面向数据发布和分析的差分隐私保护

张啸剑 孟小峰

(中国人民大学信息学院 北京 100872)

摘 要 随着数据分析和发布等应用需求的出现和发展,如何保护隐私数据和防止敏感信息泄露成为当前面临的重大挑战.基于k-匿名或者划分的隐私保护方法,只适应特定背景知识下的攻击而存在严重的局限性.差分隐私作为一种新出现的隐私保护框架,能够防止攻击者拥有任意背景知识下的攻击并提供有力的保护.文中对差分隐私保护领域已有的研究成果进行了总结,对该技术的基本原理和特征进行了阐述,重点介绍了当前该领域的研究热点:差分隐私下基于直方图的发布技术、基于划分的发布技术以及回归分析技术.在对已有技术深入对比分析的基础上,指出了差分隐私保护技术的未来发展方向.

关键词 差分隐私;数据发布;隐私保护;数据分析

中图法分类号 TP309 **DOI号** 10.3724/SP.J.1016.2014.00927

Differential Privacy in Data Publication and Analysis

ZHANG Xiao-Jian MENG Xiao-Feng

(*School of Information, Renmin University of China, Beijing* 100872)

Abstract As the emergence and development of application requirements such as data analysis and data publication, a challenge to those applications is to protect private data and prevent sensitive information from disclosure. However, most existing methods based on k-anonymity or partition-based have serious limitations because they only preserve individual privacy under special assumption of adversary's background knowledge. Differential privacy has emerged as a new paradigm for privacy protection with strong privacy guarantees against adversaries with arbitrary background knowledge. This paper surveys the state of the art of differential privacy for data publication and analysis. The mechanisms and properties of this model are described, while our focuses are put on private data releases in terms of histogram and partition techniques, and analysis based on regression skills. Following the comprehensive comparison and analysis of existing works, future research directions are put forward.

Keywords differential privacy; data publication; privacy-preserving; data analysis

1 引 言

信息技术的飞速发展使得各类数据的发布、采集、存储和分析变得方便快捷.例如,医院电子病例记录病人基本信息、疾病信息及药品购买记录;人口普查记录市民的家庭住址以及收入情况;金融业务服务会记录客户私有信息及其交易行为等.而这些数据的收集和发布直接给个人隐私造成威胁.一方面,如果数据拥有者直接发布隐含的敏感信息,而不

收稿日期:2012-12-26;最终修改稿收到日期:2013-11-25.本课题得到国家自然科学基金(61379050,91224008),国家"八六三"高技术研究发展计划项目基金(2013AA013204)及高等学校博士学科点专项科研基金(20130004130001)资助.张啸剑,男,1980年生,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为差分隐私、数据挖掘、图数据管理.E-mail:xjzhang82@126.com.孟小峰,男,1964年生,教授,博士生导师,主要研究领域为Web数据管理、移动数据管理、XML数据管理、云数据管理等.

采用适当数据保护技术，将可能造成个人的隐私泄露.另一方面，对发布后的数据进行分析也给数据的隐私带来了威胁.例如，采用数据挖掘和机器学习技术对医疗病例记录和搜索日志进行挖掘，可以获得病人所患何种疾病以及用户搜索的行为模式等敏感信息.隐私保护技术可以解决数据发布和数据分析带来的隐私威胁问题.如何发布和分析而又不泄露隐私信息是隐私保护技术的主要目的.近年来出现了许多基于 k-匿名[1]和划分的隐私保护方法(例如，l-diversity[2]、t-closeness[3]、(α, k)-anonymity[4])，尽管这些方法能够保护数据的更多细节，但是均需要特殊的攻击假设和背景知识.此外，针对上述隐私保护方法，出现了一些新的攻击模型，例如，组合攻击[5]、前景知识攻击[6]等.这些新的攻击模型对上述方法的有效性提出了严峻挑战.

目前，差分隐私[7-11]已经成为一种新的隐私保护模型，该模型不关心攻击者拥有多少背景知识，通过向查询或者分析结果中添加适当噪音来达到隐私保护效果.类似传统的隐私保护技术，实施差分隐私保护技术主要考虑两个方面的问题：(1)如何保证设计的算法满足差分隐私，以确保数据隐私不被泄露；(2)如何减少噪音带来的误差，以提高数据的可用性.

本文立足于数据库应用领域，对差分隐私保护技术的最新研究进展和研究方向进行综述.一方面对差分隐私的基本定义、特性、度量和噪音机制进行阐述；另一方面，针对数据库领域与差分隐私保护技术相关的研究方向进行分析，其中着重介绍基于差分隐私的数据发布和分析技术，而后对这些技术的优缺点、效果与效率进行综合分析与对比.目前差分隐私保护技术在数据库领域主要集中在“数据发布”、“数据挖掘”和“机器学习”3 大领域，本文着重介绍该技术在这 3 个领域中的应用.

本文第 2 节介绍差分隐私保护基础知识；第 3、4 节介绍差分隐私保护的研究方向与框架；第 5 节和第 6 节重点介绍差分隐私保护下的数据发布与分析技术，并对相应技术进行总结和分析；第 7 节和第 8 节分别介绍交互式查询处理和相应的系统；最后总结和展望未来工作.

2 基础知识

2.1 差分隐私定义

隐私是指个人、组织机构等实体不愿意被外部知晓的信息[12].例如，个人的薪资、医疗记录等.虽然出现了多种基于 k-匿名和划分隐私保护框架的保护方法，而差分隐私保护技术被公认为比较严格和强健的保护模型.该保护模型的基本思想是对原始数据、对原始数据的转换或者是对统计结果添加噪音来达到隐私保护效果.该保护方法可以确保在某一数据集中插入或者删除一条记录的操作不会影响任何计算的输出结果.另外，该保护模型不关心攻击者所具有的背景知识，即使攻击者已经掌握除某一条记录之外的所有记录的信息，该记录的隐私也无法被披露.差分隐私的形式化定义如下.

定义 1[7]. 给定数据集 D 和 D'，二者互相之间至多相差一条记录，即 $|D \Delta D'| \leq 1$.给定一个隐私算法 A，$Range(A)$ 为 A 的取值范围，若算法 A 在数据集 D 和 D' 上任意输出结果 $O(O \in Range(A))$ 满足下列不等式，则 A 满足 ε-差分隐私.

$$Pr[A(D)=O] \leq e^{\varepsilon} \times Pr[A(D')=O] \quad (1)$$

其中，概率 $Pr[\cdot]$ 由算法 A 的随机性控制，也表示隐私被披露的风险；隐私预算参数 ε 表示隐私保护程度，ε 越小隐私保护程度越高.

从定义 1 可以看出差分隐私技术限制了任意一条记录对算法 A 输出结果的影响.该定义是从理论角度确保算法 A 满足 ε-差分隐私，而要实现差分隐私保护需要噪音机制的介入.

2.2 噪音机制

噪音机制是实现差分隐私保护的主要技术，常用的噪音添加机制分别为拉普拉斯机制[13]与指数机制[14].而基于不同噪音机制且满足差分隐私的算法所需噪音大小与全局敏感性(Global Sensitive)密切相关.

定义 2[13]. 对于任意一个函数 $f: D \rightarrow R^d$，函数 f 的全局敏感性为

$$\Delta f = \max_{D, D'} \| f(D) - f(D') \|_p \quad (2)$$

其中，D 和 D' 至多相差一条记录，R 表示所映射的实数空间，d 表示函数 f 的查询维度，p 表示度量 Δf 使用的 L_p 距离，通常使用 L_1 来度量.

2.2.1 拉普拉斯机制

文献[13]提出了拉普拉斯机制，该机制通过拉普拉斯分布产生的噪音扰动真实输出值来实现差分隐私保护.

定理 1[13]. 对于任一个函数 $f: D \rightarrow R^d$，若算法 A 的输出结果满足下列等式，则 A 满足 ε-差分隐私.

$$A(D)=f(D)+\langle Lap_1(\Delta f/\varepsilon),\cdots,Lap_d(\Delta f/\varepsilon)\rangle \tag{3}$$

其中，$Lap_i(\Delta f/\varepsilon)(1\leq i\leq d)$是相互独立的拉普拉斯变量，噪音量大小与 Δf 成正比，与 ε 成反比. 算法 A 的全局敏感性越大，所需噪音越大.

从式(3)可知，$A(D)$中第 $i(1\leq i\leq d)$个元素由拉普拉斯噪音引起的标准绝对误差与方差分别为

$$error^i_{abs}=E\left|A(D)_i-f(D)_i\right|=E\left|Lap\left(\frac{\Delta f}{\varepsilon}\right)\right|=\frac{\sqrt{2}\Delta f}{\varepsilon} \tag{4}$$

$$error^i_{var}=E\left(A(D)_i-f(D)_i\right)^2=\frac{2(\Delta f)^2}{\varepsilon^2} \tag{5}$$

2.2.2　指数机制

文献[14]提出了指数机制，该机制主要是处理一些输出结果为非数值型的算法，例如，分类操作中分裂属性的选择问题[61]. 该机制的关键技术是如何设计打分函数 $u(D,r)(r\in O)$，其中 r 表示从输出域 O 中所选择的输出项.

定理 2[14]. 给定一个打分函数 $u:(D\times O)\rightarrow R$，若算法 A 满足下列等式，则 A 满足 ε-差分隐私.

$$A(D,u)=\left\{r:\left|Pr[r\in O]\propto\exp\left(\frac{\varepsilon u(D,r)}{2\Delta u}\right)\right.\right\} \tag{6}$$

其中，Δu 为打分函数 $u(D,r)$ 的全局敏感性. 由式(6)可知，打分越高，被选择输出的概率越大.

2.3　差分隐私的组合特性

差分隐私保护技术本身蕴含着序列组合性与并行组合性两种重要的组合性质[19].

性质 1[19]. 给定数据库 D 与 n 个随机算法 $A_1,\cdots,A_n$，且 $A_i(1\leq i\leq n)$ 满足 ε_i-差分隐私，则 $\{A_1,\cdots,A_n\}$ 在 D 上的序列组合满足ε-差分隐私，$\varepsilon=\sum\varepsilon_i$.

性质 2[19]. 设 D 为一个隐私数据库，被划分成 n 个不相交的子集，$D=\{D_1,\cdots,D_n\}$，设 A 为任一个随机算法满足 ε-差分隐私. 则算法 A 在$\{D_1,\cdots,D_n\}$上的系列操作满足 ε-差分隐私.

这两种性质在证明算法是否满足差分隐私以及在隐私预算分配过程中起着重要作用.

2.4　差分隐私保护方法的性能度量

满足差分隐私的保护算法需要在保护隐私的同时，又要兼顾保护后数据的可用性以及隐私预算 ε 的分配策略是否合理. 通常包括 3 个方面对隐私保护算法进行度量.

(1)算法误差. 常用的应用型误差度量方法包括相对误差[21]、绝对误差[22]、误差的方差[23]以及欧式距离[24]等. 此外，数据依赖情况下的 Q 操作，必须考虑信息缺损带来的误差①.

(2)算法性能. 一般利用时间复杂度与渐近噪音误差边界对算法的性能进行评估.

(3)ε 的合理分配. 隐私预算 ε 代表着数据隐私保护程度. 一旦耗尽 ε，将破坏差分隐私，算法本身也就失去了意义. 因此，合理的预算分配策略要尽可能使 ε 的生命周期持续长一些. 常用的分配策略[25]包括线性分配、均匀分配、指数分配、自适用性分配以及混合策略分配等.

