并通过采用多个假名的方式来隐藏用户真实身份实现隐私保护. Wang 等 [21] 提出通过分层的区块链系统聚合智能电表数据, 并通过同态加密在聚合过程中保护单个智能电表数据的隐私. 虽然文献 [21] 提出的分布式结构解决了单点故障问题, 但是其区块链中的共识机制采用的 PBFT (practical byzantine fault tolerance) 算法, 存在通信复杂度高、可扩展性低的问题. 相比而言, 我们采用 Raft 作为共识算法, 具有比 PBFT 更高的效率 (假设一个系统可能同时有 n 个节点发生故障, PBFT 要求系统至少 $3n+1$ 个节点存在, 而 Raft 只需要 $2n+1$ 个节点即可应对故障, 具有更低的复杂性和成本). Chen 等 [22] 集成雾计算和区块链, 并结合 Paillier 加密、批处理聚合签名和匿名身份验证开发了一个安全的、具有较低计算开销的数据聚合方案. 但是该方法没有提出一种有效、智能的方法来选择聚合节点. Wang 等 [23] 提出基于区块链的安全策略来保护边缘网络中的数据聚合隐私, 并开发了新的区块生成规则来提高交易吞吐量和迟延方面的系统性能. 上述方案在不同程度上解决了数据聚合的相应问题, 但仍存在不足. 本文通过集成 Paillier 同态加密和区块链来共同保护传染病监测数据聚合中的安全与隐私, 并对区块链进行改进, 以提高数据聚合的性能.

3 预备知识

这一部分简单介绍系统模型、敌手模型、安全目标、Paillier 同态密码系统、双线性对和 BLS 短签名方案.

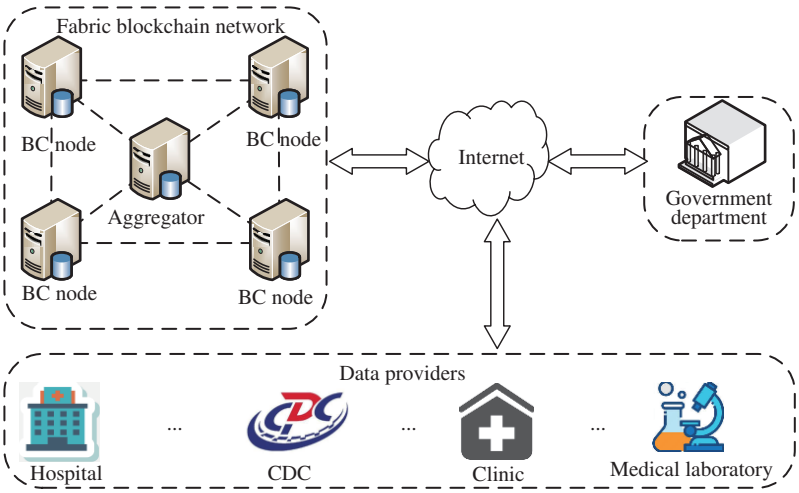


图 1 (网络版彩图) 系统模型

Figure 1 (Color online) System model

3.1 系统模型

图 1 显示本文提出的系统模型, 这也是我们提出的数据聚合方案的基础. 系统模型主要由 4 个主要实体组成: (1) 政府部门 (government department, GD); (2) 一个 Fabric 区块链网络 (Fabric blockchain network); (3) 聚合者 (aggregator, AG); (4) 数据提供者 (data providers, DP). 在我们的系统模型中, 政府部门负责启动整个系统, 分发密钥和系统参数. 此外, 它也给所有其他的实体提供注册服务. Fabric 网络包含排序服务节点 (ordering service node, OSN) 和 Peer 节点. OSN 节点通过达成共识将交易打包成区块, 而 Peer 节点共同维护一个一致性账本. 区块链节点 (即 Peer 节点和 OSN 节点) 以分布式的方式给成员用户 (即数据提供者、政府部门) 提供服务. 这种分布式架构能分散硬件在密码学计算、通信方面的负担, 使得系统更具可扩展性. 考虑到某个区块链节点可能被敌手挟持从而导致单点故障、安全等问题, 所以由 Raft 共识算法选出的主导节点来承担区块打包工作和临时担任聚合者来周期性地聚合数据上报者提供的加密传染病数据. 政府部门可利用这些数据来调整人力、物资的调度分配, 从而更好地满足疾病预防控制的需求. 数据提供者可能是医院、疾病预防与控制中心、诊所、医学检验实验室等. 他们收集原始的病例数据并生成统计数据, 通过电脑、手机等注册的智能设备周期性地将数据加密后上报到 Fabric 区块链网络. 所有这些实体通过公有的 Internet 互相通信.

3.2 敌手模型

在我们的敌手模型中, 我们认为政府部门是完全可信的组织. 聚合者的职责是辅助政府部门, 因此我们的系统模型中聚合者可以被看成一个半可信的实体. 换句话说, 它不会任意篡改传染病数据, 但是它可能想从传染病数据挖掘内嵌的私有信息, 并将其卖给感兴趣的实体.

我们假设公共 Internet 中的各个组件可能充当敌手对每个数据提供者上报的传染病数据的隐私信息感兴趣. 一个被挟持的 Internet 网络和它的各个组件 (比如路由器、交换机) 可能尝试修改或者捏造数据提供者的传染病数据. 因此 Internet 中的任何通信都可能是不安全的.

系统内部攻击者可能尝试从联盟链的账本数据中分析提取病人或者数据提供者的敏感隐私信息. 一个外部攻击者可能尝试模仿成为某一个合法的实体 (即数据提供者或聚合者) 并以它的名义发送相关数据. 此外, 外部窃听者还可能通过窃听网络传输来获取传染病数据并尝试将它们修改后转发.

