云环境下基于全同态加密的全域匿名化算法

刘君强[1)] 陈芳慧[1)] 徐从富[2)] 郭 宏[3)] 李 挺[1)]

[1)](浙江工商大学信息与电子工程学院 杭州 310018)

[2)](浙江大学计算机科学与技术学院 杭州 310027)

[3)](北京天云融创软件技术有限公司 北京 100193)

摘 要 随着云计算技术的发展和普及,人们越来越多地在云端存储数据和执行计算任务,隐私保护面临很多新挑战,其中之一是在云端对数据进行隐私保护处理的计算过程本身也可能受到攻击,这增加了隐私泄露的风险和规避风险的难度.为此,首次提出了基于全同态加密和全域泛化的匿名化算法,贡献有三点:第一,提出云环境下保护隐私的 k-anonymous 数据共享模型,对云端存储的加密数据直接进行基于全域泛化的匿名化处理,该模型优点是能够支持多种应用、多种保护原则和技术参数;第二,提出适用于全同态加密计算特点的搜索全域泛化最优解的四个优化策略,设计适合于全域泛化的数据加密存储形式,从效率上保障所提出模型的可用性;第三,提出基于全同态加密和全域泛化的匿名化算法,针对云服务提供商可能成为隐私攻击者的情况,设计匿名化算法的云端运行协议.利用分别三代全同态加密方案的开源代码项目,选用隐私保护研究常用的人口统计数据集,针对所提出的四个优化策略和匿名化算法在多种全同态加密方案下进行实验评估,表明所提出模型、策略、算法是有效的.

关键词 隐私保护;全域匿名化;k 匿名原则;全同态加密;信息安全;云计算

中图法分类号 TP309 **DOI 号** 10.11897/SP.J.1016.2019.00837

Full-Domain Anonymization Algorithm Based on Fully Homomorphic Encryption in the Cloud

LIU Jun-Qiang[1)] CHEN Fang-Hui[1)] XU Cong-Fu[2)] GUO Hong[3)] LI Ting[1)]

[1)](*School of Information and Electric Engineering, Zhejiang Gongshang University, Hangzhou* 310018)

[2)](*School of Computing Science and Technology, Zhejiang University, Hangzhou* 310027)

[3)](*Skycloud Software Co., Beijing* 100193)

Abstract With the development of the cloud computing technology and the exponential growth of data volume, people are outsourcing their data storage and computation tasks to cloud service providers, which results in new challenges to privacy preservation. One of such challenges is that the computation process to sanitize data for privacy preservation over the cloud may also suffer from attacks, which incurs additional privacy risks and makes it hard to preserve privacy with the cloud computing paradigm. To address the challenge, this paper proposes a full-domain generalization based anonymization algorithm that is secured by fully homomorphic encryption. While fully homomorphic encryption is deemed as the most promising technology to address the security and privacy issues with the cloud computing paradigm, this work is the first one that integrates the fully homomorphic encryption with data anonymization to address privacy issues over the cloud, to the best of our knowledge. The novelty of this work comes with three contributions.

收稿日期:2016-08-31;在线出版日期:2017-12-25.本课题得到国家自然科学基金(61272306,61672449)和浙江省自然科学基金(LY17F020004)资助.**刘君强**,男,1962 年生,博士,教授,主要研究领域为信息安全与隐私保护、数据挖掘.E-mail:jjliu@alumni.sfu.ca.**陈芳慧**,女,1992 年生,硕士研究生,主要研究领域为数据挖掘与隐私保护.**徐从富**,男,1969 年生,博士,副教授.主要研究领域为人工智能、机器学习、数据挖掘、信息融合.**郭 宏**,男,1964 年生,硕士,高级工程师,主要研究领域为云计算、大数据应用.**李 挺**,女,1995 年生,学士,主要研究领域为密码学与隐私保护.

First, a privacy preserving k-anonymous data sharing model over the cloud is proposed where the data are encrypted and then outsourced to a cloud service provider and full-domain generalization based anonymization is performed on the encrypted data per user request. The advantage of such a model is that it can accommodate user-specified privacy requirements, such as privacy principles and sanitization techniques and parameters, and the encrypted and sanitized data can be used for multiple purposes. Second, a data storage scheme suitable for performing full domain generalization based anonymization on the encrypted data and four optimization strategies for searching anonymization solutions in accordance with the properties of fully homomorphic encryption are proposed respectively, which facilitates the usability of the proposed data sharing model from the perspective of the algorithm performance. Third, an algorithm is proposed for performing full-domain generalization based anonymization on fully homomorphically encrypted data. Moreover, a protocol for running the anonymization algorithm over the cloud is also proposed with the assumption that the cloud service provider can be a privacy attacker. Extensive experiments are performed by using three open source projects that implement the three generations of fully homomorphic encryption schemes, respectively, and by using the well-known "census income" benchmark dataset to evaluate the four optimization strategies and the proposed anonymization algorithm, which demonstrates that the proposed model, algorithm, and protocol are effective.

Keywords privacy preserving; full-domain anonymization; k-anonymity; fully homomorphic encryption; information security; cloud computing

1 引 言

在这个信息普遍共享的时代,信息安全和隐私保护无时无刻不重要.信息安全是要防止未经授权的数据披露,而隐私保护旨在避免所披露数据被用于从中推理个人的敏感信息.

自 Samarati和Sweeney的开创性工作[1]以来,传统计算模式下隐私保护研究在原则、手段、算法等方面取得了很多成果,包括 k-anonymity[1-2]、ℓ-diversity[3]、ε-differential privacy[4]等隐私保护原则,泛化[5-11]、抑制[10-12]、Randomization[4,13]、Anatomy[14]等隐私保护手段.其中,k-anonymity 是最基本的隐私保护原则、能有效防止链接攻击[2,5-6,8-11],不仅适用于关系型数据、也适用于集合型数据[11].基于全域泛化的全域匿名化(full-domain anonymizaiton)是最基本的隐私保护手段,既能过滤隐私信息以防止攻击者进行推理、又能保留数据项的完整语义和独立性.

随着云计算技术的发展和普及,人们越来越多地在云端存储数据和执行各种计算处理任务,隐私保护面临很多新挑战,其中之一是隐私保护处理过程本身也可能受到攻击,这无疑增加了隐私泄露的风险和规避风险的难度.

幸运的是全同态加密技术为解决新挑战提供了可能性.全同态加密是一种可以直接对加密状态下的密文数据进行运算的技术.早在 1978 年 Rivest 等人[15]就提出同态加密的概念,然而此后的相关研究却一直停滞不前.直到 2009 年 Gentry[16]构造出第一个全同态加密方案后,才有了迅猛发展.

已经有研究工作将全同态加密等多种技术应用到云环境下的安全或隐私保护.一是将加密技术运用到云环境下数据挖掘过程中的隐私保护[17-20];二是将现有隐私保护技术应用于云计算[21-24];三是云环境下安全认证过程的隐私保护[25-26].但是,目前还没有解决在云端对数据进行隐私保护处理的计算过程本身受到安全攻击的问题.

本文首次提出基于全同态加密技术和全域泛化匿名化技术实施 k-anonymity 的算法,保护在云端对数据进行隐私处理的过程以防止安全攻击.主要贡献有以下几个方面:

第一,提出云环境下保护隐私的 k-anonymous 数据共享模型.考虑外包数据以加密形式长期存储在云端,在云端直接对加密数据进行基于全域泛化的匿名化处理,提供符合 k-anonymity 的数据共享服务.这个模型的优点是加密存储的数据可以有多种用途,可以采用不同隐私保护原则、技术、参数.

第二，提出适用于全同态加密计算特点的搜索全域泛化最优解的四个优化策略，从算法效率上保障所提出模型的可用性. 设计适合于全域泛化的数据加密存储形式，解决全同态加密状态下无法进行比较运算的问题.

第三，提出基于全同态加密和全域泛化的 k-anonymity匿名化算法. 针对云服务提供商(云端)是隐私攻击者的假设，设计匿名化算法云端运行协议. 真实数据集上的实验评估表明所提出的模型、策略、算法是有效的.

本文第 2 节综述相关工作；第 3 节提出云环境下保护隐私的 k-anonymous 数据共享模型；第 4 节提出适合于全同态加密的最优匿名解搜索策略；第 5 节提出云环境下基于全同态加密的全域匿名化算法和协议；第 6 节实验评价所提出模型、策略和算法；第 7 节总结全文.

2　相关工作

与本文直接相关的工作可归纳为三个方面.

传统计算模式下隐私保护研究：隐私是个人掌控自己信息被披露方式、范围、时间的权利要求. 隐私保护原则是人们对隐私权利要求及其保护水平的表述，主要有 k-anonymity[1-2]、ℓ-diversity[3]、ε-differential privacy[4]、$\ell(+)$-diversity[7]、privacy template[12]、ρ_1-to-ρ_2 privacy[13]、t-closeness[27]、m-invariance[28]等隐私保护原则. 实现隐私保护的有效手段则是数据匿名化，主要有全域泛化[1,5]、全子树泛化[6-7]、多维泛化[8-9]、抑制[10-12]、Randomization[4,13]、anatomy[14]、slicing[29]、Disassociation[30]等. 隐私保护算法通过采用特定的匿名化方法，对给定数据进行变换，使之符合所要求的隐私保护原则，有最优算法，如 Incognitio[5]、$\ell(+)$-Optimize[7]以及启发式算法，如 TDS[6]、Mondrian[8]、MinGen[10]等. 本文提出一个最优算法、运用全域泛化的匿名化方法、实施 k-anonymity.