3　主要研究方向

差分隐私作为新兴的隐私保护技术，在理论研究和实际应用方面具有非常重要的价值. 该技术首先出现在统计数据库领域，然后，又扩展到其它领域，例如机器学习、安全通信等. 数据库领域中差分隐私保护技术的主要研究方向如表 1 所示.

表 1　差分隐私保护研究方向

研究方向	示例
基于差分隐私的数据发布	Histogram[26-29,68]、DataCube[50-51,72]、Partitioning[34]、Sampling & Filtering[40]
面向数据挖掘和学习的差分隐私保护技术	DiffGen[62]、PrivBasis[58]、SmartTrunc[59]、Diff-FPM[20]、ObjectivePerb[68]、PrivateSVM[66]、TF[57]、FM[16] DiffP-C4.5[61]
基于差分隐私的查询处理	Laplace[13,18]、Privlet[23]、Linear Query[15]、Batch Query[53]
基于差分隐私的应用系统	PINQ[19]、GUPT[52]、Airavat[73]

从表 1 可以看出，该技术的研究方向是由实际应用中不同的隐私需求而决定的. 基于差分隐私的数据发布技术主要是采用非交互式框架发布敏感数据的统计信息，并且使得发布数据能够满足数据分析者的需求. 常采用的发布技术有直方图、划分以及采样-过滤等；而面向数据挖掘和机器学习的差分隐私保护技术主要解决高层隐私需求带来的问题，例如，top-k 频繁模式挖掘、分类以及回归分析等. 如何设计满足差分隐私的挖掘和学习算法是其主要目的；基于差分隐私的查询处理技术主要解决如何以较小的隐私预算与较低的误差来响应查询，例如交互式框架下的线性与批量查询；基于差分隐私的应用系统则是为了提供几种在各类应用环境中可以通用的系统.

① 这种误差通常由数据的结构变化所导致的，例如 k-d 树索引空间数据[33]、直方图相邻桶的合并[27]等.

4 基于差分隐私的数据保护框架

差分隐私下数据保护框架通常有两种：交互式框架和非交互式框架. 交互式的差分隐私保护框架也可以称之为在线查询框架，其基本结构如图 1 所示. 当数据分析者通过查询接口提交查询 Q 时，数据拥有者会根据查询需求，设计满足差分隐私的查询算法，经过差分隐私算法过滤后，把结果 O' 返回给用户. 分析者提交的查询通常包含一定的语义约束[26]，使得返回结果的可用性较低. 数据拥有者常采用后置处理(Post-Processing)[17]技术对噪音结果进行求精处理. 由于交互式框架只允许数据分析者通过查询接口提交查询，查询数目决定着该框架的误差和性能，若提交查询的数目超过某个上界，隐私预算 ε 会被耗尽，该框架则不能满足差分隐私. 该框架所支持的查询通常包括聚集查询[50]、批量查询[53]以及提交的数据挖掘任务[58]等.

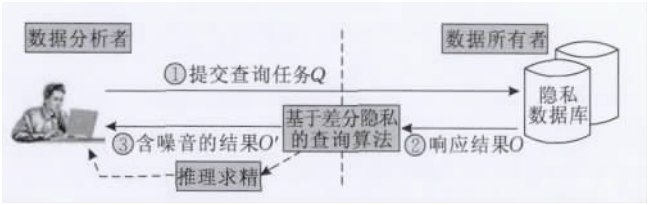

图 1 交互式框架

非交互式的差分隐私保护框架也称之为离线发布框架，其基本结构如图 2 所示. 数据拥有者通过差分隐私发布算法来发布数据库的相关统计信息. 数据分析者根据发布数据库提交查询或者挖掘任务 Q 以及得到噪音结果 O'. 非交互式发布框架下的主要研究是如何设计高效的发布算法，该类算法既满足差分隐私，又具有高的可用性. 目前，数据拥有者采用数据压缩、数据转换与采样过滤等技术对原始数据进行处理以达到缩减发布误差和查询误差的目的. 此外，数据发布过程中，合理的隐私预算分配策略也是保证差分隐私成立的关键.

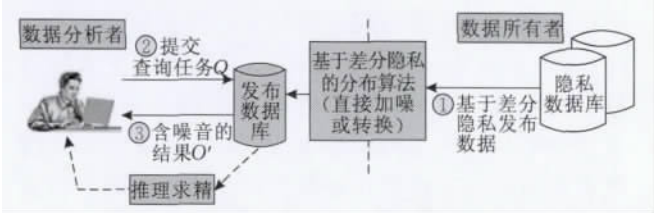

图 2 非交互式框架

差分隐私的研究均是基于上述两种框架展开的，下面从数据发布、数据挖掘与机器学习、查询处理以及相应存在系统的角度对国内外研究现状进行深入阐述.

5 发布技术的分类与性能评估

近年来，如何发布具有代表性的数据，而不披露数据的隐私已成为数据库领域的研究热点. 根据非交互式保护框架可知，数据发布方法一般可分为两类.

(1) 先对原始数据或者原始数据的统计信息添加噪音，然后对加过噪音后的数据采用规划策略(例如，二次规划、凸规划等)进行优化，最后发布优化结果. 而这类方法的隐私代价通常比较大. 本文对这类发布方法称之为策略 1. 该策略的基本思想如图 3 所示.

输入：原始数据 $X=\{x_1,x_2,\cdots,x_n\}$
1. 添加噪音对 X 进行扰动，$\widetilde{X}=\{x_1+Lap(\Delta f/\varepsilon),x_2+Lap(\Delta f/\varepsilon),\cdots,x_n+Lap(\Delta f/\varepsilon)\}$
2. 采用后置处理技术把 $\widetilde{X}$ 优化成为 $\bar{x}$

输出：优化后的发布数据 $\bar{x}$

图 3 策略 1 发布流程

(2) 先转换或者压缩原始数据，再对转换后的数据添加噪音. 这类方法主要针对如何减少发布误差以及如何提高数据可用性等. 尽管这种策略响应查询的精度较高，然而数据转换或者压缩会带来原始数据的信息缺损. 本文对这类发布方法称之为策略 2. 该策略的基本思想如图 4 所示.

输入：原始数据 $X=\{x_1,x_2,\cdots,x_n\}$
1. 将 X 转换或压缩成 $X'=\{x'_1,x'_2,\cdots,x'_n\}$，降低敏感性和噪音需求
2. 对 X' 添加噪音，则 $X''=\{x'_1+Lap(\Delta f/\varepsilon),x'_2+Lap(\Delta f/\varepsilon),\cdots,x'_n+Lap(\Delta f/\varepsilon)\}$

输出：优化后的发布数据 X''

图 4 策略 2 发布流程

基于上述两类发布策略，已有的发布技术主要分为两类：(1) 以直方图为发布标准的方法；(2) 基于划分的发布方法.

5.1 基于差分隐私的直方图发布方法

直方图使用分箱技术近似描述数据统计信息，将一个比较大的数据集按照某属性划分成不相交的桶，每个桶由一个数字表示其特征. 直方图可以分成等宽直方图(Equi-width Histogram)[29]、V-优化直方图(V-optimal Histogram)[29]等多种类型.

图 5 中的等宽直方图表示了表 2 在已知 Age

属性取值后 HIV+的分布情况. 直接发布图 5 中的直方图，会导致表 2 中个人的隐私泄露. 例如，假设攻击者知道了 Alice 的年龄为 52 岁，但不知道她是否感染 HIV+. 如果该攻击者获得了桶[50，60]中除 Alice 之外其他人的病况(例如感染 HIV+的病人计数为 2)，通过直方图的桶[50，60]计数 3，能够推理出 Alice 感染了 HIV+病毒. 下面根据不同的发布策略来分类阐述直方图的发布方法.

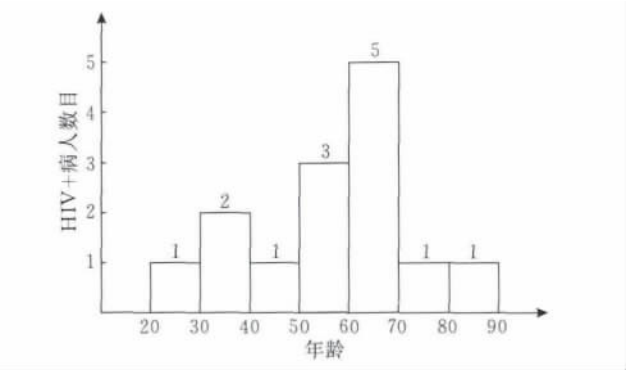

图 5　无噪音直方图发布

表 2　原始数据集

姓名	年龄	HIV+
Alice	52	Yes
Bob	43	Yes
Carol	32	No
Dave	41	Yes
…	…	…

5.1.1　基于策略 1 的直方图发布方法

策略 1 下的直方图发布方法，通常直接为每个桶的计数添加拉普拉斯噪音，进而达到扰动真实计数的效果. 由于连续的桶是相互独立的，在原始数据集中添加或者删除一条记录，最多影响直方图中 Δf 个桶的计数情况，每个桶的噪音大小为 $Lap(\Delta f/\varepsilon)$. 策略 1 下 Δf 通常为 1. 例如删除表 2 中的 Alice 记录，只影响桶[50，60]的计数，因此，图 5 中每个桶的噪音量为 $Lap(1/\varepsilon)$. 直方图通常支持单位长度(Unit-length)的范围查询①和较长(Longer-length)②的范围查询.

LP[13]是策略 1 下直方图发布的早期代表方法. 该方法结合拉普拉斯机制考虑了如何发布满足差分隐私的等宽直方图. 给定具有 n 个等宽桶的原始直方图 $H=\{H_1,H_2,\cdots,H_n\}$，$\widetilde{H}$ 表示 LP 发布的直方图，如下式所示.

$$\widetilde{H}=\{\widetilde{H}_1,\widetilde{H}_2,\cdots,\widetilde{H}_n\},\ \widetilde{H}_i=H_i+Lap(1/\varepsilon).$$

然而，LP 的缺陷是在响应长范围查询时，噪音的累加会导致很大的查询误差. 由式(5)可知，$\widetilde{H}$ 的噪音方差为 $2n/\varepsilon^2$，渐近误差边界为 $O(n/\varepsilon^2)$，因此大的 n 值会导致噪音误差过大，进而使得 $\widetilde{H}$ 的可用性很低. 为了弥补 LP 方法的不足，许多研究者提出了如何利用后置处理技术提高发布后等宽直方图的可用性和精度. 后置处理技术通常把提高 $\widetilde{H}$ 可用性转换成一致性约束③(Consistency Constraints)条件下的二次规划问题.

与 LP 不同，Boost1[26]考虑了如何利用一致性约束条件与最小二乘法对 $\widetilde{H}$ 进行约束推理. 设 $\overline{H}$ 为后置处理之后的直方图，则 Boost1 在约束条件下求目标函数 $\min\|\overline{H}-\widetilde{H}\|_2$ 的可行解或者最小解. Boost1 的渐近误差边界为 $O(\log^3 n/\varepsilon^2)$，因此，该法最终所发布的直方图 $\overline{H}$ 比 LP 发布的 $\widetilde{H}$ 的精度和可用性要高. 由目标函数 $\min\|\overline{H}-\widetilde{H}\|_2$ 可知，Boost1 一开始并没有减少每个桶的噪音量，每个桶的噪音量仍然为 $Lap(1/\varepsilon)$，只是在得到 $\widetilde{H}$ 的基础上进行了求精后置处理. 另外，该方法仅支持一维无归属直方图(Unattributed Histogram)④的发布，且只能响应单位长度的范围查询.