3.3 安全目标

基于以上描述的系统和敌手模型, 我们的方案需要满足以下安全目标.

认证和传染病数据完整性. 所有的参与我们系统的用户都需要被政府部门授权成为合法的参与者来对抗模仿攻击. 在聚合传染病数据之前, 聚合者需要认证每个数据提供者, 从而避免聚合不准确的数据. 此外, 聚合者应该能够验证来自数据提供者的数据完整性, 从而抵抗敌手对数据的未授权修改.

传染病数据机密性. 为了保护病人和数据提供者隐私, 在传染病数据聚合过程中应该保障端到端的通信信息的机密性. 一旦病人和数据提供者的敏感数据离开智能设备, 它们应该一直保持密文状态. 这样, 即使外部或者内部敌手窃听通信信道或者从联盟区块链数据库获得传染病数据, 他们也不能解读该加密的数据来获取病人或者数据提供者的私密信息.

病人和数据提供者的隐私保护. 如前所述, 保护病人和数据提供者的隐私是传染病监测系统中一个至关重要的问题. 内部敌手, 比如聚合者, 应该不能识别病人和数据提供者的真实身份, 唯独政府部门才能知道一个病人和数据提供者的真实身份. 此外即使窃听到传染病数据, 外部敌手也不能确定数据来自于哪个特定病人和数据提供者, 也不能确定两个数据是否来自同一个病人或数据提供者.

3.4 Paillier 同态密码系统

由于目前我们的数据聚合方案只需要加法操作, 而不需要乘法操作, 部分同态加密 (partially homomorphic encryption) 可以在满足我们需求的同时具有更高的效率. 因此为了保证数据聚合过程中的机密性和隐私性, 我们使用只具有加法同态性的 Paillier 密码系统 [5] 作为同态加密算法. 其基本原理如下:

密钥生成. 令 $N = st, s, t$ 为 2 个大质数. 根据欧拉函数计算 $\phi(N) = (s-1)(t-1)$; 令 $\lambda = \mathrm{lcm}(s-1, t-1)$, lcm 表示最小公倍数. 选择 $g \in Z^*_{N^2}$. 定义函数 $L(\mu) = (\mu - 1)/N$, 并通过检查 $\mu = L(g^\lambda \bmod N^2)^{-1} \bmod N$ 的存在来确保 N 整除 g 的阶. (N, g) 和 (λ, μ) 分别为公钥 PK, 私钥 SK.

加密. 给定明文消息 $M \in Z_N = \{0, 1, \ldots, N-1\}$ 并选择随机数 $r \in Z^*_N$ 满足 $\gcd(r, N) = 1$. 然后如式 (1) 使用公钥加密明文 M 得到密文.

$$C = E(M) = g^M r^N \bmod N^2. \tag{1}$$

解密. 如式 (2) 使用私钥解密密文 $C \in Z^*_{N^2}$ 得到明文 M.

$$M = D(C) = L(C^\lambda \bmod N^2)\mu \bmod N. \tag{2}$$

加法同态. 给定明文 $M_1, M_2 \in Z_N$ 对应的密文 $C_1 = E(M_1)$, $C_2 = E(M_2)$, 则可计算如下:

$$\begin{aligned} C_1 C_2 &= E(M_1)E(M_2) \\ &= g^{M_1} {r_1}^N g^{M_2} {r_2}^N \bmod N^2 \\ &= g^{M_1+M_2}(r_1 r_2)^N \bmod N^2 \\ &= E(M_1 + M_2). \end{aligned} \tag{3}$$

因此, 在只知道明文 M_1 和 M_2 对应的密文的情况下, 可计算得到 $M_1 + M_2$ 对应的密文, 从而私钥持有者通过解密可获得明文之和.

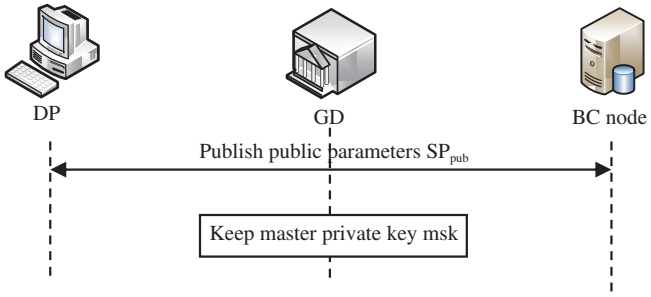


图 2 (网络版彩图) 系统初始化过程

Figure 2 (Color online) Process of the system initialization

3.5 双线性对

令 $\mathbb{G}$ 和 $\mathbb{G}_{\mathbb{T}}$ 分别表示两个阶为 p 的循环群. 令 $e:\mathbb{G}\times\mathbb{G}\to\mathbb{G}_{\mathbb{T}}$ 表示具有如下性质的映射.

双线性性. $\vee\, u,v\in\mathbb{G}$, $a,b\in\mathbb{Z}_p^*$, $e(u^a,v^b)=e(u,v)^{ab}$.

非退化性. $\exists\, u,v\in\mathbb{G}$, $e(u,v)\neq 1_{\mathbb{G}_T}$. 换句话说, 映射 e 不总是将所有的 $\mathbb{G}\times\mathbb{G}$ 对映射到 $\mathbb{G}_T$ 中的单位元.

可计算性. $\vee\, u,v\in\mathbb{G}$, 存在一个多项式时间算法可以有效地计算 $e(u,v)$.

3.6 BLS 短签名

BLS [7] 是基于双线性对的短签名方案, 相比 ECDSA, 它可以将所有交易的签名合并为一个签名并进行验证. 此外, 在相同的安全级别下, 其生成的数字签名大小只有 ECDSA 一半, 其细节如下.