全同态加密技术：Rivest 等人[15]在 1978 年就提出同态加密的概念，公钥密码算法 RSA[31]是乘法同态加密方案，Paillier 算法[32]是加法同态加密方案，但直到 2009 年 Gentry[16]构造出第一个全同态加密方案后，才迎来研究热潮. 目前的全同态加密方案大致划分为三代：第一代包括基于理想格的全同态加密方案[16]、Dijk 等人[33]基于整数的 vDGHV 方案、Smart 和 Vercauteren[34]改进的基于整数的 SV 方案，第二代包括 Brakerski 等人[35]基于 GLWE (Generalized LWE)的 BGV 方案、基于 RLWE(Ring LWE)的 BV 方案[36]，第三代包括 Gentry 等人[37]基于 LWE (Learning With Error)的 GSW 方案、Ducas等人[38]的 DM 方案、Chillotti 等人[39]的 CGG 方案.

云环境下安全与隐私研究：关于可信访问控制、虚拟安全技术、云资源控制的研究工作很多，但与本文相关的较少，大致有三类. 第一类是将同态加密或其它加密方法运用到云环境下的数据挖掘中：Hu 等人[17]提出基于同态加密的 k 近邻查询的云计算架构. Liu 等人[18]用全同态加密保护云端的数据挖掘过程. Puttaswamy 等人[19]提出自动识别工具在不影响应用程序功能情况下对进行数据加密. Zhang 等人[20]以隐私泄露上限为约束条件，通过加密部分中间数据来保护多维数据集的隐私；第二类是将传统隐私保护技术应用于云计算环境：Wang 等人[40]提出对明文数据进行隐私保护处理的新方法，与之不同的是本文直接对密文数据进行隐私保护处理. Zhang 等人[21]提出基于 MapReduce 模型在云端进行数据匿名化处理. Blass 等人[22]提出基于分布式编程模型 MapReduce 的保护隐私的搜索. Gao 等人[23]变换云端外包的图形数据，在防止近邻攻击的同时能准确回答最短路径查询. Roy 等人[24]将 differential privacy 融入“云”中的数据生成与计算阶段，防止 map reduce 计算过程中隐私泄露. 第三类是研究云环境下数据完整性验证过程的隐私保护：Wang 等人[25]研究由第三方公共审计 TPA 验证云端存储数据的完整性时，通过采用双线性合成签名方法避免 TPA 带来隐私风险. Wang 等人[26]研究通过多人环签名来验证共享数据完整性时，如何针对第三方公共审计 TPA 保护每个签名者的身份隐私. 这些工作重点在于签名与数据验证技术，与之不同的是本文重点在于运用全同态加密技术.

简言之，本文将全同态加密与保护隐私的数据共享相结合，关键是如何将全域匿名化处理表达成可以在加密状态下进行同态计算的任务，目前尚没有这方面成果.

3　云环境下保护隐私数据共享模型

考虑数据拥有者购买云存储服务存放数据、购买云计算服务进行数据隐私保护处理的场景，由于云端计算过程本身会受到安全与隐私攻击，提出基于全同态加密在云端保护隐私的 k-anonymous 数

据共享模型，如图 1 所示，模型包括两个部分：第一，数据拥有者将加密后的数据上传和长期存储在云端，提供符合 k-anonymity 隐私保护原则的数据共享服务；第二，当用户请求数据时，在云端直接对加密数据进行基于全域泛化的匿名化处理，输出符合 k-anonymity 的数据.

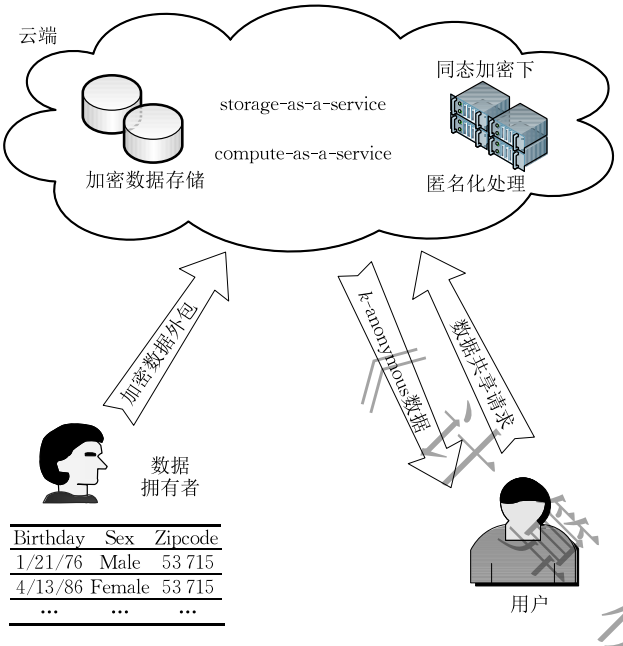


图 1 云环境下保护隐私的 k-anonymous 数据共享模型

此模型有两个优点：一是加密存储的数据可以有多种用途，比如用于各种数据挖掘任务和各种数据共享查询任务；二是易于实现多样的隐私保护，即可以采用不同的隐私参数或隐私保护原则. 此模型所有数据处理任务在云服务端完成，没有增加服务端与客户端之间的通信量.

所提出的云端隐私保护模型基于两个前提：一是云服务提供商（云端）可能是隐私攻击者，但是一定按照协议提供服务；二是用户承诺不将所获得数据表转发给包括云服务提供商在内的未经授权的他人，因此不存在合谋攻击.

3.1 隐私保护原则

数据拥有者希望对外发布和共享的数据符合 k-anonymity 隐私保护原则. 具体讲，数据是以关系表形式发布的，其中某些属性组成的一个子集称为准标识符，可能被攻击者用来进行链接攻击，即通过准标识符与其他来源的数据进行链接操作，从而造成相关人员的隐私泄露. Samarati 等人 1998 年提出 k-anonymity 原则、要求数据表每条记录与至少 $k-1$ 条其它记录在准标识符上无法区分，以有效防止链接攻击.

定义 1. 属性（attribute）. 描述一类对象的某种特征. 属性值（attribute value）则是某个对象的属性的具体描述. 属性的取值范围称为值域（attribute domain），可以表达为一个集合.

定义 2. 标识符（identifier）. 能够直接标识个人身份的属性，如学号、身份证号、银行卡号. 数据拥有者向外发布数据时，会删除或隐藏标识符.

定义 3. 准标识符（Quasi-IDentifier，QID）. 多个属性的集合，其中单个属性不能够唯一标识个人身份，但是 QID 属性联合起来就能唯一标识个人身份，如生日、地域、性别的组合.

假设准标识符 $QID=\{Q_1,\cdots,Q_n\}$，由 n 个属性组成，属性 Q_i 的值域为 $Q_i^{[0]}$，$i=1,\cdots,n$.

定义 4. QID 等价类（equivalence class）. QID 各属性 Q_i 取值的一种组合. 等价类集合，记作 EC，是 $Q_1^{[0]}\times\cdots\times Q_n^{[0]}$ 的子集. 若一条数据记录 r 的 QID 属性值与某个等价类 ec 完全相同，那么称 r 隶属于或者支持等价类 ec. 隶属于 ec 的记录条数称为等价类 ec 的支持度（support）.

定义 5. k-anonymity 隐私保护原则：要求每个 QID 等价类的支持度或者为 0、或者不低于 k. 符合此原则的数据表称为是 k-anonymous.

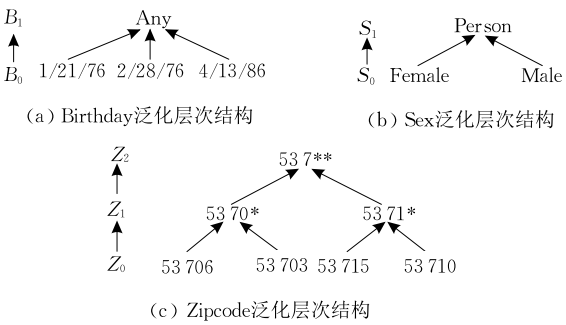
3.2 匿名化技术

实施隐私保护的主要技术是数据匿名化，通过对原始数据的隐私信息进行某种变换，使得攻击者无法推测出具体个体，从而实现个人隐私的保护.

泛化是最为常见的匿名化方法，基本思想是用范围更广的模糊的值来代替原有属性值，从而保护数据中包含的隐私信息. 全域泛化是最重要方法.

定义 6. QID 属性的泛化层次结构（generalization hierarchy）. 一个树形结构，其叶结点组成属性的值域，其余各结点是对值域的不同程度的泛化，组成泛化的值域（generalized domain）.

例 1. 图 2 所示是出生日期 Birthday、性别 Sex、邮编 Zipcode 三个属性的泛化层次结构.



(a) Birthday泛化层次结构 (b) Sex泛化层次结构

(c) Zipcode泛化层次结构

图 2 三个属性泛化层次结构

定义 7.　全域泛化(full-domain generalization).泛化处理前后同一属性的取值处于相同泛化层次,因此泛化层次结构的各叶结点与根结点等距离.可以用泛化层与叶结点层的距离值来表示该属性的泛化程度.这也称为全域匿名化(full-domain anonymization).