为了响应较长范围的计数查询，NoiseFirst[27-28]借鉴 V-优化直方图技术对 LP 所产生的直方图 $\widetilde{H}$ 进行后置处理，即是对 $\widetilde{H}$ 进行重构. V-优化直方图的精髓是采用动态规划技术合并邻接相似的桶，并最小化重构目标函数. 该目标函数通常采用平方误差和 SSE(Sum of Squared Error)度量. 差分隐私环境下的 NoiseFirst 在发布 V-优化直方图时，存在重构误差与噪音误差这两种误差，因此，该方法的目标函数由这两部分构成，如式(7)所示.

$$\min E\left(SSE(\widetilde{H},\overline{H})+\frac{2n-4k}{\varepsilon^2}\right)\qquad(7)$$

以图 5 中的直方图 H 阐述 NoiseFirst 方法的思想. 设 $\widetilde{H}$、$\overline{H}$ 分别为 LP 和 NoiseFirst 作用后的等宽直方图与 V-优化直方图，如图 6(a)和(b)所示. $H=\{1,2,1,3,5,1,1\}$，$\widetilde{H}=\{2,1,3,4,4,0,2\}$，$\overline{H}=\{2,2,2,4,4,1,1\}$. 由 SSE 计算得出 $SSE(\widetilde{H},H)=10$，$SSE(\overline{H},H)=4$，显然 V-优化直方图的误差比较小. 虽然 NoiseFirst 支持较长范围计数查询，并且渐近噪音误差边界 $O(n-k/\varepsilon^2)$ 小于 LP 的误差边界

① 注意单位长度是指直方图一个桶的大小，参见文献[13]的介绍.
② 长范围查询是指一个查询中包含多个邻接的桶，参见文献[26].
③ 一致性约束通常是指用户提交的查询语义，例如用户要求直方图的桶计数按照升序排列，参见文献[26]的介绍.
④ 注意该类直方图不考虑桶本身蕴含的语义，只关心桶所对应属性的分布情况，参见文献[26].

$O(n/\varepsilon^2)$，而该方法的缺陷在于仅支持一维的 V-优化直方图发布，并且在确定 $\overline{H}$ 的桶个数 k 时，没有采用启发式规则，只是人工设置 $k=n/10$. 因此，这样的 k 值无法均衡重构误差与噪音误差.

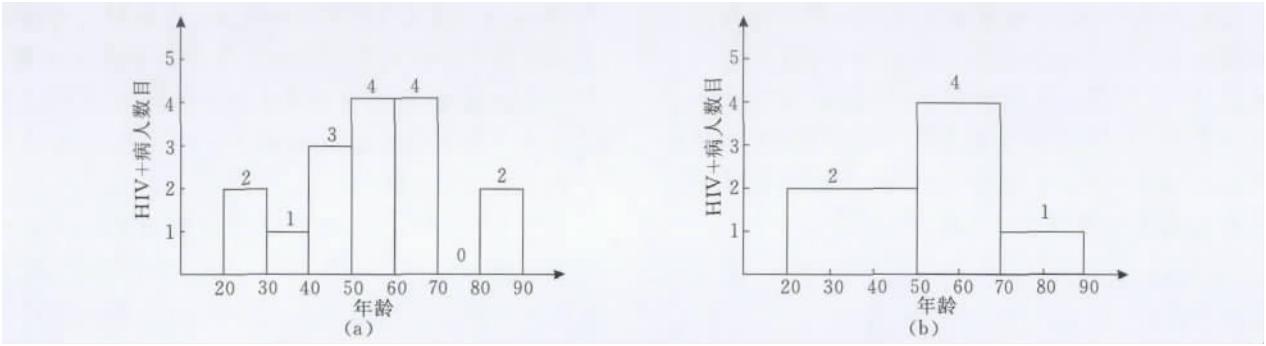

图 6 噪音直方图重构

除此之外，文献[30,39]首次提出了多维直方图发布方法 DPCube. 该方法采用单元(Cell)划分对原始数据集进行分割，同时为每个单元计数添加拉普拉斯噪音，然后采用 kd-树结构对所有的单元进行后置处理，最后获得多维 V-优化直方图. DPCube 方法的主要缺陷在于查询精度不稳定、多维范围查询偏差较大，同时发布误差也比较大.

以上介绍了策略 1 下的直方图发布方法. 表 3 对上述方法的优缺点以及性能进行了对比分析，其中计算开销是指添加噪音时的计算代价，分为“高”、“中”、“低”3 种级别；实际精度是指响应查询时返回结果的准确性，分为“好”、“一般”、“差”3 种级别；ε 的分配是指是否合理地利用了整体隐私预算，分为“合理”、“比较合理”两种级别；n 是指原始直方图桶的个数，k 是指合并之后桶的个数，d 是数据的维度.

表 3 策略 1 下直方图发布技术的对比分析

方法名称	主要优点	主要缺点	计算开销	ε 的分配	实际精度	渐近噪音误差边界
LP[13]	支持单位长度的范围查询	发布误差大，直方图的可用性低	低；仅涉及添加噪音开销	合理；统一应用所有 ε	差；主要取决于等宽直方图桶的个数	$O(n/\varepsilon^2)$
Boost1[26]	支持较高精度的单位长度范围查询	仅适用于数据独立情况下的一维直方图	中；额外的后置处理开销	合理；统一应用所有 ε	好；主要取决于后置处理方法	$O(\log^3 n/\varepsilon^2)$
NoiseFirst[27]	支持较长范围计数查询，查询精度较高	仅适用于一维直方图，无法均衡重构与噪音误差	高；额外的直方图重构开销	合理；统一应用所有 ε	好；主要取决于 V-优化方法	$O(n-k/\varepsilon^2)$
DPCube[30]	支持多维的单位长度与较长范围计数查询	发布误差大，可用性低，查询精度不稳定	高；额外的直方图重构开销	比较合理；ε 被划分成两份	一般；主要取决于维度 d	$O(n^d/\varepsilon^2)$

从表 3 可以看出，虽然 4 种方法都满足 ε-差分隐私，然而每种发布方法都有不同的特点，在不同的应用下，它们的适用范围、性能以及效果等不尽相同. 针对一维直方图，Boost1 和 NoiseFirst 效果与实际精度比 LP 和 DPCube 好，然而，Boost1 和 NoiseFirst 方法仅适用于一维直方图，扩展性比较差. 另外，NoiseFirst 方法的误差与计算开销很高，其主要原因是直方图的重构增加了额外误差.

虽然 DPCube 支持多维直方图发布，然而该方法的误差与计算开销也非常高，实际精度一般. 因此，在策略 1 下，实际精度比较好的、误差比较低的、并且能够支持单位和长范围计数查询的多维直方图发布是未来的研究方向. 此外，如何设计高效的后置处理算法也是未来的研究方向.

5.1.2 基于策略 2 的直方图发布方法

基于策略 2 的发布方法均是立足于原始直方图，在为各个桶计数添加噪音之前，先对直方图的自身结构进行重新组织，然后再对重组之后的结构添加噪音. 这种操作不但能够提供精确的长范围计数查询结果，而且也能够减少噪音误差. 原始直方图结构重组方式包括 3 类：(1) 按照层次树结构重组直方图，代表方法是 Privelet[23] 和 Boost2[26]；(2) 采用聚类重新划分直方图的各个桶，代表方法是 StructureFirst[27-28] 和 P-HPartition[31]；(3) 采用傅里叶变换有损压缩直方图，该类方法的代表是 FPA[32] 和 EDFP[31].

(1) 按照层次树结构重组直方图. 该类方法的主要目的是采用树结构对原始直方图进行组织，利

用树的层次特征精确地响应较长范围的计数查询.

Boost2[26]利用 m-ary 树对等宽直方图进行重新组织,其中 m 为该树的扇出.该方法根据 m-ary 树的高度 $1+\log_m n$ 决定重新构造直方图的敏感性,其中 n 表示桶的个数.因此,树中每个结点添加拉普拉斯噪音的大小为 $Lap(1+\log_m n/\varepsilon)$.例如图 7(a)中结点[40,50]的计数为 1,噪音量为 $Lap(1+\log_2 4/\varepsilon)$.为了能够精确地响应范围计数查询,Boost2 采用了查询语义一致性约束与最小二乘法相结合的后置处理技术,对 m-ary 树中的噪音计数进行约束推理.

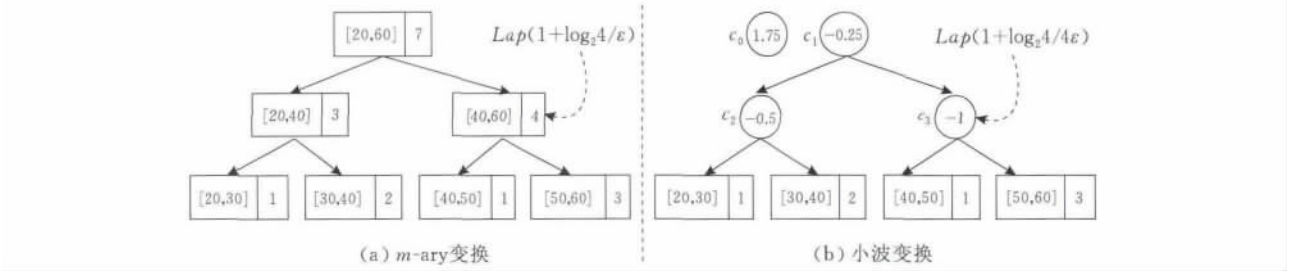

图 7　直方图重组示例

与 Boost2 不同,Privelet[23]采用哈尔小波(Haar Wavelet)变换对原始等宽直方图进行转换.哈尔小波变换可以看着一颗完全二叉树构造过程,在原始数据库中添加或者删除一条记录,至多影响 $\Delta f=1+\log_2 n$ 个小波系数的变化,其中 n 表示直方图中桶的个数.Privelet 采用式(8)计算每个小波系数 c_i,其中,$a_1(a_2)$表示结点 n_i 的左(右)子树中所有叶子结点中计数的均值.

$$c_1 = a_1 - a_2/2 \tag{8}$$

以图 5 中直方图的前 4 个桶计数为例来说明 c_i 的计算.例如图 7(a)中 $c_1=-0.25=(1.5-2)/2$,c_0 表示基系数,是所有叶子结点的均值.

根据式(3)与敏感性 Δf 可知,每个小波系数所添加的噪音大小为 $Lap((1+\log_2 n)/(\varepsilon\cdot W_{\text{Haar}}(c_i)))$,其中,$W_{\text{Haar}}(c_i)=2^{h-i+1}$,$h$ 表示小波树高度,i 表示系数 c_i 所在树的层数.

采用式(9)把小波系数逆推某个叶子结点 v 中的桶计数 $b_{(v)}$.

$$b_{(v)} = c_0 + \sum_{i=1}^{h}(f_i\cdot c_i) \tag{9}$$

其中,若 v 在 c_i 的左(右)子树中,则 $f_i=1(-1)$.

例如图 7(b)中,结点 c_3 的小波系数 -1 所添加的噪音量为 $Lap(1+\log_2 4/4\varepsilon)$.该结点的噪音系数值为 $-1+Lap(1+\log_2 4/4\varepsilon)$,根据式(9)可以推导出长范围[40,50]的噪音计数.