密钥生成. 选择 2 个阶为质数 p 的循环群 $\mathbb{G}$, $\mathbb{G}_{\mathbb{T}}$, 和双线性对 $e:\mathbb{G}\times\mathbb{G}\to\mathbb{G}_{\mathbb{T}}$. 选择随机点 $u\in\mathbb{G}$ 作为群的生成元. 此外, 选择哈希 (Hash) 函数 $H:\{0,1\}^*\to\mathbb{G}$, 和私钥 $\mathrm{sk}\in\mathbb{Z}_p^*$, 并计算得到对应的公钥 $\mathrm{pk}=\mathrm{sk}\cdot u\in\mathbb{G}$.

签名. 假设 $M\in\mathcal{M}$ 是待签名的消息. 计算 $h=H(M)\in\mathbb{G}$ 和 $v=\mathrm{sk}\cdot h\in\mathbb{G}$.

验证签名. 验证者接收到 M 和 v 后, 会计算 $h=H(M)$, 然后检查等式 $e(v,u)=e(h,\mathrm{pk})$ 是否成立.

4 LBPDA 方案

这一部分详细介绍基于联盟区块链的隐私保护传染病数据聚合方案的运行机制. 该方案由几个阶段组成, 包括系统初始化、注册、零知识证明和交易生成、区块生成和数据聚合、聚合数据解读.

4.1 系统初始化

初始化过程如图 2 所示. 首先政府部门通过 Paillier 同态密码系统生成对应的公私钥对 $\{\mathrm{PK_P},\mathrm{SK_P}\}=\{(N,g),(\lambda,\mu)\}$. 然后, 政府部门选择定义在有限域 F_p 下的椭圆曲线 E. 其次, 政府部门构造双线性映射 $e:\mathbb{G}\times\mathbb{G}\to\mathbb{G}_{\mathbb{T}}$, $\mathbb{G}$ 和 $\mathbb{G}_{\mathbb{T}}$ 都是阶为 p 的循环群, 其中群 $\mathbb{G}$ 的元素为 E 上的点, 群 $\mathbb{G}_{\mathbb{T}}$ 的元素来自于有限域 F_{p^2}. 之后, 政府部门定义一个单向哈希函数 $H:\{0,1\}^*\to\mathbb{G}$, 并随机选择 $G\in\mathbb{G}$ 作为群 $\mathbb{G}$ 的生成元. 此外, 我们假设一个特定区域的数据提供者的数量为 σ. 最后政府部门发布系统公共参数 $\mathrm{SP_{pub}}=\{N,g,p,q,\mathbb{G},\mathbb{G}_{\mathbb{T}},e,E,H,G,\sigma\}$ 和本地安全存储主私钥 $\mathrm{msk}=\{\lambda,\mu\}$.

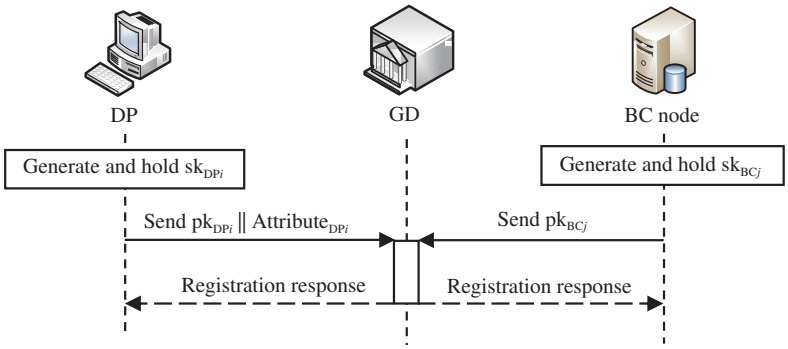


图 3 (网络版彩图) 注册过程

Figure 3 (Color online) Process of the registration

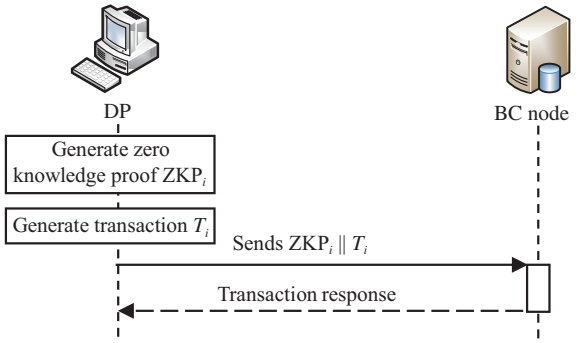


图 4 (网络版彩图) 零知识证明和交易生成过程

Figure 4 (Color online) Process of the zero knowledge proof and transaction generation

4.2 注册

当一个数据提供者的智能设备第 1 次参与 LBPDA 系统, 以及区块链节点第 1 次启动服务, 它们会被要求注册到政府部门从而起到认证的作用. 注册过程如图 3 所示, 描述如下.

数据提供者用户注册. DP_i 选择一个随机数 $\mathrm{sk}_{\mathrm{DP}_i} \in Z_p^*$ 作为它的私钥, 并为了认证, 注册相关的公钥 $\mathrm{pk}_{\mathrm{DP}_i} = \mathrm{sk}_{\mathrm{DP}_i} \cdot G$ 到政府部门. 此外, 数据提供者也在政府部门注册自己的属性 $\mathrm{Attributes}_{\mathrm{DP}_i}$, 经过验证后其属性以电子证书 (以下称此证书为凭证) 的形式签发, 存储在数据提供者的智能设备中.

区块链节点注册: 同样的, 区块链节点 BC_j 随机选择一个数 $\mathrm{sk}_{\mathrm{BC}_j} \in Z_p^*$, 和注册对应的公钥 $\mathrm{pk}_{\mathrm{BC}_j} = \mathrm{sk}_{\mathrm{BC}_j} \cdot G$ 到政府部门.