给定准标识符 $QID=\{Q_1,Q_2,\cdots,Q_n\}$,属性 Q_i 泛化层次结构高度为 h_i,原始值域为 $Q_i^{[0]}$,第 d_i 层泛化值域为 $Q_i^{[d_i]}$,$0\leq d_i\leq h_i$,$i=1,\cdots,n$.泛化等价类集合 $EC\subseteq Q_1^{[d_1]}\times Q_2^{[d_2]}\times\cdots\times Q_n^{[d_n]}$,某个等价类 $ec=\{v_1^{[d_1]},v_2^{[d_2]},\cdots,v_n^{[d_n]}\}$,$v_i^{[d_i]}=f_i(v_i^{[0]},d_i)\in Q_i^{[d_i]}$,$f_i$ 函数取决于 Q_i 的泛化层次结构.

定义 8.　数据表(Record Table,RT).数据记录的集合.未经泛化处理的记录组成的表称为原始数据表(Original RT,ORT).ORT 经过泛化处理转换成泛化数据表(Generalized RT,GRT).

原始数据表的结构是 $ORT[Q_1^{[0]},\cdots,Q_n^{[0]}]$,泛化数据表的结构则是 $GRT[Q_1^{[d_1]},\cdots,Q_n^{[d_n]}]$,其中 $d_1,\cdots,d_n$ 是各个属性的泛化层次.

例 2.　表 1 是原始数据表,有 Birthday、Sex 和 Zipcode 三个 QID 属性,显然不符合 2-anonymity.如果将三个属性分别泛化至第 1、0、1 层,那么有四个等价类:{Any,Female,5370 * }、{Any,Female,5371 * }、{Any,Male,5370 * }和{Any,Male,5371 * },支持度分别为 2、1、2 和 1,仍然不符合 2-anonymity.如果将三个属性分别泛化至第 1、0、2 层,则有二个等价类{Any,Female,537 * * }和{Any,Male,537 * * },支持度分别为 3 和 3,表 2 所示是相应的泛化数据表,符合 2-anonymity.

表 1　原始数据表 ORT

Birthday 属性	Sex 属性	Zipcode 属性
1/21/76	Male	53715
4/13/86	Female	53715
2/28/76	Male	53703
1/21/76	Male	53703
4/13/86	Female	53706
2/28/76	Female	53706

表 2　泛化数据表 GRT

Birthday 属性	Sex 属性	Zipcode 属性
Any	Male	537 * *
Any	Female	537 * *
Any	Male	537 * *
Any	Male	537 * *
Any	Female	537 * *
Any	Female	537 * *

3.3　隐私保护算法

隐私保护算法根据给定数据表、隐私保护原则、隐私参数、匿名化技术,找到一个问题解,即如何通过匿名化使数据表转换成符合隐私原则的形式.当采用全域泛化方法时,问题解可以表达为所谓的泛化格点,算法执行过程则是搜索所谓泛化格的过程.

定义 9.　泛化格(generalized lattice).所有 QID 属性的泛化程度可表为一个距离向量 $(d_1,\cdots,d_n)$,其中 d_i 是 Q_i 的泛化层次,$0\leq d_i<h_i$,$i=1,\cdots,n$.通过对某个分量的加 1 运算,从一个距离向量可衍生出另一个距离向量.根据这种衍生关系,所有距离向量组成一个层次结构,称为泛化格,距离向量则称为格点.由格点 *node* 衍生而来的格点称为 *node* 的子女格点.

泛化格具有层次性,用 *lattice*(0)表示 0 层,只包含根格点,其距离向量的各个分量均为 0,即所有属性都没有泛化.一般地,*lattice*(i)表示 i 层($i>0$)的格点集合,它由 *lattice*($i-1$)衍生得到.

例 3.　图 3 所示是由图 2 的 3 个属性 Birthday、Sex、Zipcode 组成 QID 的泛化格.*lattice*(0)只有唯一的根格点(0,0,0),其子女格点(1,0,0)、(0,1,0)、(0,0,1)组成 *lattice*(1).

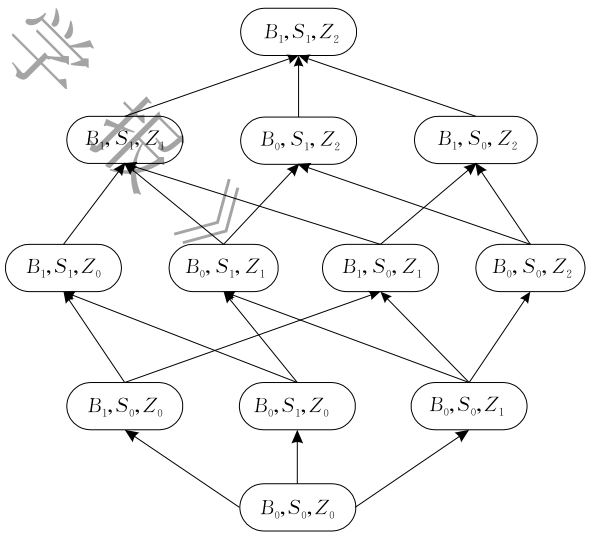


图 3　三个 QID 属性构建的泛化格

泛化格中符合 k-anonymity 原则的格点称为可行解,精度最高的可行解称为最优解.一般来讲,隐私保护算法并不依赖于具体的精度标准.这里介绍通用精度标准 Precision metric(简称 Prec)[10],它根据泛化层次的高度进行度量.计算公式是 $Prec(GRT)=1-\dfrac{\sum_{i=1}^{n}\frac{d_i}{h_i}}{n}$,其中 ORT 是原始数据表,GRT 是泛化

数据表，$N=|ORT|=|GRT|$ 是数据表的记录条数. n 为 ORT 表中属性个数. h_i 是属性 Q_i 的泛化层次，d_i 表示 Q_i 属性的当前泛化高度.

例 4. 根据图 2 知，图 3 格点(1,0,2)的精度为 $Prec=1-\frac{1/2+0/2+2/3}{3}=0.611$.

算法的目标是寻找符合 k-anonymity 原则的最优泛化格点，以下性质有助于剪裁泛化格(搜索空间)、提高算法效率.

性质 1. 泛化性质. 如果泛化格点 $node$ 符合 k-anonymity，那么 $node$ 的所有子女格点都符合 k-anonymity. 同时，$node$ 的精度不低于子女格点的精度.

性质 2. 子集性质. 若 QID 符合 k-anonymity，那么 QID 属性子集也符合 k-anonymity；若 QID 属性子集不符合 k-anonymity，那么其超集也不符合.

3.4 运用全同态加密保护隐私处理过程

所提出的云端保护隐私的 k-anonymous 数据共享模型(图 1 所示)采用全同态加密技术. 全同态加密是一种加密方案，允许在密文上进行计算，直接得到密文形式的结果，该结果解密后等于相应的明文计算结果. 通常，我们用布尔电路来表达计算.

定义 10. 全同态加密方案(fully homomorphic encryption scheme). 由 KeyGen、Encrypt、Decrypt、Evaluate 四个基本函数和 Recrypt 辅助函数定义.

KeyGen(λ)密钥函数，生成密钥对(pk,sk)，其中 pk 是公钥、sk 是私钥.

Encrypt($pk,m_1,\cdots,m_s$)加密函数，用公钥 pk 加密明文二进位 $m_i\in\{0,1\}$ 得到密文 c_i，$i=1,\cdots,s$.

Decrypt($sk,c_1,\cdots,c_t$)解密函数，用私钥 sk 对密文 c_i 进行解密，还原出明文 $m_i\in\{0,1\}$，$i=1,\cdots,t$.

Evaluate($pk,f,c_1,\cdots,c_u$)同态计算函数，根据公钥 pk 和明文计算函数 f，对密文输入 $c_1,\cdots,c_u$ 做同态计算，使得 Decrypt(sk,c_f)$=f(m_1,\cdots,m_u)$，其中 c_f=Evaluate($pk,f,c_1,\cdots,c_u$)，m_i=Decrypt(sk,c_i)，$i=1,\cdots,u$.

Recrypt(pk_2,Encrypt(pk_2,sk_1),Encrypt($pk_1,m_1,\cdots,m_v$))重加密函数，将 pk_1 加密的密文转换成用 pk_2 加密的密文，其中(pk_1,sk_1)和(pk_2,sk_2)是两对密钥，$m_i\in\{0,1\}$，$i=1,\cdots,v$，重加密的实质就是同态解密.

全同态加密方案有两个特点：一是同态计算需要给定由与门和非门组成的二进制计算电路(程序)；二是出于安全性的要求，密文添加随机噪音，无法通过比较两个密文的大小关系来确定相应明文的大小关系.

上述两个特点直接影响到所提出云环境下保护隐私的 k-anonymous 数据共享模型的具体设计.

4 基于全同态加密的全域泛化策略

根据寻找符合 k-anonymity 的全域泛化最优解的目标，提出符合全同态加密两个特点的泛化格搜索策略、加密数据外包形式、基于全同态加密的支持度的计算方法.

4.1 泛化格搜索优化策略

寻找最优解的基本思路是从泛化格的根格点出发，按宽度优先次序搜索泛化格，即先 $lattice(0)$、再 $lattice(1)$、依次类推. 对搜索到的每个格点进行 k-anonymity 原则判定. 如果相应的泛化数据表中有某个等价类的支持度大于 0 并小于 k，则格点不符合 k-anonymity，否则为符合. 从符合 k-anonymity 的泛化格点中选取精度最高的格点.