Privelet 与 Boost2 的优点在于能够比较精确地响应较长范围的计数查询,而它们的不足在于查询敏感性比较高,实际的性能比较差.通常直方图实际的桶个数 n 会非常大,层次树的中间结点数目可能达到指数级别,这样直接导致响应长范围计数查询的质量很差.而且 Boost2 方法仅适用于一维直方图的发布.此外,Privelet 与 Boost2 均没有讨论层次树的扇出 m 对最终发布精度的影响.扇出 m 的选择决定着树的高度以及每层应分得的隐私预算.因此,如何基于层次树结构重组方法,发布多维直方图是未来的研究方向;此外,如何设定合理的扇出 m 也是未来的一个研究方向.

(2)采用聚类重新划分直方图的各个桶.该类方法主要考虑如何利用基于有损压缩的聚类技术来有效减少直方图的查询敏感性.

由式(3)可知,可以从两方面来提高直方图的发布精度:(1)减少查询敏感性;(2)合理地分配隐私预算.StructureFirst[27-28]发现利用 V-优化直方图合并原始等宽直方图中近邻相似的桶,可以有效地减少查询敏感性.设原始直方图有 n 个单位长度的桶,相应的查询敏感性为 Δf,V-优化合并之后桶的个数为 k.设桶 $H_i(1\le i\le k)$ 合并了 p 个单位长度的桶,则 Δf 变成了原来的 $1/p$,噪音需求量为 $Lap(\Delta f/p\varepsilon_2)$.例如图 8(a)中的查询敏感性 $\Delta f=1$,图 8(b)中 3 个较长范围的桶,由于合并操作使得查询敏感性变成了 $\Delta f=1/3$ 和 $\Delta f=1/2$.

在构建 V-优化直方图时,如何选择合并桶的边界以及如何避免合并操作所导致的隐私泄露是问题的关键.StructureFirst 采用指数机制解决了上述两种问题,如式(10)所示.

$$Pr(r_j=q)\propto\exp\left(-\frac{\varepsilon_1 SSE(q,j,r_{j+1})}{2(n-1)(2F+1)}\right) \tag{10}$$

其中,ε_1 为分得的隐私预算,F 为原始桶计数的上界,$SSE(q,j,r_{j+1})$表示选择边界 q 的打分函数.

StructureFirst 对原始直方图进行有损压缩而得到了 V-优化直方图,因此该方法的误差同样是由重构误差与噪音误差构成.虽然该方法能够精确地

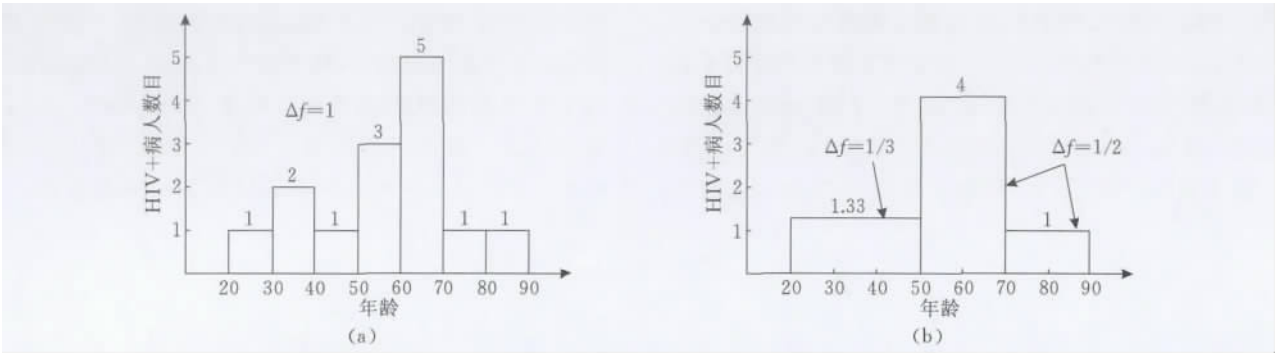

图 8 发布前直方图重组

响应较长范围的计数查询，然而该方法的缺陷也比较明显：(1) 只是采用简单的启发式规则确定合并桶的个数 k，而这种方法没有顾及到重构误差和噪音误差之间的平衡；(2) 对于一些桶个数比较多的直方图，该方法效率非常低.

P-HPartition[31] 结合自适应的层次聚类技术弥补了 StructureFirst 的不足. 该方法结合贪婪二等分策略，自适应地对原始的 n 个桶自顶向下进行分割，并利用指数机制找出每次的二分分割点. 设 k 为最终的分割得到的桶数，k 个桶对应着 k 个聚簇 $C_k=\{C_1^k,C_2^k,\cdots,C_k^k\}$. 该方法的误差也是由重构误差和噪音误差组成，如式(11)所示.

$$Error_{(C_k)}=RE_{C_k}+\frac{k}{\varepsilon} \tag{11}$$

$$RE_{C_k}=\sum_{i=1}^{n}\sum_{H_j\in C_i^k}|H_j-\overline{C_i^k}|,\ 1\leq j\leq n$$

$$\overline{C_i^k}=\sum_{H_j\in C_i^k}\frac{H_j}{|C_i^k|} \tag{12}$$

其中，k/ε 为噪音误差；RE_{C_k} 表示重构误差，如式(12)所示，$\overline{C_i^k}$ 表示聚簇 C_i^k 的均值.

从式(11)可以看出，P-HPartition 采用层次聚类找到最佳的合并桶个数 k 来平衡重构误差与噪音误差，其效率优于 StructureFirst 方法.

综上，第(2)类方法主要是转换直方图的组织形式来发布直方图，而总体桶的实际数量并没有减少，P-HPartition 所发现的 k 个聚簇蕴含着原始直方图的 n 个桶. P-HPartition 与 StructureFirst 均未考虑离群点(Outlier)对直方图发布的影响，例如图 8(a)中，桶[40,50]的计数为 10^6，则在划分簇时，离群点会导致非常高的全局敏感性. 此外，上述两种方法均未考虑原始 n 个桶计数的顺序，而计数的顺序直接影响到分割效果. 例如，$H=\{1,2,1,3,5,1,1\}$，如果对 H 排序后$\{1,1,1,1,2,3,5\}$再分割，最终的误差会更小. 虽然文献[74]提出了排序-分割方法 GS，但是该方法只是简单的排序后等宽分割，该分割方法可能导致大的误差. 因此，基于聚类重新划分直方图的方法下，如何发布携带离群点的直方图以及如何设计自适应的排序-分割方法是未来的重要研究方向.

(3) 采用傅里叶变换发布直方图. 该类方法是利用基于有损压缩的离散傅里叶转换(Discrete Fourier Transform，DFT)来发布直方图.

FPA[32] 是该类方法的典型代表. 给定一个直方图 $H=\{H_1,\cdots,H_n\}$，该方法首先对 H 实施 DFT 操作，即 $F=\mathrm{DFT}(H)$；第 2 步，为了避免高频度桶对发布质量的影响，该方法从 F 中选出 k 个系数作为 F^k，即是剔除 $n-k$ 个其它的系数①. 接着对 F^k 中的 k 个系数添加拉普拉斯噪音，使 F^k 变成 $\widetilde{F}^k$. 最后，在 $\widetilde{F}^k$ 中补充 $n-k$ 个 0，即 $\widetilde{F}=\langle\widetilde{F}^k,0_{n-k},\cdots,0_n\rangle$，再执行 DFT 的逆操作 IDFT，即是 $\widetilde{H}=\mathrm{IDFT}(\widetilde{F})$ 操作.

FPA 的误差也是由噪音误差$(2k^2/\varepsilon^2)$与重组误差$\left(\sum_{i=k+1}^{n}|F_i|^2\right)$组成. 虽然该方法对直方图进行了相应压缩，然而如何挑选 k 至关重要，较大的 k 导致噪音误差增加，较小 k 的导致重组误差增加. 针对该问题，文献[31]基于指数机制设计了一种自适应的 k 挑选方法 EFPA，该方法根据 FPA 的发布误差设计了一种更有效的打分函数 $u(H,k)$，如式(13)所示. 将 $u(H,k)$ 代入式(6)即可获得 k 的挑选概率.

$$u(H,k)=\sqrt{\sum_{i=k+1}^{n}|F_i|^2}+\frac{\sqrt{2}k}{\varepsilon} \tag{13}$$

由式(11)与式(13)可知，第(2)类与第(3)类方法均是通过调整相应参数来均衡重构误差与噪音误差. 而如何选择均衡参数是个大的挑战. 虽然，聚类重复划分方法能够找出比较合适的均衡参数，但是原始直方图的桶个数未得到压缩；通过傅里叶变换

① 注意剔除高频度桶的原因是为了缩减查询敏感性，参阅文献[32].

的采样技术虽然能够压缩原始直方图，但是找出的均衡参数未必最佳，两类方法各有优缺点.

以上阐述了策略 2 下的直方图发布方法. 表 4 介绍了上述 7 种方法的性能对比分析，其中 k 表示转换后直方图桶的个数；ε 的分配、计算开销与实际精度与表 3 分类级别相同.

表 4　策略 2 下直方图发布技术的对比分析

方法名称	转换技术	主要优点	主要缺点	计算开销	ε 的分配	实际精度	时间复杂度
Boost2[26]	层次树变换	支持单位长度与较长范围的查询	仅适用一维等宽直方图	中；额外的后置处理开销	合理；统一应用所有 ε	一般；主要取决于桶的个数	$O(n+\log n)$
Privelet[23]		支持较长范围计数查询，精度较高	实际的可用性比较差	中；额外的傅里叶逆操作		差；主要取决于 k 值选择	$O(n+k)$
StructureFirst[27]	聚类变换	支持较长范围计数查询，查询精度较高	无法均衡重构与噪音误差，无法处理离群点	高；额外直方图重构开销	比较合理；用规划策略找出最佳 ε 配置	一般；主要取决于 V-优化方法	$O(kn^2)$
GS[74]		支持较长范围计数查询	仅采用等宽的分割方法	中；采取的采样技术	比较合理；ε 被均分成两等份	好；取决于采样粒度	$O(n^2)$
P-HPartition[31]		支持较长范围计数查询，效率高	无法处理离群点，仅适用一维直方图	高；额外直方图聚类开销	比较合理；ε 被均分成两等份	好；主要取决于 k-均值聚类	$O(n^2)$
EFPA[31]	傅里叶变换	支持较长范围计数查询，精度高	仅适用一维直方图转换，扩展性差	中；采取了采样技术		好；取决于 k 的自适应选择	$O(n\log n)$
FPA[32]		支持单位长度的范围查询	仅适用一维直方图转换，查询精度低	高；额外的傅里叶变换	合理；统一应用所有 ε	差；主要取决于 k 的挑选	$O(n\log n)$

总体看来，上述 7 种方法均采用转换技术对原始直方图进行重新组织来响应长范围查询、或者提高查询精度. 不同的转换技术有着不同的特点：树变换与傅里叶变换虽然能够支持长范围查询，但是只支持一维直方图的发布，实际的可用性比较差；聚类变换能够启发式地发布直方图，但是通常也仅仅支持一维直方图，并且现有的方法没有考虑到离群点和计数顺序问题. 因此，在发布策略 2 下，基于上述转换技术，如何支持多维直方图的发布以及支持携带离群点的直方图发布是未来的研究方向；此外，如何均衡重构误差与噪音误差也是未来的研究方向.

5.2　基于差分隐私的划分发布方法

该类方法通常基于发布策略 2，考虑如何设计支持数据划分的索引结构，并依据索引结构发布隐私数据. 常用的索引划分结构分为基于树结构与基于网格(Grid)结构的划分. 而这两种划分均要考虑是否在原始的基础数据(Underlying Data)上划分，如果是在基础数据上进行的划分，则称为数据依赖的划分，该类划分可能会使得划分结构自身泄露数据隐私；如果是在查询空间上的划分，而没有涉及到基础数据，则称之为数据独立的划分.