4.3 零知识证明和交易生成

每个数据提供者 DP_i 会收集传染病数据 D_i, $i = \{1, 2, \ldots, \sigma\}$ 并输入智能设备. 图 4 显示了零知识证明和交易生成过程, 其详细步骤如下.

零知识证明生成: 为了保证身份的匿名性和不可链接性, 设备生成一个零知识证明 (zero knowledge proof, ZKP_i) 来证明自己拥有在政府部门注册过的凭证, 并且只选择性的公开自己想公开的属性. ZKP_i 因为是零知识的, 所以不会向任何人透露任何额外信息. 此外, 同一个 DP_i 生成的不同 ZKP_i 不能进行关联分析.

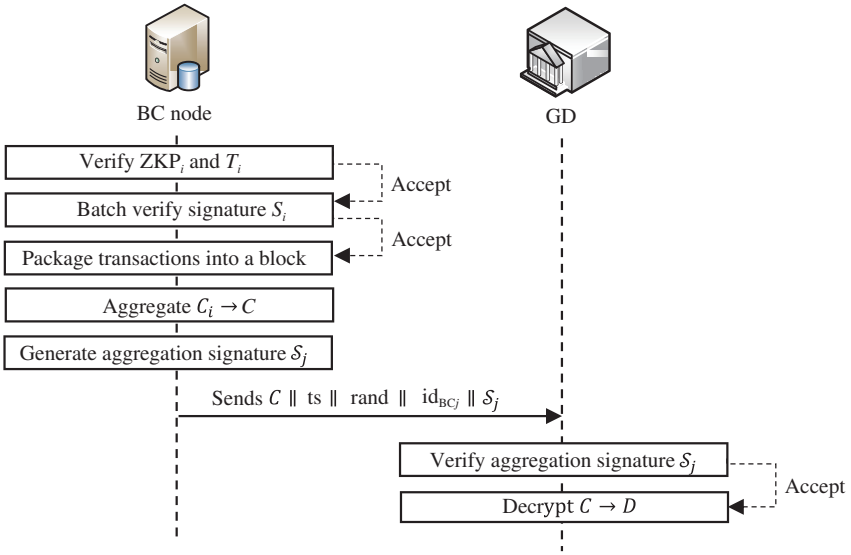


图 5 (网络版彩图) 区块生成、数据聚合和解密

Figure 5 (Color onlin) Process of the block generation, data aggregation and decryption

交易生成. 为了保证数据的机密性, 设备会使用 Paillier 对应的公钥 $\mathrm{PK_P}$ 对数据 D_i 进行加密:

$$C_i = g^{D_i} \cdot {r_i}^N \bmod N^2. \tag{4}$$

为了保证数据的完整性和数据来源的可靠性, 设备使用 BLS 短签名算法对密文数据进行签名:

$$\mathcal{S}_i = \mathrm{sk}_{\mathrm{DP}_i} \cdot H(C_i \| \mathrm{ts} \| \mathrm{rand} \| \mathrm{id}_{\mathrm{DP}_i} \| V_n). \tag{5}$$

最终 DP_i 通过智能设备将相关数据封装成一个交易 T_i, 并发送到最近的区块链节点, T_i 的结构如下:

$$T_i = C_i \| \mathrm{ts} \| \mathrm{rand} \| \mathrm{id}_{\mathrm{DP}_i} \| V_n \| \mathcal{S}_i, \tag{6}$$

其中 ts 表示交易生成的时间戳, rand 表示一个随机数, 它们一起用于抵抗重放攻击. $\mathrm{id}_{\mathrm{DP}_i}$ 表示 DP_i 的身份, 其对除了数据提供者的所有其他实体匿名. V_n 表示交易版本号.

4.4 区块生成和数据聚合

如图 5 所示, 在收到来自数据提供者的 ZKP_i 和 T_i 后, 首先区块链节点使用从政府部门获取的公钥验证零知识证明 ZKP_i 是否符合访问控制策略和检查 T_i 的时间戳 ts、随机数 rand、身份 $\mathrm{id}_{\mathrm{DP}_i}$ 和交易版本号 V_n 的有效性. 验证都通过后, T_i 被广播到联盟链网络被其他区块链节点存储到本地内存池. 最后联盟链网络通过 Raft 共识算法临时推选出的主导节点, 一方面对交易进行批量签名验证后, 将它们打包成区块, 并广播到联盟链网络被所有记账区块链节点 (即 Peer 节点) 添加到本地区块链账本, 另一方面主导节点也起到聚合者的作用, 在无需解密传染病数据的情况下, 对签名验证通过的交易中的数据进行聚合.

批量签名验证. 如式 (7) 所示, 当共识时间到达, 主导节点开始利用 BLS 短签名方法的批量验证特性来验证所有新增加的交易中的签名.

$$e\left(G, \sum_{i=1}^{\sigma} \mathcal{S}_i\right) = \prod_{i=1}^{\sigma} e(\mathrm{pk}_{\mathrm{DP}_i}, H(C_i \| \mathrm{ts} \| \mathrm{rand} \| \mathrm{id}_{\mathrm{DP}_i} \| V_n)). \tag{7}$$

该等式的正确性来自于双线对的特性, 证明如下:

$$\begin{aligned} e\left(G,\sum_{i=1}^{\sigma}\mathcal{S}_i\right) &= e\left(G,\sum_{i=1}^{\sigma}\mathrm{sk}_{\mathrm{DP}_i}\cdot H(C_i\|\mathrm{ts}\|\mathrm{rand}\|\mathrm{id}_{\mathrm{DP}_i}\|V_n)\right)\\ &= \prod_{i=1}^{\sigma} e(G,\mathrm{sk}_{\mathrm{DP}_i}\cdot H(C_i\|\mathrm{ts}\|\mathrm{rand}\|\mathrm{id}_{\mathrm{DP}_i}\|V_n))\\ &= \prod_{i=1}^{\sigma} e(\mathrm{sk}_{\mathrm{DP}_i}\cdot G,H(C_i\|\mathrm{ts}\|\mathrm{rand}\|\mathrm{id}_{\mathrm{DP}_i}\|V_n))\\ &= \prod_{i=1}^{\sigma} e(\mathrm{pk}_{\mathrm{DP}_i},H(C_i\|\mathrm{ts}\|\mathrm{rand}\|\mathrm{id}_{\mathrm{DP}_i}\|V_n)). \end{aligned} \tag{8}$$