基本策略是剪裁搜索空间：当访问某一层格点时，按精度从高到低次序进行，如果泛化格点 $node$ 符合 k-anonymity，那么泛化格的性质 1，同层余下的格点不必再访问，并且 $node$ 的所有子女格点可以从下一层格点中排除.

本节提出针对全同态加密下搜索泛化格的四个优化策略，进一步剪裁搜索空间、降低判定 k-anonymity原则的计算量.

策略 1. 等价类集合自动选取策略(Strategy ECType). 等价类集合选取方式有：(1)将数据表每条记录作为等价类，即等价类集合 $EC\leftarrow RT$，记为 ECType_1；(2)选取各 QID 属性泛化值域的笛卡尔积为等价类集合，即 $EC\leftarrow\times_i Q_i^{[d_i]}$，记为 ECType_2；(3)自动选取策略是如果 RT 的记录条数小于笛卡尔积的尺寸，选取 ECType_1，否则 ECType_2.

要强调两点：首先，ECType_1 的等价类可能重复出现，但是能自动剔除支持度为 0 的；ECType_2 的等价类不重复，但支持度可能为 0. 第二，加密状态下为判定格点是否符合 k-anonymity 而计算每个等价类支持度时，需要访问数据表中的所有记录.

本策略旨在减少需要计算支持度的等价类条数，减少数据表访问和判定 k-anonymity 的计算量.

例 5. 考虑 2-anonymity、表 1 数据表、图 2 泛化层次结构、图 3 泛化格. 当访问根格点(0,0,0)时，如果按 ECType_2 生成等价类有 24 条，经计算，前

5 个等价类支持度均为 0，第 6 个为 1. 此时判定不符合 2-anonymity，已扫描数据表 6 遍，共 36 条记录. 如果按 ECType_1，数据表直接选作等价类集合，共 6 条，第一条等价类的支持度为 1，即判定不符合 2-anonymity，此时只遍历了 1 遍数据表，共 6 条记录. 采用本策略，共访问数据表 14 遍(84 条记录)，即可找到最优 2-anonymity 格点(1,1,0)，不采用本策略则要访问数据表 24 遍(144 条记录).

策略 2. 等价类支持度计算策略(Strategy recSum). 给定泛化格点，令 N 是(泛化)数据表记录条数，可以分两种情况来优化等价类支持度的计算过程. (1) 在计算 ECType_2 等价类时，令 $recSum$ 是已经计算的等价类的支持度之和. 如果 $recSum=N$，则剩余等价类的支持度必定均为 0. 若 $N\text{-}recSum<k$，则剩余的等价类中必有支持度大于 0 并小于 k 的. 前一种情况下，格点已符合 k-anonymity，后一种则不符合，因此不需要再做统计；(2) 在计算 ECType_1 等价类时，由于密文状态下无法分辨等价类是否相同，无法剔除重复的等价类. 相应的 $recSum$ 计算如下：支持度互不相等者一定是不同等价类，其支持度直接计入 $recSum$；相等的支持度合并计入，即如果支持度 $supp$ 出现 $times$ 次，则按 $\lceil times/supp\rceil\times supp$ 计入 $recSum$.

例 6. 接续例 5. 当访问泛化格(1,1,0)时，如果采用 ECType_2，共包含 4 个等价类，前 3 个等价类的支持度均为 2，和数为 6(数据表记录数)，无需统计第 4 个等价类；如果采用 ECType_1，那么 6 条记录直接选为等价类，由于在加密状态下，无法知道(不同)等价类实际上只有 3 个. 但是计算前 5 个等价类得到的支持度均为 2，得 $recSum=\lceil 5/2\rceil\times 2=6$，因此无需统计第 6 个等价类.

策略 3. 改进型子集策略(Strategy subSet). 泛化格点 $node$ 是表示所有 QID 属性泛化程度的距离向量 $(d_1,\cdots,d_n)$，$node$ 的子向量通过选取某些分量形成，可用指示向量 $\boldsymbol{vec}=(x_1,\cdots,x_n)$ 表征，$\boldsymbol{vec}[i]=x_i=1$ 选取第 i 分量、0 则不选取. 根据泛化格的性质 2，应该按照指示向量的非 0 分量个数的递增序逐一判定 $node$ 共 2^n-1 个子向量如下.

(1) 如果 $node$ 的某个子向量不符合 k-anonymity，该格点 $node$ 也不符合，剩下的子向量和 $node$ 都无需再判定；(2) 如果 $node$ 的某个子向量不符合 k-anonymity，下一步只需要考虑由该子向量相应分量加 1 所衍生的子女格点，其余子女格点无需考虑(必定不符合)；(3) 不同泛化格点可能有共同子向量，通过记录符合 k-anonymity 的格点子向量，即记录格点以及表征子向量的指示向量 $\boldsymbol{vec}$，可以避免重复计算.

例 7. 接续例 5. 访问根格点(0,0,0)时，首先考虑单个属性子向量，用指示向量(0,0,1)、(0,1,0)、(1,0,0)表征，计算表明各个子向量均符合 2-anonymity，记录根格点和三个指示向量. 然后，考虑两个属性子向量，用指示向量(0,1,1)、(1,0,1)、(1,1,0)表征. 指示向量(0,1,1)表征的子向量不符合 2-anonymity. 因此，根格点也不符合，下一步需要判定的是根格点第 2 和第 3 分量分别加 1 得到的子女格点(0,1,0)、(0,0,1)，无需考虑子女格点(1,0,0). 当访问格点(0,1,0)时，指示向量为(0,0,1)对应子向量不必重复计算，因为已经记录该子向量符合 2-anonymity. 不采用策略需遍历 8 个格点，采用本策略只遍历 6 个格点，并且各个格点计算量更低.

请注意，策略 1 和策略 2 也适用于格点子向量相应的子等价类集合选取和子等价类支持度计算.

策略 4. 零子等价类标记策略(Strategy markEC0). (子)等价类是泛化格点(子向量)各个泛化属性取值的一种组合. 如果子等价类的支持度为 0，那么包含它的(子)等价类的支持度也为 0. 同时，不同泛化格点(子向量)之间可能有公共子向量，一个泛化格点的子等价类可能被另一个泛化格点的(子)等价类所包含. 因此，在采用改进型子集策略时，通过标记支持度为 0 的子等价类，可以大大减少支持度为 0 的 ECType_2 等价类的计算量.

例 8. 接续例 5. 判定格点(0,0,0)指示向量(0,0,1)对应子向量时，发现子等价类{53 710}的支持度为 0，应而标记该子等价类. 判定格点(0,1,0)指示向量为(0,1,1)对应子向量时，发现子等价类{Person,53 710}是{53 710}的超集，无需计算即可判定该子等价类的支持度为 0.

4.2　加密数据外包形式

搜索泛化格的过程中，需要快速地确定每个泛化格点对应的泛化数据表. 为此，根据全同态加密的两个特点，提出将数据表转化为扩展数据表、再编码、加密、然后外包的形式.

定义 10. 扩展数据表 exRT (extended RT). 给定原始数据表 $ORT[Q_1^{[0]},\cdots,Q_n^{[0]}]$，将每个 QID 属性扩展为多列、每列对应属性的一个泛化值域，形成扩展数据表 $\text{exRT}[Q_1^{[0]},\cdots,Q_1^{[h_1-1]},\cdots,Q_n^{[0]},\cdots,Q_n^{[h_n-1]}]$. 对于原始数据表每条记录 $\{v_1^{[0]},\cdots,v_n^{[0]}\}$，将各属性值映射到各泛化层次：$v_i^{[j]}=f_i(v_i^{[0]},j)$，

$j=1,\cdots,h_i-1,i=1,\cdots,n$，得扩展记录$\{v_1^{[0]},\cdots,v_1^{[h_1-1]},\cdots,v_n^{[0]},\cdots,v_n^{[h_n-1]}\}$.

表 3 扩展数据表 exRT

Birthday 属性		Sex 属性			Zipcode 属性	
B_0	B_1	S_0	S_1	Z_0	Z_1	Z_2
1/21/76	Any	Male	Person	53715	5371*	537**
4/13/86	Any	Female	Person	53715	5371*	537**
2/28/76	Any	Male	Person	53703	5370*	537**
1/21/76	Any	Male	Person	53703	5370*	537**
4/13/86	Any	Female	Person	53706	5370*	537**
2/28/76	Any	Female	Person	53706	5370*	537**

例 9. 表 3 所示是表 1 的数据表的扩展形式.

定义 11. 二进制扩展数据表 bRT(binary RT). 为各个 QID 属性每个泛化层的值域单独指定二进制编码，将扩展数据表每个单元取值替换为对应编码，得到二进制扩展数据表 bRT. 二进制编码可节约存储空间、便于进行加密、提高计算效率.

定义 12. 加密的扩展数据表 e^2RT(encrypted extended RT). 为二进制扩展数据表 bRT 每列$Q_i^{[j]}$，$i=1,\cdots,n,j=0,\cdots,h_i-1$，单独生成密钥对$(pk_{i,j},sk_{i,j})$，然后用公钥 $pk_{i,j}$ 对该列加密，得到加密的扩展数据表 e^2RT.