5.2.1　基于树结构划分

(1) 数据独立的树划分

Quad-Post[33] 采用完全四分树对二维几何查询空间进行自顶向下划分[37]. 划分查询空间时，完全四分树需要满足：(1) 所有的叶子-根路径具有相同长度；(2) 所有中间结点具有相同的扇出.

由于四分树自身的结构不会对基础数据造成隐私泄露，只针对叶子节点中的计数值添加拉普拉斯噪音即可. 为了提高数据的发布精度，Quad-Post 同时考虑了两方面因素：(1) 隐私预算 ε 的合理分配；(2) 如何最大化查询精度.

Quad-Post 基于优化策略采用均匀分配与几何分配技术来合理分配隐私预算. 在采用上述两种分配策略时，该方法充分利用了差分隐私的并行组合与序列组合性质. 叶子-根路径上的预算分配符合序列组合性，每条路径上的隐私预算之和为 ε，即 $\sum_{i=1}^{h}\varepsilon_i=\varepsilon$，其中 h 为树的高度，ε_i 表示每一层所分得的隐私预算；由于每一层上的结点相互独立，预算分配符合并行组合性.

为了最大化查询精度，Quad-Post 采用最小二乘法无偏估计对最终噪音响应结果进行了后置处理. 在响应查询时，该方法存在着噪音误差与均匀假设误差①. 该方法的优点在于能够合理地分配隐私预算，噪音误差较低；而缺点在于仅适应于二维空间数据，均匀假设误差比较高.

(2) 数据依赖的树划分

该类划分通常存在树自身的结构会披露基础数据隐私的问题. kd-standard[33] 结合 kd-树对基础空间数据进行了划分. 在发布 kd-树时，如何根据数据

① 一般假设数据在均匀分布的情况下响应查询，而数据实际分布与均匀分布存在偏差，该偏差带来的误差为均匀假设误差，参阅文献[33].

空间的中值数确定分割线是划分的关键.若不采用差分隐私保护此分割过程,中值数可能会被披露.例如在响应图 9 的 Q 查询时,分割线 l 披露了中值数 a 的真实值.*kd*-standard 分别采用指数机制与噪音均值机制[36]确定中值数,而这两种保护机制中相对噪音误差最小的是指数机制,操作如下:

假设 $C=\{x_1,x_2,\cdots,x_n\}$ 为区域 $[a,b]$ 中一个升序排列的集合,x_m 为 C 的实际中值数.任取一个数值 $x(x\in[a,b])$,则 x 被选取的概率如式(14)所示.

$$Pr(A(C)=x)\propto\exp\left(-\frac{\varepsilon}{2}|rank(x)-rank(x_m)|\right) \tag{14}$$

其中 $rank(x)$ 函数表示 x 在集合 C 中的排名,A 为选择算法.因此使用指数机制即可以挑选出与 x_m 最接近的数值,也可以保护 x_m 的隐私.

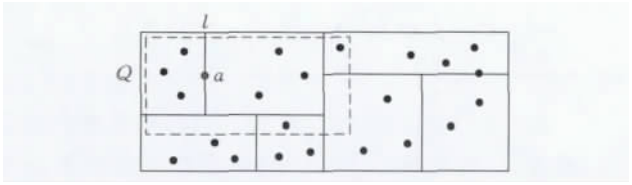

图 9 *kd*-树索引结构

值得注意的是 *kd*-standard 中的隐私预算也被分成了两部分,一部分用来确定中值数,剩余部分用来为每个分割空间中的数据计数添加噪音.

此外,文献[36]基于*kd*-树和多方安全计算提出了一种记录匹配方法 *kd*-noisemean,该方法采用近似噪音均值替代中值数来分割数据空间.然而该方法的通信代价与噪音误差比较高.

综上,Quad-Post、*kd*-noisemean 以及 *kd*-standard 虽然在理论上能够发布相应的索引结构,然而 3 种方法却存在着共同缺陷:在自顶向下分割数据空间时,如何确定分割停止条件至关重要.分割过细会引入过多的噪音数据;反之范围计数查询的响应精度会非常低.因此,如何设计分割条件(例如,基于区域密度的分割条件)是未来的一个研究方向.

而基于数据依赖的树划分方法中,DiffPart[34]最具有代表性,该方法采用自顶向下的方式随机地分割基于泛化技术①的分类树(Taxonomy Tree)[37]来发布集值型数据.4 个关键性因素决定着 DiffPart 分割的性能与效率:(1)如何防止树结构本身泄露数据隐私;(2)如何确定非叶子结点的分割条件;(3)如何分配隐私预算;(4)如何控制空结点(∅)的个数.

该方法采用拉普拉斯机制防止树结构泄露隐私信息.例如在图 10(b)中,非叶子结点 $V1$ 实际包含 4 条 10(a)中的记录,如果直接分割,则计数值 4 就会被泄露.因此,分割某个非叶子结点时,采用拉普拉斯机制扰动该结点的真实计数.例如结点 $V1$ 的噪音计数为 $N_{V1}=4+Lap(6/\varepsilon)$.

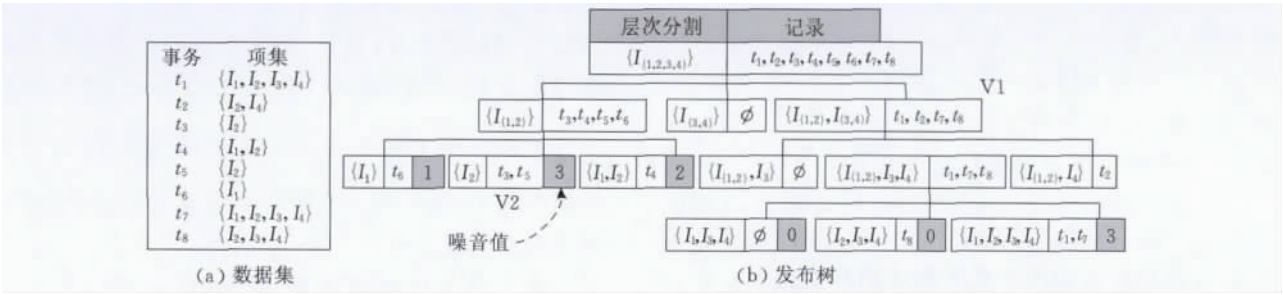

图 10 分类树结构

噪音标准差②作为分割非叶子结点与发布叶子结点的条件,对于非叶子结点 V_i,若其噪音计数 $N_{Vi}\geq\frac{\sqrt{2}C_1\times h(V_i)}{\varepsilon'}$,则分割该结点,其中 C_1 为常数,$h(V_i)$ 为 V_i 所在的层数,ε' 为 V_i 所分得的隐私预算.

为了合理地分配隐私预算,该方法提出了一种自适应分配策略.给定预算 ε,先一分为二,其中 $\varepsilon/2$ 引导非叶子结点分割,$\varepsilon/2$ 用来叶子结点计数信息的发布.每条根-叶子路径上分割剩余的预算均被追加到叶子结点中去.

此外,每层分割会随机生成大量的空结点,即真实计数为 0 的结点,而这些空结点不但消耗隐私预算,还会影响最终的发布精度.为了限制产生大量的空结点,该方法借鉴独立的布尔测试和二项分布,生成 k 个空结点.设 n 为非叶子结点 s_i 生成的总空结点数,从 n 中选择 k 个空结点可看作一个二项分布 $B(n,p)$,则其分布函数如式(15)表示.

$$Pr(x)=\begin{cases}0, & \forall x<\lambda\\ 1-\exp(\varepsilon'\lambda-\varepsilon'x), & \forall x\geq\lambda\end{cases} \tag{15}$$

① 泛化是对数据进行更概括和更抽象的描述,参见文献[12].
② 噪音标准差来自于拉普拉斯分布,参见式(3).

其中，λ 表示选择阈值 $\lambda=\sqrt{2}C_2/\varepsilon'$，$C_2$ 表示常数，ε' 表示选择操作所分得的隐私预算.

尽管 DiffPart 较好地分割集值型数据，并支持 top-k 频繁模式挖掘，然而该方法却存在两点不足：(1) 仅支持计数查询；(2) 在泛化集值数据时，没有考虑不同项之间的语义关联，或者相似关系，而是随机泛化，进而导致发布数据的可用性较低.

除此之外，Hybrid-Bus[35] 借鉴前缀树与分类树发布轨迹数据，并且利用前缀树本身所蕴含的固有约束，设计了一种一致性约束推理策略来增强发布精度. 但是该方法却忽视了轨迹自身携带的时间戳，导致发布轨迹数据的可用性较低；文献[25]同样采用树划分方式，提出了一种序列数据发布方法 n-gram. 与文献[34-35]不同，该方法通过抽取所有变长的 n-gram，并结合前缀树索引发布序列数据. 同时该方法采用马尔科夫假设(Markov Assumption)自适应地分配隐私预算. 另外，文献[71]也提出了一种支持多维数据的树分割方法 DP-tree，该方法利用嵌入树(Nested Tree)索引多维数据来支持范围计数查询，但是该方法易受到树的扇出影响.

以上内容介绍的方法，均是采用迭代分割索引树的方式发布隐私数据. 数据依赖的与数据独立的树划分存在各自的优缺点. 虽然数据依赖的划分方法不存在均匀假设误差，但是如何防止树自身泄露隐私以及如何分配隐私预算是非常大的挑战. 虽然数据独立的划分方法不会导致树结构自身泄露隐私，而如何有效控制发布误差是个非常大的挑战.

5.2.2　基于网格结构划分

在均衡噪音误差与均匀假设误差时，划分粒度的选择至关重要. 而基于树划分的方法均没涉及如何设置划分粒度的大小来平衡这两种误差.

UG[38] 对二维空间数据均匀地划分成 $m\times m$ 个等宽格单元，结合划分粒度 m 为每个单元添加拉普拉斯噪音. 如图 11(a)给出一个 3×3 的格划分，单元格 $y7$ 噪音计数为 $153+n7$. 给定一个查询框 Q，响应 Q 的误差由两部分构成，如式(16)所示.

$$Error(Q)=\frac{\sqrt{2r}m}{\varepsilon}+\frac{\sqrt{r}N}{mc_0} \tag{16}$$

其中，N 表示数据集的大小，r 表示查询框 Q 面积与 $m\times m$ 之间的比率，c_0 为一个常数.

$\frac{\sqrt{2r}m}{\varepsilon}$ 表示噪音误差，由绝对误差度量，$\frac{\sqrt{r}N}{mc_0}$ 为均匀假设误差. 为了均衡这两种误差，UG 设置划分粒度 $m=\sqrt{\frac{N\varepsilon}{C}}$，其中 $c=\sqrt{2}c_0$.

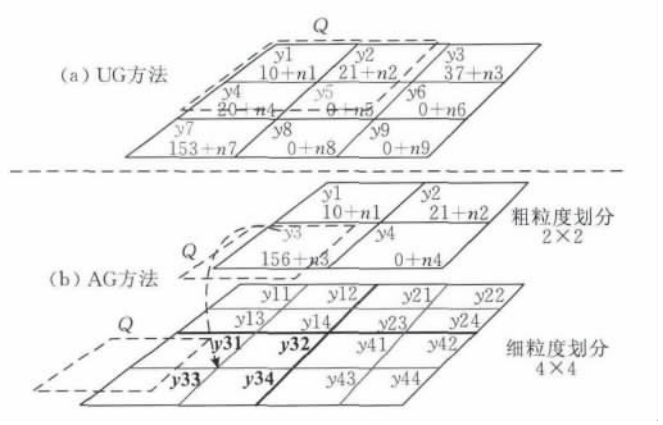

图 11　网格结构划分

虽然 UG 能够比较合理地设定划分粒度 m，但是却没有考虑数据分布的密度和稀疏性. 若某个单元过于稀疏，甚至计数为零，如图 11(a)中的 $y5$ 所示，会导致过大的噪音误差；反之一个单元过于密集，该单元划分不够彻底，如图 11(a)中的 $y7$ 所示，会导致比较大的均匀假设误差.