区块生成. 验证通过的交易被打包成区块后广播到联盟链网络中, 最终每个区块链节点都将新生成区块链接到本地区块链账本, 从而维护一个一致性的账本. 需要特别指出的是, 在将交易打包到区块过程中, 无需保留每个交易的签名字段, 只需保存一个所有签名聚合 $\sum_{i=1}^{\sigma}\mathcal{S}_i$ 以证明该区块所有交易的合法性, 这样可以轻量化区块链的数据存储开销. 因为一旦新生成的区块被连接到区块链, 之后区块链的不可篡改, 不可删除的特性可保证该区块中所有交易中的数据完整性和认证.

数据聚合. 因为推选出的主导节点具有较高的可信性, 可以临时充当聚合者节点, 可以基于 Paillier 的加法同态性质, 对新增区块中加密的传染病数据 C_i 进行数据聚合, 其过程如下:

$$C=\prod_{i=1}^{\sigma}C_i \bmod N^2. \tag{9}$$

聚合签名. 然后聚合者使用它的私钥 $\mathrm{sk}_{\mathrm{AG}}$ 根据式 (10) 对聚合的数据进行签名.

$$\mathcal{S}_{\mathrm{AG}}=\mathrm{sk}_{\mathrm{AG}}\cdot H(C\|\mathrm{ts}\|\mathrm{rand}\|\mathrm{id}_{\mathrm{AG}}). \tag{10}$$

最终, 数据 $C\|\mathrm{ts}\|\mathrm{rand}\|\mathrm{id}_{\mathrm{AG}}\|\mathcal{S}_{\mathrm{AG}}$ 被发送到政府部门.

4.5 聚合数据解读

收到聚合者发来的数据后, 政府部门在验证其数据完整性、来源和确定它是非重放数据的前提下对其进行解密, 从而最终得到聚合的数据总数, 详细过程如下.

签名验证. 政府部门首先检查来自聚合者数据的时间戳 ts、随机数 rand、身份 $\mathrm{id}_{\mathrm{AG}}$ 的有效性. 然后政府部门通过验证等式 (11) 是否成立来验证签名.

$$e(G,\mathcal{S}_{\mathrm{AG}})=e(\mathrm{pk}_{\mathrm{AG}},H(C\|\mathrm{ts}\|\mathrm{rand}\|\mathrm{id}_{\mathrm{AG}})). \tag{11}$$

传染病聚合数据解读. 聚合的密文数据 C 如式 (12) 所示, 因为其满足 Paillier 密码系统的密文形式, 所以政府部门可以用式 (13) 对其解密.

$$C=\prod_{i=1}^{\sigma}C_i \bmod N^2=\prod_{i=1}^{\sigma}g^{D_i}\cdot {r_i}^N \bmod N^2=g^{\sum_{i=1}^{\sigma}D_i}\cdot\left(\prod_{i=1}^{\sigma}r_i\right)^N \bmod N^2, \tag{12}$$

$$D=\sum_{i=1}^{\sigma}D_i=\frac{L(C^{\lambda} \bmod N^2)}{L(g^{\lambda} \bmod N^2)} \bmod N. \tag{13}$$

最终, 政府部门可获取聚合的传染病数据 D.

5 安全性分析

本节针对第 3 节描述的敌手模型和安全目标给出了详细的安全属性分析, 主要集中考虑机密性和隐私保护、认证和数据完整性.

5.1 机密性和隐私保护

在 LBPDA 方案中, 我们利用 Paillier 密码系统来加密机密数据, 并基于它的加法同态性质来聚合密文数据. 数据的机密性和隐私保护以及身份隐私可以由如下几个方面保证.

首先, 在交易生成阶段, 数据提供者提供的传染病数据 D_i 被加密成标准的 Paillier 密码系统密文形式. 因为 Paillier 密码系统基于决定性的复合阶剩余假设从而被证明是语义安全的, 可以抵抗选择明文攻击 [5], 从而可以保证机密信息不会被泄露. 其次, 在数据聚合阶段, 聚合者不能在无法获取私钥 (λ, μ) 的前提下从密文中恢复明文信息, 但是可以直接聚合收到的密文, 且得到的结果是有效的 Paillier 密码系统密文形式. 因此, 即使聚合者节点是半可信的, 数据提供者上报的传染病数据的机密性和隐私性都可以得到保证.

在提出的方案中, 数据提供者无需获取单个病人的身份信息, 只需要收集特定类型的统计数据. 内部或者外部窃听者都无法从这些数据识别一个病人的身份, 所以病人身份隐私得到保护. 此外, 我们采用身份混淆 (identity mixer) 机制实现交易发起者 (即数据提供者) 的匿名性以及不可关联性. 匿名性表示只有政府部门知道数据提供者的真实身份而其他实体只知道数据提供者的匿名身份. 不可关联性表示当一个数据提供者发送多笔交易时, 无法揭示这些交易是否发送自同一个数据提供者. 因此数据提供者的身份隐私也同样不会受到内部或者外部攻击者的侵害.