此外，生成密钥对(pk_{comp},sk_{comp})，用公钥 pk_{comp} 加密 bRT 的所有列，得到加密的扩展数据表的副本，记作e^2RT_{comp}，也可以通过重加密e^2RT得到e^2RT_{comp}. 考虑到全同态加密计算的特点和安全性的要求，全同态加密下匿名化处理在e^2RT_{comp}上进行，数据共享发布则通过部分解密 e^2RT 实现.

因此，外包到云端存储的是加密的扩展数据表 e^2RT 及其副本e^2RT_{comp}、QID 泛化层次结构及其编码规则. 实际应用中，e^2RT 和e^2RT_{comp}是逐步积累的，不一定是整批上传云端的.

4.3 加密数据全同态计算等价类支持度

需要判定泛化格点 *node* 的某个由指示向量 ***vec*** 表征的子向量是否符合 *k*-anonymity 的基础工作是在加密的扩展数据表中全同态计算各个(子)等价类 *ec* 的支持度，相应的伪代码如过程 1 所示，其中加密投影子表 *T* 根据泛化格点 $node=(d_1,\cdots,d_n)$和指示向量 $\boldsymbol{vec}=(x_1,x_2,\cdots,x_n)$选取$e^2RT_{comp}$相关列而来.

过程 1. *countingFHE*.

输入：(子)等价类 *ec*，加密投影子表 *T*.

输出：(子)等价类 *ec* 的支持度的密文 *supp*.

1. *supp*←0 in $\lceil\log_2|T|\rceil+1$ cipherbits;
2. FOR each record *t* in *T* DO
3. *IsEqual*←1 in one cipherbit;
4. FOR $i=1$ to sizeof(*t*) DO
5. *cmp*←*ec*[*i*]+*t*[*i*]+1;
6. *IsEqual*←*IsEqual*×*comp*;
7. END FOR
8. *supp*←*supp*+*IsEqual*;
9. END FOR
10. return *supp*;

过程 1 首先初始化$\lceil\log_2|T|\rceil+1$密文位的密文计数器 *supp*(第 1 行)，然后将(子)等价类 ***ec*** 与 *T* 的每条记录 *t* 作比对(第 2～9 行). 其中，第 4～7 行逐位比对 ***ec*** 与 *t*，如果对应位相等则 *cmp* 为 1 的密文(第 5 行)，如果 ***ec*** 与 *t* 相等则 *IsEqual* 为 1 的密文(第 6 行)，累加 *IsEqual* 得到 ***ec*** 的支持度密文(第 8 行).

注意到第 1、3、5、6、8 行是加密数据的全同态计算. 过程 1 的时间复杂度取决于$time_1\times|\boldsymbol{ec}|\times|T|$和$time_2\times|T|$，其中$time_1$是比对单个密文位所需时间(第 5～6 行)，$time_2$是一次密文计数器加 1 操作所需时间(第 8 行).

5 云端同态加密的匿名化算法和协议

5.1 基于全同态加密全域泛化算法

提出基于全同态加密和全域泛化的匿名化算法(fully homomorphic encryption based full-domain generalization for anonymization)，针对云端保护隐私的 *k*-anonymous 数据共享模型(第 3 节)，根据泛化格搜索优化策略、加密数据形式和全同态计算方法(第 4 节)，找出既符合 *k*-anonymity 又有较高精度的全域泛化解.

算法 1 所示是基于全同态加密和全域泛化的匿名化算法的伪代码，其输入是加密的扩展数据表e^2RT_{comp}，QID 属性$(Q_1,\cdots,Q_n)$泛化层次结构，*k*-anonymity 隐私参数 *k*(密文形式).

本算法从根格点开始自底向上逐层搜索泛化格，*Nodeset* 是当前层拟搜索的格点集合，*mxNode* 是已搜索格点中符合 *k*-anonymity 且精度最高者，*goodSubNodes* 按照策略 3(3)记录符合 *k*-anonymity 的格点子向量(第 1 行). *Nodeset* 非空时，搜索泛化格当前层(第 2～20 行). 其中，*succNodes* 和 *failNodes* 分别记录当前层符合 *k*-anonymity 的格点和不符合的格点(第 3 行).

本算法按精度降序逐一处理 *Nodeset* 中每个格点 *node*(第 4～15 行). 根据策略 3，调用函数 *guideVector*(*i*)按非 0 分量个数的递增序逐一生成表征各个子向量的指示向量 ***vec***，判定相应格点子向量是否符合 *k*-anonymity(第 5～14 行). 如果(*node*，***vec***)对应

的格点子向量 ***sbNode*** 是已记录的符合 k-anonymity 的格点子向量，则根据策略 3(3)开始下一个格点子向量(第 6～7 行). 如果格点子向量不符合 k-anonymtiy，则格点 *node* 以及指示向量 ***vec*** 一并记入 *failnodes*，并开始下一个格点(第 8～10 行). 如果 ***vec*** 的分量全 1，则 *node* 记入 *succNodes*，否则根据策略 3(3)将格点子向量 ***sbNode*** 记入 *goodSubNodes* (第 11～13 行).

本算法根据当前层泛化格点搜索结果，准备下一层拟搜索格点(第 16～19 行). 其中，$children(y)$ 生成 y 的所有子女格点，$children(y,\boldsymbol{vec})$ 为 ***vec*** 所有非 0 分量对应的 y 分量加 1 生成其子女格点.

算法 1. 全同态加密全域匿名化算法 *fheFDA*.

输入：加密数据 e^2RT_{comp}，QID 泛化层次结构，参数 k.

输出：符合 k-anonymity 的最高精度泛化格点 *mxNode*.

1. $Nodeset\leftarrow\{(0,\cdots,0)\}$; $mxNode\leftarrow\varnothing$; $goodSubNodes\leftarrow\varnothing$;
2. WHILE $Nodeset\neq\varnothing$ DO
3. $succNodes\leftarrow\varnothing$; $failNodes\leftarrow\varnothing$;
4. FOR each $node=(d_1,\cdots,d_n)\in Nodeset$ DO
5. FOR $\boldsymbol{vec}=guideVector(i)$ with $i=1$ to 2^n-1 DO
6. IF $\boldsymbol{sbNode}=(node,\boldsymbol{vec})\in goodSubNodes$
7. THEN continue; ENDIF
8. IF !$isKanonymous(\boldsymbol{sbNode}, e^2RT_{comp}, k)$ THEN
9. $failNodes\leftarrow failNodes\cup\{\boldsymbol{sbNode}\}$; break;
10. END IF
11. IF($\forall i$ $\boldsymbol{vec}[i]=1$) $succNodes\leftarrow succNodes\cup\{node\}$;
12. ELSE $goodSubNodes\leftarrow goodSubNodes\cup\{\boldsymbol{sbNode}\}$;
13. END IF
14. END FOR
15. END FOR
16. $preNodes\leftarrow\{x\in children(y)\mid y\in succNodes\}$;
17. $newNodes\leftarrow\{x\in children(y,\boldsymbol{vec})\ x\notin preNodes,(y,\boldsymbol{vec})\in failNodes\}$;
18. $mxNode\leftarrow\mathbf{argmax}_x Prec(x)\mid x\in succNodes\cup\{mxNode\}$;
19. $Nodeset\leftarrow\{x\in newNodes\mid Prec(x)>Prec(mxNode)\}$;
20. END WHILE
21. return *mxNode*;

过程 2 所示是过程 *isKanonymous* 的伪代码，被算法 1(第 8 行)调用来判定格点子向量 ***sbNode*** 是否符足 k-anonymity，其输入是格点子向量 $\boldsymbol{sbNode}=(node,\boldsymbol{vec})$，加密的扩展数据表 e^2RT_{comp}，隐私参数 k(密文形式).

过程 *isKanonymous* 根据格点 $node=(d_1,\cdots,d_n)$ 及其子向量的指示向量 $\boldsymbol{vec}=(x_1,\cdots,x_n)$ 选取 e^2RT_{comp} 的相关(扩展)属性列从而得到加密的泛化数据表的投影 T，简称加密投影子表(第 1 行)，初始化$\lceil\log_2|T|\rceil+1$ 密文位的密文计数器 *recSum*(第 2 行)，根据策略 1(Strategy ECType)，等价类集合 *EC* 或者直接由数据表生成(第 3 行)，或者是指示向量分量非 0 即 $\boldsymbol{vec}[i]=1$(扩展)属性值域的笛卡尔积(第 4 行). 计算每条(子)等价类 *ec* 支持度来判定 $\boldsymbol{sbNode}=(node,\boldsymbol{vec})$ 是否符合 k-anonymity(第 5～14 行).

过程 *isKanonymous* 通过静态集合变量 $setEC_0$ 记录所有支持度为 0 的子等价类，包括相应格点、子向量的指示向量、子等价类取值，实现策略 4 Strategy markEC0(第 6，10 行).

过程 *isKanonymous* 调用 *countingFHE* 在加密数据中全同态计算一个(子)等价类的支持度的密文(第 7 行). 通过累计已经计算的(子)等价类的支持度，实现策略 2 Strategy recSum(第 8，11～12 行)，其中第 8 行是全同态加密下的累计.

过程 2. *isKanonymous*.

输入：$\boldsymbol{sbNode}=(node,\boldsymbol{vec})$，数据表 e^2RT_{comp}，参数 k.

输出：true 如果 ***sbNode*** 符合 k-anonymity，否则 false.