AG[38] 给出了一种自适应划分策略来避免单元过于密集与过于稀疏问题. 以 $\alpha\cdot\varepsilon(0<\alpha<1)$ 大小的预算给出一个 $m_1\times m_1$ 的粗粒度划分，针对每个粗粒度单元，以 $(1-\alpha)\varepsilon$ 大小的预算自适应地分割成 $m_2\times m_2$ 个细粒度单元. 如图 11(b)所示，$y3$ 被划分成 2×2 个单元. 若一个查询框 Q 穿过第一层划分的某个粗粒度单元，则在第二层划分中最多 $(m_2)^2/4$ 个单元落入 Q 中. Q 的查询误差同样由噪音误差和均匀假设误差组成，如式(17)所示.

$$Error(Q)=\frac{\sqrt{2}}{(1-\alpha)\varepsilon}\cdot\sqrt{\frac{(m_2)^2}{4}}+\frac{NC}{m_2c_0} \tag{17}$$

其中，NC 表示第一层被穿过格单元的噪音计数(如图 11(b)中 $y3$ 的噪音计数为 $156+n3$).

为了自适应地均衡 Q 的查询误差，AG 方法设置 $m_2=\sqrt{NC(1-\alpha)\varepsilon/\sqrt{2}c_0/2}$.

综上，UG 和 AG 方法的优点在于能够比较合理地给出空间数据的格划分粒度，并且能够均衡噪音误差与均匀假设误差. 而这两种算法存在各自的不足：UG 方法由于没有考虑到数据本身的稀疏性，可能会导致过大的均匀假设误差；AG 方法虽然能够根据数据稀疏性自适应地设置空间数据的划分粒度，然而该方法却没有给出相应的启发式规则来区分数据稠密(Dense)与稀疏(Sparse)之间的边界. 另外，UG 和 AG 方法仅适用于二维数据

的发布.因此,结合数据的稀疏性,如何利用启发式规则指导自适应地划分网络粒度是未来的研究方向.

以上内容介绍了策略 2 下基于划分的发布方法.表 5 分别对上述方法进行分析对比.其中,表 5 中的 n 表示记录的个数,$|I|$ 表示项集域的大小.总体来看,大多数已有的方法是采用数据依赖的树划分索引结构发布隐私数据,而这些方法大都受到实际的数据维度影响,导致计算开销比较高、实际的可用性比较低.另外,一些基于数据独立的树划分发布方法没有顾及到如何均衡噪音误差与均匀假设误差.基于网格结构划分的发布方法,虽然顾及到了上述两种误差的均衡,但是仅局限于二维空间数据,并且在设计均衡策略方面,没有考虑到如何利用启发式规则来自适应地设置均衡参数.因此,如何设计支持高维的、支持数据依赖的树划分方法是未来的研究方向;如何设计具有启发式的均衡参数设置方法也是未来的研究方向.此外,上述所有基于划分的发布方法,均假设原始数据是静态的,只考虑一个数据快照(Snapshot).而实际应用中,无论是集值数据、序列数据,还是空间数据通常是动态的,因此如何设计支持动态数据的划分发布方法同样是未来的研究方向.

表 5 基于划分发布技术的对比分析

方法名称	划分技术	主要优点	主要缺点	ε 的分配	实际精度	时间复杂度
Quad-Post[33]	树划分	支持数据独立的范围计数查询	无法均衡噪音误差与均匀假设误差	合理;统一应用所有 ε	一般;主要取决于后置处理技术	$O(n\log n)$
kd-strandard[33]		支持数据依赖的范围计数查询,查询精度较高	一般局限于二维以内的数据	比较合理;ε 被划分成两份	一般;取决于如何均衡两份预算	$O(n\log n)$
k-noisemean[36]		支持多维数据依赖的计数查询	通信代价比较高	合理;统一应用所有 ε	差;主要取决于噪音均值与通信代价	$O(n\log n)$
Diffpart[34]		支持多维数据依赖的计数查询,查询精度较高	没有考虑集值型数据中项的语义关联	合理;自适应分配 ε	一般;主要取决于集值型数据的维度	$O(n\|I\|)$
Hybrid-Bus[35]		支持多维数据依赖的计数查询,查询精度高	忽视了轨迹数据的时间属性,实用性差	比较合理;ε 被划分成两份	一般;主要取决于轨迹序列的维度	$O(n\|I\|)$
n-gram[25]		支持多维数据依赖的计数查询,查询精度高	容易受到序列维度影响,一般序列长度为 5	合理;自适应分配 ε	好;主要取决于序列维度、预算策略	$O(n\|I\|)$
DP-tree[71]		支持多维数据依赖的范围计数查询,查询精度高	效率易受到树扇出的影响.	合理;自适应分配 ε	好;取决于后置处理方法、预算策略	$O(\log^3 n)$
UG[38]	网格划分	支持范围计数查询,均衡噪音与均匀假设误差	没有考虑数据分布的稀疏性	合理;统一应用所有 ε	一般;主要取决于划分粒度的选择	$O(n)$
AG[38]		支持数据依赖的范围查询,自适应均衡两种误差	均衡误差时,没有采用启发式方法	比较合理;ε 被划分成两份	好;主要取决于两层划分粒度的选择	$O(n)$

通过以上分析可知,现有的发布技术大多关注直方图技术与划分技术的关系数据发布.目前,针对不同数据类型,研究者提出了不同的发布方法,本文所述工作仅是其中一部分,还有很多未提及的工作,例如,稀疏数据的发布[40-43]、图数据的发布[44-47]、流数据发布[48-49]以及数据立方体发布[50-51]等工作.

虽然已经提出了较多的发布方法,但是所存在的问题依然很多.针对相应的问题,我们指出了未来的一些研究方向,参见各类方法的对比分析.

6 数据分析方法的分类与评估

数据分析的目的在于从数据中抽取或者学习到有价值的模型和规则.模型与规则中的敏感信息可能导致个人隐私泄露,进而使得隐私保护的数据挖掘和机器学习得到广泛关注[55-56].以前一些基于匿名隐私保护模型下的挖掘和学习问题,在差分隐私保护模型下又得到新的探索.

6.1 基于数据挖掘的分析方法

6.1.1 基于差分隐私的模式挖掘技术

频繁模式挖掘是数据分析的主要技术之一,其目的是找出频繁出现在数据集中的模式.然而频繁模式本身的内容以及相应的频度有可能泄露用户隐私信息.基于差分隐私的模式挖掘主要是如何保护模式的频度不被披露.

TF[57]是基于差分隐私保护技术的 top-k 频繁模式挖掘典型代表.该方法的基本思路可以概括为以下两步:(1)从所有长度不小于 l 的频繁项集中选择出 k 个模式;(2)对 k 个模式的频度添加拉普拉斯噪音.

假设 D 为一个事务数据库,含 n 条事务,$|I|$ 表示项集域的大小,隐私预算 ε 被均分为二.TF 的第

2 步相对容易实现，只要对每个模式的频度添加大小为 $Lap(2k/n\varepsilon)$ 的噪音即可. 实现该方法第 1 步的关键是如何在候选集合 C 中挑选出 k 个模式，而候选集合的规模 $|C|\approx|I|^l$. 由于 $|C|$ 通常规模比较大，如果通过枚举所有模式来挑选 k 个模式，计算量非常大. TF 通过截断频率(Truncated Frequency)技术与指数机制从 C 中挑选 k 个模式. 对于任意一个模式 p，其截断频率 $f'(p)=\max(f(p),f_k-\gamma)$，$\gamma$ 为调节参数，f_k 表示 C 中第 k 个最频繁的模式. 因此，TF 每次以概率 $Pr(p)\propto\exp(\varepsilon nf'(p)/4k)$ 从 C 中选择一个模式.

图 12(c)给出了 TF 经过指数机制筛选以及拉普拉斯机制加噪音之后所获得的 $l=2$，top-3 的频繁模式. TF 通过 $f(p)>f_k-\gamma$ 缩减候选集 C 空间. 然而，由于 $\gamma>4kl\ln|I|/\varepsilon n$，随着 k 值以及事务维度的增加，有可能使得 $f_k-\gamma\leq0$，进而使得删减条件 $f(p)>f_k-\gamma$ 失效. 若图 12(a)中设置 $f_k-\gamma=0$，则图 12(b)中至少有 16 个候选集. 在此情况下 TF 的挖掘效率与准确性非常低.

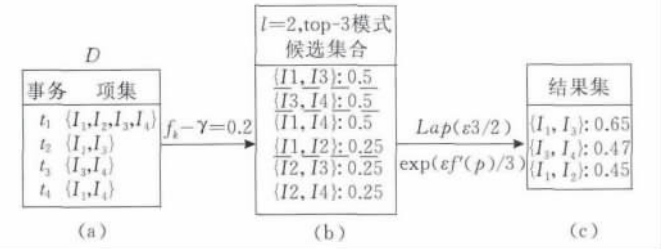

图 12　TF 方法发布流程

为了弥补 TF 的不足，PrivBasis[58] 结合 θ-基和映射技术挖掘 top-k 频繁模式. 该方法首先在数据集 D 中找出所有频度大于等于 θ 的频繁项 F，即把 D 映射到集合 F 中. 然后基于 F 构建所有的 θ-频繁对 P，根据集合 F 和 P 构建 θ-基集合. 最后根据该集合创建所有 θ-频繁项集的候选集合 C，并对集合 C 中的所有项集的频度添加噪音.

关键问题是如何构建 θ-基集合. PrivBasis 借助于极大团(Maximal Clique)思想来生成 θ-基集合. F 和 P 分别作为结点和边生成图 $G(F,P)$，然后找出图 $G(F,P)$ 的所有极大团. 每个极大团可视为一个 θ-基. 图 13 给出了 θ-基集合一个生成实例，$\{I_1,I_3,I_4\}$ 为事务数据集 D 的 θ-基集合，而基于 θ-基集合可以生成 top-k 频繁模式. 例如，以 $\{I_1,I_3,I_4\}$ 为 θ-基集合，可以生成 $l=2$，top-3 模式为 $\{I_1,I_3\}$，$\{I_1,I_4\}$ 与 $\{I_3,I_4\}$.

此外，文献[59]依据长事务记录会导致高查询敏感性的缺陷，提出了一种基于事务截断技术的贪婪方法 SmartTrunc，该方法利用阈值和动态权重频率对每条记录进行局部转换(Local Transformation)，通过截断长记录来降低查询敏感性，进而提高模式的可用性. 然而，该方法仅对分布比较极端的数据集(例如，数据集包含大量的短事务记录)有效.