5.2 认证和数据完整性

LBPDA 方案可以认证数据提供者和聚合者发送的数据的来源. 数据提供者和临时充当聚合者的区块链节点都将自己的真实信息和公钥在政府部门注册, 从而保证系统中的数据发送方和接收方都是合法实体. 我们的方案中不管是数据提供者提供的交易数据还是聚合者聚合的密文数据都通过 BLS 短签名方案签名. 考虑到 BLS 短签名方案是可证明安全的, 在随机预言机模型下的自适应选择明文攻击中是不可伪造的, 其安全性是基于计算的迪菲 – 赫尔曼 (computational Diffie-Hellman) 假设的. 这样一来, 即使内部或者外部攻击者能篡改消息, 签名验证程序都不能得到正确的结果, 从而被篡改的数据都被视为无效, 数据完整性得到保证. 此外, 数据一旦存储到区块链上, 区块链的特性可以保证链上数据的完整性和不可伪造性.

6 性能评估

本节评估 LBPDA 方案时间和存储开销方面的性能. 我们使用改进 Fabric 的 LBPDA 与原始 Fabric 进行比较来证明我们方案的有效性. 不失一般性地, 我们使用 JPBC (Java pairing-based cryptography) 库来实施 Paillier 密码系统、ECDSA 和 BLS. 区块链部分基于 Fabric v2.2.1 实施来验证它的可行性和有效性. 数据提供者节点、区块链节点、政府部门节点各自安装在 2.20 GHz 4vCPU 8 GB 内存、运行 CentOS 7.4 的虚拟机上. Fabric-sdk-java[1)]和 Fabric-chaincode-java[2)]分别用于开发 Fabric 区块链客户端和智能合约.

1) Fabric-sdk-java. https://github.com/hyperledger/fabric-sdk-java.
2) Fabric-chaincode-java. https://github.com/hyperledger/fabric-chaincode-java.

表 1 密码操作的记法和时间开销

Table 1 Notations and time overhead of cryptographic operations

Notation	Description	Time cost (ms)
TS_E	Signature in ECDSA	1.011
TV_E	Signature verification in ECDSA	1.618
TS_B	Signature in BLS	21.370
TV_B	Signature verification in BLS	612.474
TE_P	Encryption in Paillier	232.135
TD_P	Decryption in Paillier	2.118
TM_P	Multiplication in Paillier	0.015

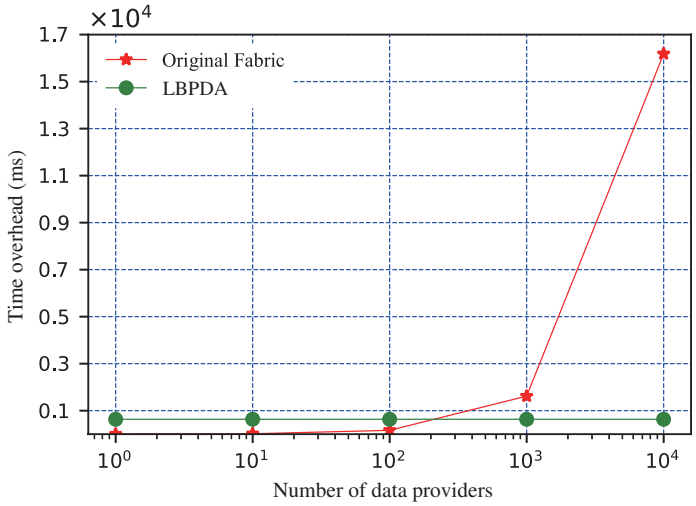


图 6 (网络版彩图) 两种方案签名和验证签名时间开销对比

Figure 6 (Color online) Comparison of signature and verification time overhead between two schemes

6.1 时间开销分析

如表 1 所示, 我们测量了几个主要的密码学操作的详细处理时间. 需要提及的是 ECDSA 和 BLS 中的群阶 p 的大小是 256 位, Paillier 使用的参数 N 大小为 1024 bit. 所有的处理时间是 10000 次重复实验的平均.

我们假设数据提供者以周期性的、准时的方式对数据加密和签名后上报. 原始 Fabric 采用 ECDSA 作为其签名算法, 本文的 LBPDA 方案中的签名方案采用 BLS. 原始 Fabric 和 LBPDA 的签名和验证签名的总体时间开销分别为 $TS_E + \sigma TV_E$ 和 $TS_B + TV_B$. 图 6 描述了 Fabric 改进前后随着数据提供者数量 σ 变化的签名和验证签名总体时间开销变化趋势. 从图 6 可以看出, LBPDA 的时间开销接近于常数, 这主要是因为验证签名的时间开销跟数据提供者的数量无关. 此外, 图 6 显示当数据提供者的数量较小时 (比如 < 110), LBPDA 方案时间开销比原始 Fabric 大, 这主要是因为 BLS 验证单个签名中的双线对操作时间开销较大. 但随着数据提供者数量增大, LBPDA 采用批量签名验证方式的优势更加明显, 其时间开销可以控制在一秒钟内.

表 2 交易中的参数

Table 2 Parameters in the transaction

Parameters	Signature	Payload		
		Header		Data
		ChannelHeader	SignatureHeader	Array of actions
Length (bytes)	71	72	36	56

表 3 区块中的参数

Table 3 Parameters in the block

Parameters	Header			Data	Metadata
	Number	PreviousHash	DataHash	Array of transactions	Metadata
Length (bytes)	4	32	32	$\sigma \times 235$	48

6.2 存储开销分析

我们假设数据提供者上报的 σ 个交易都通过检验且被打包到一个区块. Fabric 中的交易和区块中各部分结构所占存储空间分别如表 2 和 3 所示. 交易包含两个部分: 交易发送者的签名 (signature) 和负载 (payload). 负载包含数据头 (header) 和动作 (action) 数组. 数据头包括通道数据头 (ChannelHeaeder) 和签名数据头 (SignatureHeader), 通道数据头由交易 ID (4 bytes)、时间戳 (4 bytes) 和通道信息 (64 bytes) 组成. 签名数据头包括 mspid (32 bytes) 和一个随机数 (4 bytes). 区块包含 3 部分: 数据头 (header)、数据 (Data)、元数据 (Metadata). 其中数据头包括当前区块的编号 (4 bytes) 和上一个区块和当前区块的哈希值 (32 bytes); 数据包含交易数组; 元数据包含与当前区块相关的元数据 (48 bytes).