1. $T\leftarrow$projection of e^2RT_{comp} according to $(node,\boldsymbol{vec})$;
2. $recSum\leftarrow 0$ in $\lceil\log_2|T|\rceil+1$ cipherbits;
3. IF $|T|<|\times_{\boldsymbol{vec}[i]=1} Q_i^{[d_i]}|$ THEN $EC\leftarrow T$;
4. ELSE $EC\leftarrow\times_{\boldsymbol{vec}[i]=1} Q_i^{[d_i]}$; ENDIF
5. FOR each $ec\in EC$ DO
6. IF $\exists ec_0\in setEC_0$ s. t. $ec\subseteq ec_0$ THEN continue;
7. ELSE $supp\leftarrow countingFHE(ec,T)$; END IF
8. $recSum+=supp$;
9. SWITCH $result=judge(supp,recSum,|T|,k)$
10. CASE 1: $setEC\ 0\leftarrow setEC_0\cup\{ec\}$;
11. CASE 2,4: return false;
12. CASE 3: return true;
13. END SWITCH
14. END FOR
15. return true;

过程 *isKanonymous* 委托数据拥有者的代理来协助计算 $judge(supp,recSum,|T|,k)$(第 9 行). 代理得到数据拥有者的授权，可以解密 *supp*、*recSum* 和 k，能够计算并返回 5 种可能结果之一：(1) 支持度 *supp* 为 0 或 2. 支持度大于 0 小于 k 或 3. 累计支持度之和 *recSum* 大于等于记录数 $|T|$ 或 4. $|T|>recSum>|T|-k$ 或 5. 其它情况.

5.2 匿名化算法的云端运行协议

提出基于全同态加密和全域泛化的匿名化算法的云端运行协议，如图 4 所示. 结合云环境下保护隐私的 k-anonymous 数据共享模型的两个前提（第 3 节）、全同态加密技术的两个特点（第 3.4 节）和云端数据加密存储形式（第 4.2 节）以及匿名化算法流程，本协议运行如下.

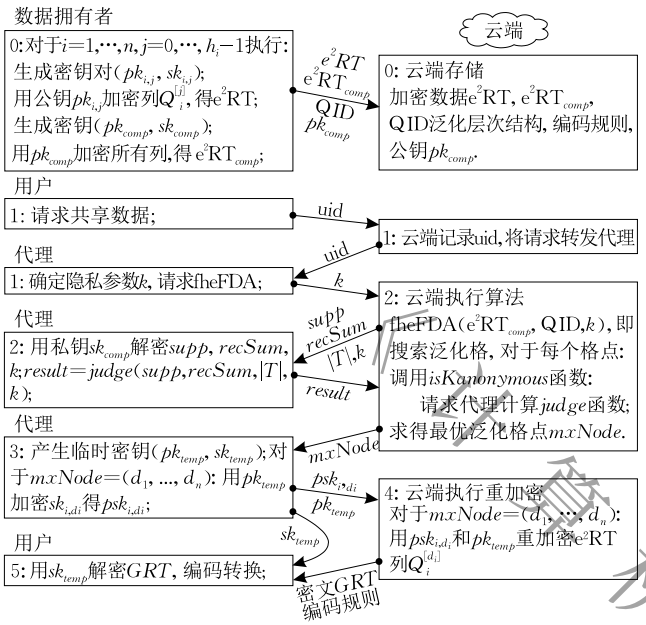


图 4 匿名化算法的云端运行协议

在协议的准备阶段，云端已经存储加密的扩展数据表 e²RT 及其副本 e²RT$_{comp}$、QID 泛化层次结构及其编码规则（Step 0）. 每当用户要共享数据时，请求先发送到云端、再转发数据拥有者的代理，然后由代理确定隐私参数 k、并指示云端开始执行匿名化算法（Step 1）. 执行算法过程中，云端将请求代理协助完成简单的比较运算（Step 2）.

算法结束时，云端将符合 k-anonymity 且精度最高的泛化格点 $mxNode=(d_1,\cdots,d_n)$ 传送给代理，由代理生成临时密钥对（pk_{temp}，sk_{temp}）、并用临时公钥 pk_{temp} 加密列 $Q_i^{[d_i]}$ 的私钥 $sk_{i,di}$ 得 $psk_{i,di}$（Step 3）.

云端根据代理发来的 pk_{temp} 和 $psk_{i,di}$ 重加密 e²RT 对应 $mxNode=(d_1,\cdots,d_n)$ 的列，得到泛化数据表 $GRT[Q_1^{[d_1]},\cdots,Q_n^{[d_n]}]$ 的密文形式（Step 4）.

用户从云端下载密文 GRT，从代理获得临时私钥 sk_{temp}，经过解密和编码转换，最终得到 k-anonymous 的数据表（Step 5）.

6 实验结果与分析

本文所提出算法是第一个基于全同态加密和全域泛化的隐私保护算法. 重点评价本文所提出四个优化策略以及本文所提出算法在多种全同态加密方案下的运行效率. 算法采用 C++语言编程. 全同态加密方案选用 GitHub 开源项目 libScarab、HElib、tfhe，分别实现第一代全同态加密方案 SV10[34]、第二代方案 BGV12[35]、第三代方案 CGG16[39].

数据集选用隐私保护研究领域常用的 Adult，来自加州大学欧文分校的机器学习库，有完整记录 45 222 条，主要属性的名称、底层属性值个数、泛化层次数如表 4 所示. 实验在 Dell 低端服务器上进行，配置为 CPU Intel Xeron E3-1225，主频 3.2GHz，内存 8GB，运行 CentOS release 6.3 操作系统.

表 4 Adult 数据集的属性与泛化层次

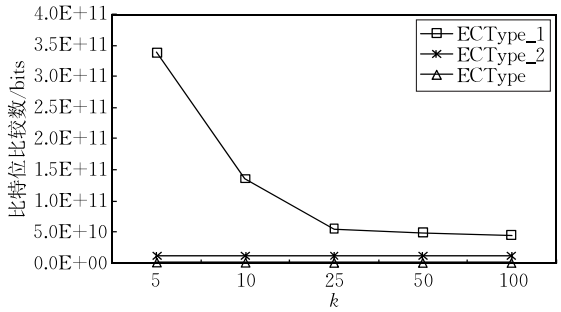
	属性名称	属性值个数	泛化层次数
1	Age	81	4
2	Hours per week	101	4
3	Native country	41	4
4	Sex	2	2
5	Race	5	3
6	Relationship	6	3
7	Education num	20	4
8	Education	16	5
9	Occupation	14	3

6.1 四种优化策略的评估

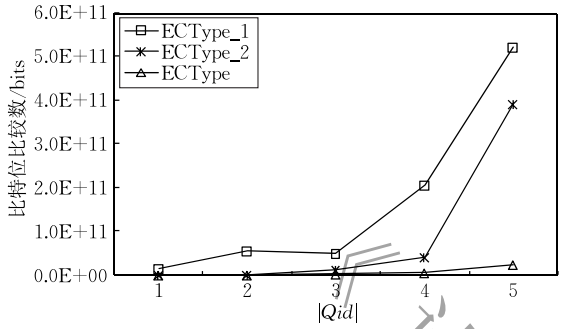
本文所提出算法的时间复杂度取决于被搜索的泛化格点个数、判定每个格点及其子向量时计算支持度的（子）等价类条数、每条（子）等价类的计算量三个因素. 第三个因素又取决于数据表记录数、每条记录的密文二进位数. 因此，算法时间复杂度归结为：$time_1$×密文位比对数+$time_2$×密文计数器加 1 操作总次数. 其中，$time_1$ 是比对单个密文位所需时间，$time_2$ 是一次密文计数器加 1 操作所需时间，均取决于实验软硬件平台以及全同态加密方案的具体实现. 因此，评价四种策略时“密文位比对数”是比“绝对时间”更有价值的指标.

首先，分析策略 1（Strategy ECType），如图 5 所示，ECType_1 表示固定采用第一种等价类生成方式，ECType_2 表示固定采用第二种，ECType 则是自动选择生成方式.

图 5(a)所示是随隐私参数 k 变化的密文位比对数. 当 k=5 时，ECType_1∶ECType_2∶ECType 是 120∶5∶1. 当 k=100 时，ECType_1∶ECType_2∶ECType 是 21∶6∶1. 图 5(b)所示是随准标识符 QID 属性个数变化的密文位比对数. 当|Qid|=1 时，ECType_1∶ECType_2∶ECType 是 686∶1∶1.



(a) 随隐私参数k变化的密文位比对数($|Qid|=3$)

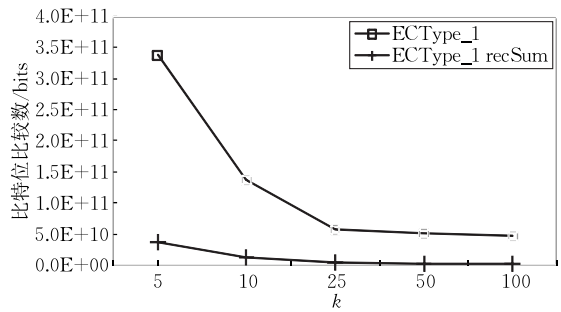


(b) 随准标识符QID个数变化的密文位比对数($k=50$)

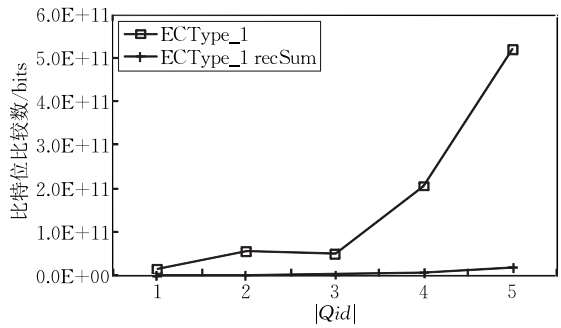
图 5　策略 1 Strategy ECType 分析

当$|Qid|=5$时，ECType_1∶ECType_2∶ECType 是 22∶16∶1. 实验表明：自动选取等价类集合生成方式远远优于固定采用某一种方式.