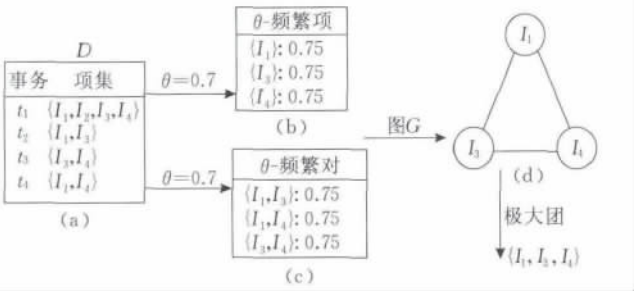

图 13　θ-基集合生成过程

综上，上述方法存在各自的不足，TF 处理较大的 k 值或者较长的 l 时性能与效率比较差. 虽然 PrivBasis 的性能和效率优于 TF，但是该方法存在难以兼顾隐私保护与模式可用性的不足；存在破坏原始 Top-k 模式频度特征(例如，频度大小按降序排列等)使得扰动后的模式频度发生较大偏差的问题. SmartTrunc 扩展性比较差. 此外，TF 与 PrivBasis 未考虑记录本身长度带来的影响. 因此，如何设计同时兼顾 Top-k 模式的隐私性与可用性的方法是未来的研究方向.

6.1.2　基于差分隐私的分类技术

分类技术在数据预测分析中起着关键作用，该技术的目的是找出描述和区分数据类或概念的模型，分类模型的典型代表是决策树(Decision Tree)，该结构是一种树形的分类模型，树内结点表示在某个属性上的测试，而叶结点表示一个类. 结合差分隐私与决策树的代表方法分别是 SuLQ-based ID3[60]、DiffP-C4.5[61] 以及 DiffGen[62]. 这 3 种方法在生成分类器时类似于 ID3[63]，主要是考虑决策树各个结点上分割属性的选择问题.

上述 3 种方法均采用了信息增益(Information Gain)选择分割属性，并递归地构建决策树. 给定训练数据集 D，包含 K 个类 C_k. 设属性 A 有 n 个不同的值，则 A 的取值将 D 划分成 n 个子集 $D_1,\cdots,D_n$. 设 D_{ik} 表示子集 D_i 中属于类 C_k 的记录集合. 则属性 A 对 D 的信息增益 $InfoGain(D,A)$ 可表示为

$$InfoGain(D,A)=H(D)-H(D|A),$$

其中，$H(D)$ 表示数据集 D 的经验熵，如式(18)所示. 而 $H(D|A)$ 则表示属性 A 对 D 的经验条件熵，

如式(19)所示.

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|} \tag{18}$$

$$H(D|A)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|} \tag{19}$$

基于交互式框架的 SuLQ-based ID3 直接采用拉普拉斯机制对式(18)和(19)中的各个计数(例如 $|C_k|$,$|C_{ik}|$,$|D|$ 以及 $|D_i|$)添加噪音,然后计算属性 A 对 D 的 $InfoGain$ 值. 然而,若 D 的属性个数比较多时,该方法不得不把隐私预算分成若干份来计算各个属性对 D 的 $InfoGain$ 值,这样形成的分类器不但累积噪音大,而且浪费了大量的隐私预算.

针对 SuLQ-based ID3 的缺点,DiffP-C4.5 利用指数机制来挑选分割属性 A,其中 $InfoGain(D,A)$ 作为属性 A 的打分函数来计算其被选中的概率值. 该方法不必分割所分得的隐私预算,而是全部利用这些预算选择最好的分割属性. 然而,该方法是基于交互式查询接口构建的决策树,一旦大量的分析者提交查询时,该方法的分类精度就会降低,即该方法只能支持少量的分析查询.

SuLQ-based ID3 和 DiffP-C4.5 均采用交互式的分类方法. 不同于上述两种方法,DiffGen 结合指数机制与 InfoGain 来确定分割属性,借助于分类树自顶向下地把数据集 D 中所有记录划分到叶子结点中去,然后对叶子结点中的计数值添加拉普拉斯噪音.

虽然 DiffGen 的分类精度无论从理论和实际应用角度均高于 SuLQ-based ID3 和 DiffP-C4.5,但是该方法仍然存在着不足. 由于每一个分类属性对应一个分类树,当 D 中的分类属性的维度非常大时,该方法不得不维护大量的分类树. 大量的分类树会导致基于指数机制的选择方法效率很低,并且有可能耗尽隐私预算. 因此,如何对具有高纬度分类属性的数据集进行分类以及如何设计有效的隐私预算分配策略是未来的研究方向.

6.1.3 基于差分隐私的聚类技术

聚类同样是数据分析的主要技术,它是把数据对象划分成多个簇的过程,而在聚类过程中数据隐私可能被泄露,例如,均值(Means)、中心点(Center)与中值(Median)等. 文献[64]结合采样与聚集技术提出了一种满足差分隐私的 k-均值聚簇中心发布方法 Pk-means,该方法给出了聚类敏感性的度量方法以及聚类误差的下界. 此外,在 k-均值聚类过程中,隐私预算 ε 的设置也非常关键,文献[65]提出了两种分配方法:(1)迭代次数 n 已知情况下,每一轮聚类预算为 ε/n;(2)迭代次数不知道的情况下,每次所分配的预算为上次剩余预算的一半.

虽然上述两种聚类方法均满足 ε-差分隐私,但实际应用性比较差. 当数据集合很大时,k 值的选择是 NP 问题,选择 k 值操作有可能泄露真实的数据点,并且每次选择均要消耗隐私预算. 因此,如何利用指数机制挑选上述两种方法的 k 值是未来的研究方向.

6.2 基于机器学习的分析方法

该类分析的目的是采用统计学习方法对已知敏感数据进行分类,或者对未知新敏感数据进行预测和分析,其典型代表是支持向量机(Support Vector Machines)分类、回归分析,例如线性回归(Linear Regression)、逻辑斯谛回归(Logistic Regression)等. 下面进行分类阐述.

6.2.1 基于差分隐私的回归分析

回归分析是机器学习中常用的数据分类分析方法,该分析是确定输入数据集中两种或两种以上属性间相互依赖的定量关系. 常用的回归分析方法包括逻辑斯谛回归与线性回归. 例如,图 14(a)中的线性回归,其目的是要找出"医疗费用"与"年龄"的线性关系,即是找出该线性模型,使得病人分布点到该直线的平方误差和最小. 而逻辑斯谛回归则是比较事件的发生概率与不发生概率的大小来进行分类. 图 14(b)显示了"HIV+"与"年龄"逻辑回归关系,利用对数几率(Log Odds)表示了糖尿病患者的分布情况.

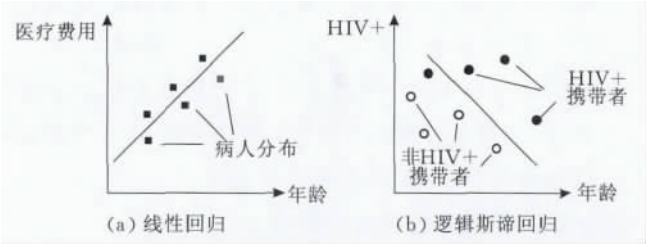

图 14 回归分析示例

下面进行阐述基于差分隐私保护的逻辑斯谛回归与线性回归.

回归分析通常有两类函数:一是预测函数;二是目标函数,或者称为风险函数. 无论是线性回归或逻辑斯谛回归分析,通常归结为目标函数的最优化问题. 下面是该问题的形式化表示.

给定训练数据集 $D=\{t_1,t_2,\cdots,t_n\}$,有 n 个元组,每个元组包含 $d+1$ 个属性 $X_1,\cdots,X_d,Y$,其中 $X_i\in R_n$,$Y\in\{0,1\}$,或者 $[-1,1]$. 元组 $t_i=(x_i,$

y_i)，其中 $\boldsymbol{x}_i$ 表示 $(x_1, x_2, \cdots, x_d)$ 向量. 假设 $\rho(x_i)$ 表示预测函数，该函数通常由向量 x_i 与其相应的权重向量 w^* 的参数化形式表示. 式(20)和式(21)分别表示 D 上的线性回归预测函数和逻辑斯谛回归预测函数.

$$\rho(x_i)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}^* \tag{20}$$

$$\rho(y_i=1\,|\,\boldsymbol{x}_i)=\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}^*)/(1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}^*)) \tag{21}$$

从式(20)和式(21)可以看出，只要得到权重向量 w^*，即可以对元组 t_i 进行分类. 而向量 w^* 通常表示为下列公式，其中，$f(t_i, w)$ 表示目标函数.

$$\boldsymbol{w}^*=\arg\min_{w}\sum_{i=1}^{n}f(t_i,w) \tag{22}$$

线性回归与逻辑斯谛回归下的目标函数分别由式(23)与式(24)表示.

$$f(t_i,w)=(y_i-\boldsymbol{x}_i^{\mathrm{T}}w)^2 \tag{23}$$

$$f(t_i,w)=\log(1+\exp(\boldsymbol{x}_i^{\mathrm{T}}w))-y_i\boldsymbol{x}_i^{\mathrm{T}}w \tag{24}$$

然而，直接发布 w^* 值会泄露预测函数 $\rho(x_i)$ 与 D 中的数据信息. 例如图 14(a)中，如果已知 w^* 值与红点的 x_i 值，即可推测出该点被分类到直线下方.

文献[66]借鉴拉普拉斯机制与逻辑斯谛回归提出了一种 w^* 计算方法 LPLog，该方法利用式(22)求出 w^* 之后，添加拉普拉斯噪音，利用含噪音的 $\widetilde{F}$ 计算 $\rho(x_i)$ 的值. 然而，由于回归分析的输入 x_i 与输出 $\rho(x_i)$ 存在紧密的关联性，使得计算 w^* 敏感性的代价非常高，导致预测精度较低.

不同于 LPLog，文献[67-68]提出了一种直接扰动目标函数方法 ObjectivePerb，该方法对 D 中 n 个元组目标函数的均值添加噪音，如式(25)所示，其中 b 是来自拉普拉斯分布的噪音向量. 在扰动目标函数 $\bar{f}_D(w)$ 的基础上，利用式(26)求出 w^*，其中 Δ 为一个常数.

$$\bar{f}_D(\boldsymbol{w})=\frac{1}{n}\sum_{i=1}^{n}f(t_i,w)+\frac{1}{n}\boldsymbol{b}^{\mathrm{T}}w \tag{25}$$

$$\boldsymbol{w}^*=\arg\min\bar{f}_D(\boldsymbol{w})+\frac{1}{2}\Delta\,\|w\|^2 \tag{26}$$

然而，ObjectivePerb 在添加噪音时，噪音量的大小仍然由 w^* 的敏感性决定，而不是由目标函数 $f_D(w)$ 自身的敏感性决定，因此，该方法在计算 w^* 的敏感性时，代价也非常大. 同时，该方法只适用于具有较强约束条件、凸函数特性以及双可微特征的目标函数，通用性较差. 例如，该方法不适合标准逻辑斯谛回归，即是当 D 的分类属性 y_i 属于布尔类型时($y_i\in\{0,1\}$)，ObjectivePerb 方法失效.

针对上述方法的不足，文献[16]提出了一种函数机制 FM(Functional Mechanism)分别实现了差分隐私保护下的线性与逻辑斯谛回归分析. 由式(23)可知，$f_D(w)=\sum f(t_i,w)$. FM 机制首先通过噪音扰动 $f_D(w)$ 得到扰动的目标函数 $\bar{f}_D(w)$，然后基于式(22)求出 w^*.