我们将传染病统计数据按综合特征, 比如严重急性呼吸道感染 (severe acute respiratory infection, SARI) 和流感样疾病 (influenza-like illness, ILI) 及年龄段 (比如 <2, $2-4$, $5-17$, $18-27$, $28-44$, $45-64$, $65+$) 分组. 因此每个数据提供者生成的交易包含 $2\times 7=14$ 个动作, 每个动作 (action) 以键值对 (4 bytes) 的方式存储数据. 此外, 经过测试, ECDSA 和 BLS 签名大小分别为 $\mathrm{SS_E}=71$ bytes 和 $\mathrm{SS_B}=33$ bytes. 因此原始 Fabric 与提出的 LBPDA 方案中区块的存储开销分别 $116+\sigma\times(\mathrm{SS_E}+72+36+14\times 4)$ 和 $116+\sigma\times(72+36+14\times 4)+\mathrm{SS_B}$.

正如前面所述, 原始 Fabric 的交易中的签名算法采用 ECDSA, 而我们的 LBPDA 中的将签名算法由改成了 BLS 短签名. 一方面利用 BLS 的签名聚合功能, 交易中可以删除签名字段, 只在区块存储一个所有交易的签名聚合. 毕竟我们只需保证区块中聚合签名的正确性, 因为区块一旦添加到区块链, 其无法篡改的特性可以保障区块中的所有交易都是通过验证的. 另一方面因为 BLS 的实际签名大小差不多是 ECDSA 的一半, 可以进一步减少区块的存储开销.

图 7 描述了两个方案随着数据提供者数量变化的区块存储开销变化趋势. 从图 7 可以看出, 本文的 LBPDA 方案具有更小的存储开销, 且数据提供者的数量 σ 越大, 优势越明显, 当 σ 达到 10000 时, 可以减少 30.2% 的区块存储开销.

7 结论

本文设计了一个分布式的、安全的、保护隐私的传染病数据聚合方案. 方案集成了联盟区块链和同态加密技术来保证数据的机密性、完整性、认证和隐私. 区块链提供了一个分布式的信息物理系统

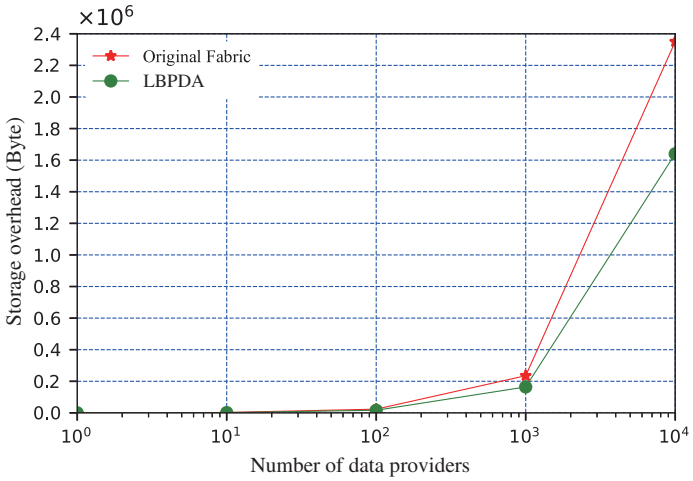


图 7 (网络版彩图) 两种方案区块存储开销对比

Figure 7 (Color online) Comparison of block storage overhead between two schemes

来提高传染病监测数据聚合系统的鲁棒性. Paillier 密码系统加法同态性质允许数据提供者在不透露原始传染病统计数据的情况下追踪各个区域的传染病状况. 在本文提出的方案中, 政府部门能在不破坏任何病人和数据提供者的隐私的同时监测特定区域的传染病数据总数. 安全性分析验证提出的方案如何满足认证、数据完整性、机密性和隐私保护. 仿真实验证明我们方案的改进是有效的, 相比基于原始 Fabric 的隐私保护方案具有更低的存储和迟延开销. 我们现在致力于将区块链、密码学、信息安全等技术引入医疗信息系统, 未来将致力于拓展我们的数据聚合方案, 比如设计出支持包含求和、均值、方差、单因素方差分析等多种统计分析功能的数据聚合函数来提升系统监测能力, 并利用它进行传染病爆发检测与追踪.

参考文献

1 Hu J. Privacy-preserving data integration in public health surveillance. Dissertation for Ph.D. Degree. Université d'Ottawa: University of Ottawa, 2011

2 Bellod Cisneros J L, Aarestrup F M, Lund O. Public health surveillance using decentralized technologies. Blockchain in Healthcare Today, 2018. https://core.ac.uk/download/pdf/189889509.pdf

3 Guan Z T, Si G L, Zhang X S, et al. Privacy-preserving and efficient aggregation based on blockchain for power grid communications in smart communities. IEEE Commun Mag, 2018, 56: 82–88

4 Chattu V K, Nanda A, Chattu S K, et al. The emerging role of blockchain technology applications in routine disease surveillance systems to strengthen global health security. Big Data Cognitive Comput, 2019, 3: 25

5 Paillier P. Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of International Conference on the Theory and Applications of Cryptographic Techniques, 1999. 223–238

6 Dong G S, Chen Y X, Fan J, et al. Research on privacy protection strategies in blockchain application. Comput Sci, 2019, 46: 29–35 [董贵山, 陈宇翔, 范佳, 等. 区块链应用中的隐私保护策略研究. 计算机科学, 2019, 46: 29–35]