第二，分析策略 2(Strategy recSum). 本策略与 ECType_1 等价类生成方式联合使用效果特别好. 图 6(a)所示是随 k 变化的密文位比对数. 当 $k=5$ 时，ECType_1∶ECType_1 recSum＝9∶1. 当 $k=100$ 时，ECType_1∶ECType_1 recSum＝54∶1. 图 6(b)所示是随$|Qid|$变化的密文位比对数. 当$|Qid|=1$时，ECType_1 ∶ ECType_1 recSum＝328 ∶ 1. 当$|Qid|=5$ 时，ECType_1 ∶ ECType_1 recSum＝30 ∶ 1.



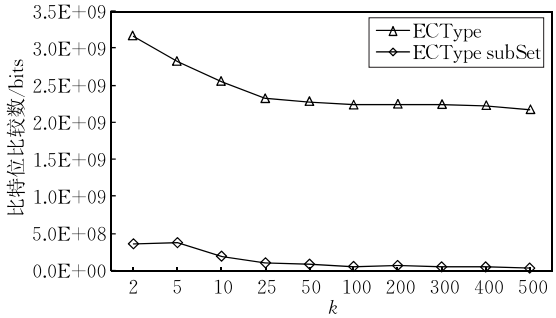
(a) 随隐私参数k变化的密文位比对数($|Qid|=3$)



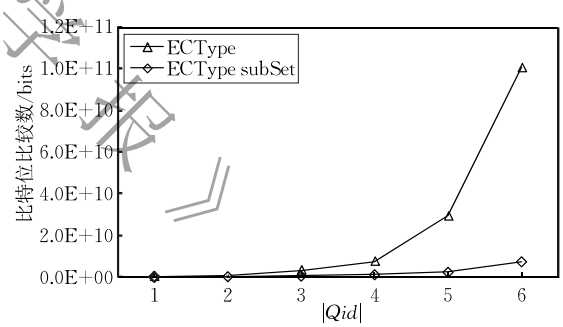
(b) 随准标识符QID属性个数变化的密文位比对数($k=50$)

图 6　策略 2 Strategy recSum 分析

第三，分析策略 3 Strategy subset. 其优化效果非常显著. 在采用策略 1 Strategy ECType 情况下，图 7 对比不使用和使用本策略的结果. 如图 7(a)所示是随 k 变化的密文位比对数. 当 $k=2$ 时，ECType∶ECType subSet＝8 ∶ 1. 当 $k=500$ 时，ECType∶ECType subSet＝53∶1. 图 7(b)所示是随$|Qid|$变化的密文位比对数. 当$|Qid|=1$时，ECType∶ECType subSet＝1∶1. 当$|Qid|=6$时，ECType∶ECType subSet＝14∶1.



(a) 随隐私参数k变化的密文位比对数($|Qid|=3$)

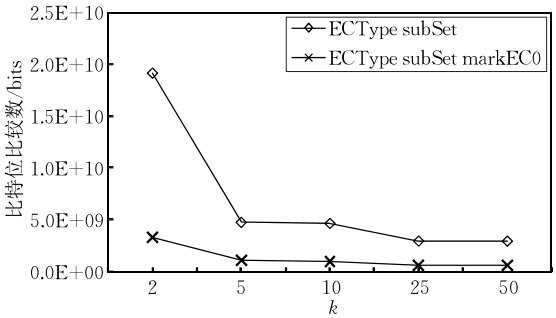


(b) 随准标识符QID属性个数变化的密文位比对数($k=2$)

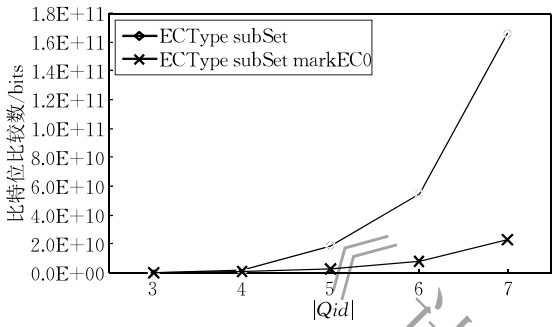
图 7　策略 3 Strategy subSet 分析

最后，分析策略 4 Strategy markEC0. 图 8 所示是一起使用 Strategy subSet 和 Strategy ECType 情况下，不使用和使用本策略的结果. 如图 8(a)所示是随 k 变化的密文位比对数. 当 $k=2$ 时，ECType subSet∶ECType subSet markEC0＝6∶1. 当 $k=50$ 时，ECType subSet ∶ ECType subSet markEC0＝4∶1.

如图 8(b)所示是随$|Qid|$变化的密文位比对数. 当$|Qid|=3$时，ECType subSet ∶ ECType subSet markEC0＝1 ∶ 1. $|Qid|=7$ 时，ECType subSet ∶ ECType subSet markEC0＝7 ∶ 1. 如果固定采用



(a) 随隐私参数k变化的密文位比对数($|Qid|=5$)



(b) 随准标识符QID属性个数变化的密文位比对数($k=2$)

图 8 策略 4 Strategy markEC0 分析

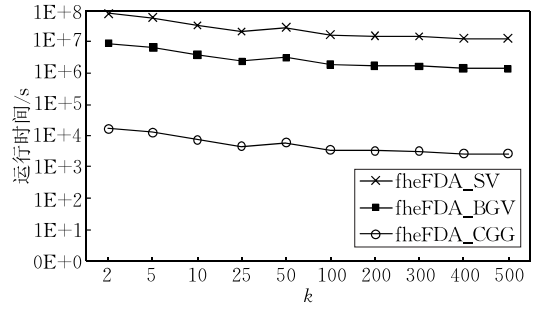
ECType_2 生成等价类方式,本策略效果将更显著.

6.2 匿名化算法效率评价

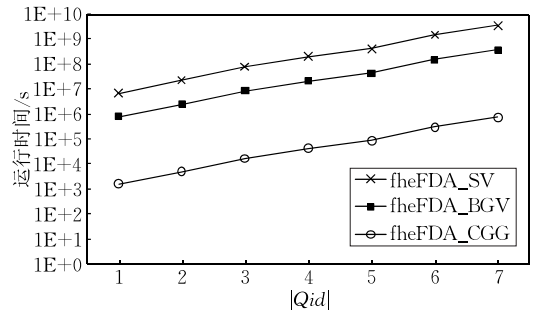
分析本文所提出算法在三种全同态加密方案下的运行效率. 图 9 所示,fheFDA_SV、fheFDA_BGV、fheFDA_CGG 分别是本文算法在 SV10 方案、BGV12 方案、CGG16 方案下运行时间.

首先,分析随隐私参数 k 变化的算法效率. 如图 9(a)所示,总趋势是运行时间随着 k 增加而减少. 原因是随 k 增加,符合 k-anonymity 的泛化格点应该在较高层,这使得搜索的格点个数增加,但是每个泛化格点的计算量却在下降(低层格点只需检查小部分等价类就能判定不符合 k-anonymity,而高层格点等价类总数少). 总的来讲,后者作用强于前者,所以运行时间是下降的趋势,但也有前者强于后者的例外情况,比如在 $k=50$ 时,两者运行时间都略有上升. 当 k 超过 10 之后,fheFDA_CGG 的运行时间在 46 min 到 2 h 之间,fheFDA_CGG 的运行效率分别比 fheFDA_BGV 和 fheFDA_SV 高 2 到 3 个数量级.

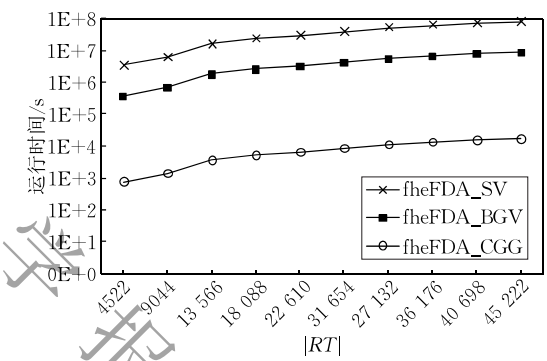
第二,分析随准标识符 QID 属性个数变化的算法效率. 如图 9(b)所示,随着 $|Qid|$ 增大,算法运行时间也增大. 原因是 $|Qid|$ 增大有两个作用,一是泛化格点数增加,即搜索空间变大,二是每个泛化格点需检查的等价类总数也在增加. 例如,fheFDA_CGG 的运行时间在 $|Qid|=1$ 时为 27 min,当 $|Qid|=2$ 时为 86 min. fheFDA_CGG 比 fheFDA_BGV 快 2 个数量级,fheFDA_BGV 比 fheFDA_SV 快 1 个数量级.