7 Boneh D, Lynn B, Shacham H. Short signatures from the Weil pairing. J Cryptology, 2004, 17: 297–319

8 Yuan B, Lin W, McDonnell C. Blockchains and electronic health records. Mcdonnell Mit Edu, 2016

9 Zheng Z B, Xie S A, Dai H N, et al. An overview of blockchain technology: architecture, consensus, and future trends. In: Proceedings of the 6th International Congress on Big Data, 2017, 557–564

10 Androulaki E, Bargeer A, Bortnikov V, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. In: Proceedings of the 13th EuroSys Conference, 2018. 1–15

11 Lu R X, Liang X H, Li X, et al. EPPA: an efficient and privacy-preserving aggregation scheme for secure smart grid

communications. IEEE Trans Parallel Distrib Syst, 2012, 23: 1621–1631
12 He D, Kumar N, Zeadally S, et al. Efficient and privacy-preserving data aggregation scheme for smart grid against internal adversaries. IEEE Trans Smart Grid, 2017, 8: 2411–2419
13 Boneh D, Goh E J, Nissim K. Evaluating 2-DNF formulas on ciphertexts. In: Proceedings of Theory of Cryptography Conference, 2005. 325–341
14 Li S H, Xue K P, Yang Q Y, et al. PPMA: privacy-preserving multisubset data aggregation in smart grid. IEEE Trans Ind Inf, 2018, 14: 462–471
15 Zhang J L, Zhao Y C, Wu J, et al. LVPDA: a lightweight and verifiable privacy-preserving data aggregation scheme for edge-enabled IoT. IEEE Int Things J, 2020, 7: 4016–4027
16 El Emam K, Hu J, Mercer J, et al. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. J Am Med Inf Assoc, 2011, 18: 212–217
17 Xu J, Xue K P, Li S H, et al. Healthchain: a blockchain-based privacy preserving scheme for large-scale health data. IEEE Int Things J, 2019, 6: 8770–8781
18 Wang S, Wang J, Wang X, et al. Blockchain-powered parallel healthcare systems based on the ACP approach. IEEE Trans Comput Soc Syst, 2018, 5: 942–950
19 Dagher G G, Mohler J, Milojkovic M, et al. Ancile: privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. Sustain Cities Soc, 2018, 39: 283–297
20 Li C T, Shih D H, Wang C C, et al. A blockchain based data aggregation and group authentication scheme for electronic medical system. IEEE Access, 2020, 8: 173904
21 Wang Y X, Luo F J, Dong Z Y, et al. Distributed meter data aggregation framework based on Blockchain and homomorphic encryption. IET cyber-phys syst, 2019, 4: 30–37
22 Chen S G, Yang L, Zhao C X, et al. Double-blockchain assisted secure and anonymous data aggregation for fog-enabled smart grid. Engineering, 2020. doi: 10.1016/j.eng.2020.06.018
23 Wang X D, Garg S, Lin H, et al. A secure data aggregation strategy in edge computing and blockchain empowered Internet of Things. IEEE Int Things J, 2020. doi: 10.1109/JIOT.2020.3023588

Lightweight-blockchain based privacy-preserving data aggregation for epidemic disease surveillance

Baiji HU, Yuancheng LI*, Fang FANG & Xingyu SHANG

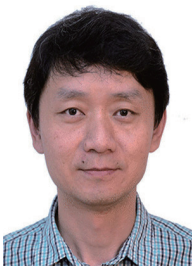
School of Control and Computer Engineering, North China Electric Power University, Beijing 102206*, China*
* Corresponding author. E-mail: ycli@ncepu.edu.cn

Abstract As the COVID-19 pandemic is raging worldwide, epidemic surveillance helps prevent the spread of the epidemic. Protecting the privacy of patients and data providers in the monitoring system can relieve them of their concerns about the leakage of private information, thereby improving the epidemic data collection capabilities of the system. In this article, we propose a lightweight-blockchain based privacy-preserving data aggregation scheme (LBPDA) for epidemic disease surveillance, which can aggregate data without relying on a trusted third party. Especially, to protect data privacy in the data aggregation process, the case count data is aggregated based on the Paillier cryptosystem's homomorphism. Besides, to reduce the time and storage overhead, we improved the adopted Hyperledger Fabric, thus lightening the data aggregation process. Finally, we simulated the proposed scheme and analyzed its security and performance to verify its feasibility and effectiveness. The results showed that the proposed scheme could meet the government department's requirements to aggregate patient case count data for epidemic disease surveillance while protecting the privacy of patients and data providers. Moreover, we also proved that the lightweight of blockchain is effective by comparison.

Keywords blockchain, epidemic disease surveillance, privacy-preserving data aggregation, Paillier cryptosystem, Hyperledger Fabric

Baiji HU was born in 1992. He is a Ph.D. candidate at North China Electric Power University, Beijing, under the supervision of professor Yuancheng LI. His main research interests include blockchain and cryptography.

Yuancheng LI was born in 1970. He received his Ph.D. degree from the University of Science and Technology of China, Hefei, in 2003. Currently, he is a professor and Ph.D. supervisor at North China Electric Power University, Beijing. His main research interests include the information security of power grid, cloud computing, and big data.



Fang FANG was born in 1976. He received his Ph.D. degree from North China Electric Power University, Beijing, in 2005. Currently, he is a professor and Ph.D. supervisor at North China Electric Power University, Beijing. His current research interests include modeling and control of power generation units, optimal configuration and operation of combined cooling, heating and power (CCHP) systems, and forecasting of integrated energy systems.

Xingyu SHANG was born in 1992. He is an M.S. candidate at North China Electric Power University, Beijing. His main research interests include blockchain and information security.