(a) 随隐私参数k变化的效率分析($|Qid|=3, |RT|=45\,222$)



(b) 随准标识符QID变化的效率分析($k=2, |RT|=45\,222$)



(c) 随数据记录条数变化的效率分析($|Qid|=3, k=2$)

图 9 本文算法效率分析

最后,分析随数据记录量 $|RT|$ 变化的算法效率. 如图 9(c)所示,运行时间随着 $|RT|$ 平稳增加. 原因是 $|RT|$ 的增加直接使得每个等价类支持度的计算量增加,也意味着策略 1 Strategy ECType 自动选择等价类集合生成方式的作用减弱.

简言之,实验评估表明所提出的基于全同态加密和全域泛化的隐私保护算法是可行的,所提出的四个优化策略大大提高算法效率. 所报告的是低档硬件平台单进程单线程执行的实验结果,在高性能硬件平台并行执行的时间将更短.

7 结 论

针对在云端执行匿名化处理过程本身可能受到攻击造成隐私泄露的情况,提出一种基于全同态加

密和全域泛化实施 k-anonymity 的匿名化算法及其在云端运行的协议.提出适合于全同态加密特点的搜索最优匿名解的四个优化策略及相应的数据加密存储形式.云环境下保护隐私数据共享模型的优点是数据可以有多种用途、可以采用不同的隐私保护技术.利用全同态加密方案的开源项目在隐私保护研究经典数据集上进行实验,结果表明所提出模型、策略、算法是有效的.

未来研究工作有两个方面:一是研究分布式计算模型进一步提高全同态加密计算的效率;二是研究不同隐私保护原则和匿名化技术与全同态加密技术的集成.

参　考　文　献

[1] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information//Proceedings of the 17th Symposium on Principles of Database Systems. Seattle, USA, 1998: 188

[2] Sweeney L. K-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570

[3] Machanavajjhala A, Gehrke J, Kifer D, et al. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 3

[4] Dwork C. Differential privacy//Proceedings of the 33rd International Colloquium on Automata, Languages and Programming. Venice, Italy, 2006: Part II: 1-12

[5] LeFevre K, DeWitt D J, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Baltimore, USA, 2005: 49-60

[6] Fung B C M, Wang K, Yu P S. Anonymizing classification data for privacy preservation. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(5): 711-725

[7] Liu J, Wang K. On optimal anonymization for l^+-diversity//Proceedings of the 26th IEEE International Conference on Data Engineering. Long Beach, USA, 2010: 213-224

[8] LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional k-anonymity//Proceedings of the 22nd International Conference on Data Engineering. Atlanta, USA, 2006: 25

[9] Xu J, Wang W, Pei J, et al. Utility-based anonymization for privacy preservation with less information loss. ACM SIGKDD Explorations Newsletter, 2006, 8(2): 21-30

[10] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 571-588

[11] He Y, Naughton J. Anonymization of set-valued data via top-down, local generalization//Proceedings of the 35th International Conference on Very Large Data Bases. Lyon, France, 2009: 934-945

[12] Wang K, Fung B C M, Yu P S. Handicapping attacker's confidence: An alternative to k-anonymization. Knowledge and Information Systems, 2007, 11(3): 345-368

[13] Evfimievski A, Gehrke J, Srikant R. Limiting privacy breaches in privacy preserving data mining //Proceedings of the 22nd ACM Symposium on Principels of Database Systems. San Diego, USA, 2003: 211-222

[14] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation//Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul, Korea, 2006: 139-150

[15] Rivest R, Adleman L, Dertouzos M. On data banks and privacy homomorphisms. Foundations of Secure Computation, 1978, 4(11): 169-180

[16] Gentry C. Fully homomorphic encryption using ideal lattices//Proceedings of the 41st Annual ACM Symposium on Theory of Computing. Bethesda, USA, 2009: 169-178

[17] Hu H, Xu J, Ren C, et al. Processing private queries over untrusted data cloud through privacy homomorphism//Proceedings of the 27th International Conference on Data Engineering. Hannover, Germany, 2011: 601-612

[18] Liu J, Li J, Xu S, Fung B. Secure outsourced frequent pattern mining by fully homomorphic encryption//Proceedings of the 17th International Conference on Big Data Analytics and Knowledge Discovery. Valencia, Spain, 2015: 70-81

[19] Puttaswamy K P N, Kruegel C, Zhao B Y. Silverline: Toward data confidentiality in storage-intensive cloud applications//Proceedings of the 2nd ACM Symposium on Cloud Computing. Cascais, Portugal, 2011: 10

[20] Zhang X, Liu C, Nepal S, et al. A privacy leakage upper-bound constraint based approach for cost-effective privacy preserving of intermediate datasets in Cloud. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(6): 1192-1202

[21] Zhang X, Yang L T, Liu C, Chen J. A scalable two-phase topdown specialization approach for data anonymization using mapreduce on cloud. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(2): 363-373

[22] Blass E-O, Pietro R D, Molva R, Önen M. Prism—privacy preserving search in mapreduce//Proceedings of the 12th International Conference on Privacy Enhancing Technologies. Vigo, Spain, 2012: 180-200

[23] Gao J, Yu J, Jin R, et al. Neighborhood-privacy protected shortest distance computing in cloud//Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. Athens, Greece, 2011: 409-420

[24] Roy I, Setty S T V, Kilzer A, et al. Airavat: Security and privacy for mapreduce//Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation. San Jose, USA, 2010: 297-312

[25] Wang C, Wang Q, Ren K, et al. Privacy-preserving public auditing for data storage security in cloud computing//Proceedings of the 29th IEEE International Conference on

Computer Communications. San Diego, USA, 2010: 525-533

[26] Wang B, Li B, Li H. Oruta: Privacy-preserving public auditing for shared data in the cloud. IEEE Transactions on Cloud Computing, 2014, 2(1): 43-56

[27] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy Beyond *k*-Anonymity and l-Diversity//Proceedings of the 23rd International Conference on Data Engineering. Istanbul, Turkey, 2007: 106-115

[28] Xiao X, Tao Y. M-invariance: Towards privacy preserving republication of dynamic datasets//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing, China, 2007: 689-700

[29] Li T, Li N, Zhang J, Molloy I. Slicing: A new approach for privacy preserving data publishing. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3): 561-574

[30] Terrovitis M, Liagouris J, Mamoulis N, Skiadopolous S. Privacy preservation by disassociation//Proceedings of the VLDB Endowment. Istanbul, Turkey, 2012, 5(10): 944-955

[31] Rivest R, Shamir A, Adleman L. A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 1978, 21(2): 120-126

[32] Paillier P. Public-key cryptosystems based on composite degree residuosity classes//Proceedings of the 1999 International Conference on the Theory and Application of Cryptographic Techniques. Prague, Czech Republic, 1999: 223-238

[33] Dijk M V, Gentry C, Halevi S, et al. Fully homomorphic encryption over the integers//Proceedings of the29th Annual International Conference on the Theory and Applications of Cryptographic Techniques. French Riviera, 2010: 24-43

[34] Smart N, Vercauteren F. Fully homomorphic encryption with relatively small key and ciphertext sizes//Proceedings of the 13th International Conference on Practice and Theory in Public Key Cryptography. Paris, France, 2010: 420-443

[35] Brakerski Z, Gentry C, Vaikuntanathan V. Fully homomorphic encryption without bootstrapping. ACM Transactions on Computation Theory, 2014, 6(3): 13

[36] Brakerski Z, Vaikuntanathan V. Fully homomorphic encryption from ring-LWE and security for key dependent messages//Proceedings of the 31st Annual Cryptology Conference. Santa Barbara, USA, 2011: 505-524

[37] Gentry C, Sahai A, Waters B. Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based//Proceedings of the 33rd Annual Cryptology Conference. Santa Barbara, USA, 2013: 75-92

[38] Ducas L, Micciancio D. FHEW: Bootstrapping homomorphic encryption in less than a second//Proceedings of the 34th Annual International Conference on the Theory and Applications of Cryptographic Techniques. Sofia, Bulgaria, 2015: 617-640

[39] Chillotti I, Gama N, Georgieva M, Izabachène M. Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds//Proceedings of the 22nd International Conference on the Theory and Application of Cryptology and Information Security. Hanoi, Vietnam, 2016: 3-33

[40] Wang H. Privacy-preserving data sharing in cloud computing. Journal of Computer Science and Technology, 2010, 25(3): 401-414

LIU Jun-Qiang, born in 1962, Ph. D., professor. His research interests include security and privacy preserving, and data mining.

CHEN Fang-Hui, born in 1992, M. S. candidate. Her research interests include privacy protection and data mining.

XU Cong-Fu, born in 1969, Ph. D., associate professor. His research interests include artificial intelligence, machine learning, data mining, and information fusion.

GUO Hong, born in 1964, M. S., senior engineer. His research interests include cloud computing and big data.

LI Ting, born in 1995, B. S. Her research interests include cryptography and privacy protection.

Background

This work is part of a research project titled "Security and privacy protection in knowledge discovery and data mining using cloud services", which are supported by the National Natural Science Foundation of China (Nos. 61272306, 61672449) and the Natural Science Fund of Zhejiang Province (LY17F020004). The research project aims at addressing the three categories of issues in data mining services using cloud computing. The first category of issues is content security, i. e., confidentiality issues with respect to the content of data and knowledge. The second is content privacy, i. e., privacy issues with respect to the content shared with a legitimate party. The third is behavior privacy, i. e., privacy issues with respect to the behavioral patterns of individuals who consume data mining services. This work integrates the fully homomorphic encryption into the anonymization process to address the second category of issues, i. e., content privacy.