在加噪过程中，FM 机制不是通过 w^* 的敏感性控制噪音量，而是通过 $f_D(w)$ 本身的敏感性. 假设 $w=(w_1, w_2, \cdots, w_d)$ 为一个权重向量. 文献[48]首先把目标函数 $f_D(w)$ 表示成多项式形式，如式(27)所示.

$$f_D(\boldsymbol{w})=\sum_{t_i\in D}f(t_i,w)=\sum_{j=1}^{J}\sum_{\phi\in\Phi_j}\sum_{t_i\in D}\lambda_{\phi t_i}\phi(\boldsymbol{w}) \tag{27}$$

其中，$\lambda_{\phi t_i}\in R$ 表示变量 $\phi(w)$ 的系数，Φ_j 表示 w_1，w_2，$\cdots$，w_d 所有乘积集合，$\phi(w)\in\Phi_j$.

同理，给定 D 的近邻 D'，D' 上的目标函数 $f_{D'}(w)$ 也可以表示如式(27)的多项式形式. 然后根据式(2)求出 $f_D(w)$ 的敏感性，如式(28)所示.

$$\|f_D(\boldsymbol{w})-f_{D'}(\boldsymbol{w})\|_1\le 2\max_{t}\sum_{j=1}^{J}\sum_{\phi\in\Phi_j}\|\lambda_{\phi t}\|_1 \tag{28}$$

利用 $f_D(w)$ 的敏感性以及预算 ε，为多项式的每个系数 $\lambda_{\phi t_i}$ 添加拉普拉斯噪音，然后可以得到扰动目标函数 $\bar{f}_D(w)$ 以及 w^*. 与文献[63-64]中的方法相比，FM 机制绕过了对 w^* 敏感性的计算.

假设 D 的分类属性 $y_i\in[-1, 1]$，利用式(22)、式(27)与式(28)可求出 w^*，进而可以求出满足差分隐私的线性回归分类模型. 由式(20)可知，线性回归的目标函数展开式是多项式，与 FM 机制恰好吻合. 然而，逻辑斯谛回归的目标函数(如式(21)所示)并不能表示成多项式的形式. 基于此类情况，文献[16]借鉴截断泰勒展开式(Truncated Taylor Expansions)技术，提出了一种标准逻辑斯谛回归①目标函数近似多项式表示方法，如式(29)所示.

$$\hat{f}_D(\boldsymbol{w})=\sum_{i=1}^{\pi}\sum_{k=0}^{2}\frac{f^{(k)(0)}}{k!}(\boldsymbol{x}_i^{\mathrm{T}}w)^k-\Big(\sum_{i=1}^{n}y_i\boldsymbol{x}_i^{\mathrm{T}}\Big)w \tag{29}$$

基于式(28)和式(29)，可以计算出 $\hat{f}_D(w)$ 的敏感性，并可由式(22)推导出 w^*.

此外，文献[69]凭借直方图技术，也提出了一种直接绕开 w^* 敏感性分析的逻辑斯谛回归分析方法 DPME，该方法利用拉普拉斯机制在训练集 D 上生成多维的噪音直方图，利用直方图重新合成训练集 $\bar{D}$，然后基于 $\bar{D}$ 计算 w^*. 然而，DPME 方法只适应

① 注意标准逻辑斯谛回归是指 D 的分类属性 y_i 属于布尔类型时的回归，参加文献[54].

于维度比较小的训练集，一旦训练集的维度比较大时，该方法的回归预测精度明显降低.

综上，3 种回归分析方式均存在各自的不足. 基于拉普拉斯机制的回归分析方法，其回归分类精度比较低，噪音误差比较高；基于扰动机制的回归分析方法仅适用于特定的目标函数，存在很大的局限性；虽然函数机制的性能和效率优于前两种方式，并且能够弥补前两种方式存在的缺陷，然而，该机制自身的缺陷也非常明显. 该机制目前仅适用于线性表示的目标函数，而对于实际应用中复杂的目标函数，例如 Cox 回归(Cox Regression)，该机制的回归分析效果很差. 因此，如何设计处理复杂目标函数的差分隐私回归分析是未来的研究方向.

6.2.2 基于差分隐私的支持向量机

支持向量机也是常用的数据分类技术，该技术通常处理输入空间为非线性的分类问题. 利用核技巧(Kernel Trick)中的非线性变换将输入空间映射到一个特征空间. 然后在特征空间中，利用线性问题的解法求解支持向量. 假设非线性分类问题被转换后的线性问题可以形式化为下列公式. 给定一个训练集 $D=\{(x_i,y_i)\}_{i=1}^{n}$，$x_i\in \mathrm{R}^d$，$y_i\in\{+1,-1\}$. 同样假设 $w=(w_1,w_2,\cdots,w_d)$ 为一个权重向量.

$$f_D(w)=\frac{1}{2}\|(w)\|^2+\frac{C}{n}\sum_{i=1}^{n}l(y_i,\rho_w(x_i)) \quad (30)$$

$$\rho_w(x_i)=\sum_{i=1}^{m}w_i y_i k(\cdot,x_i) \quad (31)$$

其中，$f_D(w)$ 为目标函数，$\rho_w(x_i)$ 表示预测函数，$k(\cdot,x_i)$ 表示核函数，$l(\cdot)$ 表示损失函数.

因此，类似于回归分析的求解，求解支持向量机问题转换为目标函数的凸二次规划问题，即在约束条件下求最优解.

$$w^*=\arg\min_{w} f_D(w) \quad (32)$$

然而，由于输入空间中的训练集 D 含有敏感性信息，直接发布 w^* 同样会泄露 $f_D(w)$ 与训练集 D 的敏感信息.

文献[66]结合差分隐私保护技术提出了一种支持向量机分类方法 PrivateSVM，该方法利用拉普拉斯噪音扰动法向量 $w(w^*=w+Lap(\lambda))$，使得方法 PrivateSVM 满足 ε-差分隐私. 然而，该方法因 w^* 的敏感性过高会导致较大的噪音量，进而导致比较低的分类精确. 与 PrivateSVM 不同，文献[69]提出了一种对目标函数加噪音的分类方法 ObjectiveSVM，该方法采用拉普拉斯分布 $Lap(b)$ 产生随机噪音 b. 然后把 b 添加到风险函数 $f_D(w)$ 而得到扰动目标函数 $\bar{f}_D(w)$. 根据式(32)即可计算出 w^*.

虽然 ObjectiveSVM 方法的分类精度高于 PrivateSVM 方法，然而该方法的缺陷是其目标函数必须具有特定性质，即是具有凸函数特性以及双可微特征的目标函数.

综上，两种差分隐私下的支持向量机分类技术的缺陷比较明显，基于拉普拉斯机制的向量机分类精度低、噪音大，而基于扰动目标函数的支持向量机只适用于特定的目标函数. 因此，如何设计通用的并且能够适用于多种目标函数的扰动机制是未来的研究方向.

自从差分隐私保护技术出现以后，数据挖掘与机器学习领域出现了许多数据分析研究工作. 针对所提及到的数据分析方法，文中给出了不同技术的特点比较，如表 6 所示. 从表 6 可以看出，不同的分析算法优缺点各有差异，例如，基于 ID3 的分类比基于 SVM 的算法容易实现；回归分析易实现，但隐私预算较高等. 目前，差分隐私下的数据分析依然存在很多问题，例如，回归分析仅适用于比较简单的目标函数；分类分析不太适合大规模数据集，以及增量更新环境下的分类等.

表 6 差分隐私下的数据分析方法比较

分析技术	主要优点	主要缺点	代表方法	典型应用
回归分析	满足差分隐私；简单方便；回归误差小；精度较高	所需隐私预算较多；寻找最优解代价较高；计算开销大；实现较难	LPLog[66]，FM[16]，DPME[69] ObjectivePerb[68]	各种回归操作，如线性回归、逻辑斯谛回归
分类分析	满足差分隐私；适合较小训练集；精度较高；实现简单	难以应对较大训练集；SVM 的系数会引起较大误差；	SuLQ-based ID3[60]，DiffP-C4.5[61]，DiffGen[62]，PrivateSVM[66]，ObjectiveSVM[68]	各种回归操作，如ID3、C4.5分类、支持向量机分类
频繁模式分析	利用模式可以合成原始数据集，且满足差分隐私；	长模式导致噪音增加；误差较大；只适合记录长度较短的数据	TF[57]，SmartTrunc[59] PrivBasis[58]	用户行为模式分析、搜索日志分析、推荐系统等.
聚类分析	满足差分隐私；分析结果直接简单；实现简单	聚类误差下界难定；实现比较难；误差较高；只适合个人隐私保护	Pk-means[65] Pk-median[70]	各种聚类操作，如k-均值聚类

7　查询处理方法的分类与评估

数据库领域中的查询处理分为非交互式和交互式查询. 第 5 节所有的非交互式发布方法均是为了满足用户的查询需求. 例如，直方图发布方法 Boost[26]、NoiseFirst[27]、Privelet[23]、*P*-HPartition[31] 是为了能精确地响应范围计数、计数等查询；基于划分方法 Quad-Post[33]、*kd*-standard[33]、DiffPart[34]、AG[38] 等均是为了响应计数、频繁模式挖掘等查询. 这些支持非交互式查询处理的发布方法在第 5 节已经详细阐述，在支持相应的查询时，这些方法未考虑查询负载(Query workload)之间的线性关联. 而交互式下的查询方法通常考虑了这种关联关系，并且以批量形式处理.

表 7　原始数据集

姓名	地名	HIV+
Alice	*A*	Yes
Bob	*B*	Yes
Carol	*C*	No
…	…	…

表 8　HIV+统计

地名	HIV+计数
A	3200
B	1200
C	4000
…	…

线性查询(Linear Query)的线性方程形式表示为 $q(\boldsymbol{D})=w_1x_1+w_2x_2+\cdots+w_nx_n$，其中 $\boldsymbol{D}=\{x_1,x_2,\cdots,x_n\}$ 表示一个计数向量，例如表 8 中的 HIV+计数向量；w_i 表示每个计数相应的权重. 表 8 为基于表 7 中的属性“地名”而得到的 HIV+统计数据. 假设给出一条线性查询为“查询所有地区的 HIV+总数”，则该查询可以表示为 $q=x_A+x_B+x_C$.

而线性查询的批量处理是指同时提交多个线性查询，即 $Q(\boldsymbol{D})=\{q_1,q_2,\cdots,q_m\}$. 如何在差分隐私保护下响应 Q 是一个很大的挑战. 目前处理批量查询的方法有矩阵机制和低秩机制. 下面我们针对两种批量处理方式进行分类介绍和比较.

7.1　基于矩阵机制的批量查询处理

给定 $Q(\boldsymbol{D})=\{q_1,q_2,\cdots,q_m\}$，则 $Q(\boldsymbol{D})$ 可以表示成 $Q(\boldsymbol{D})=\boldsymbol{WD}$，其中 $\boldsymbol{W}_{m\times n}$ 表示查询负载矩阵(Workload matrix). 直接基于拉普拉斯机制防止 $Q(\boldsymbol{D})$ 隐私泄露的方法：直接为 $\boldsymbol{WD}$ 添加噪音，如式(33)所示.

$$\boldsymbol{A}(\boldsymbol{W},\boldsymbol{D})=\boldsymbol{WD}+(\Delta_W/\varepsilon)^m\tilde{\boldsymbol{b}} \tag{33}$$

其中，$\tilde{b}$ 为 $Lap(1)$ 形成的 m 维列向量.

矩阵机制[15,53,78]利用查询计划(Query Plan)和优化策略对批量查询进行求精处理. 该机制首先基于 $\boldsymbol{W}$ 产生相应的查询计划 $\boldsymbol{A}$，$\boldsymbol{A}$ 为满秩的矩阵. 然后基于 $\boldsymbol{A}$ 请求 $\boldsymbol{D}$ 的噪音结果对 $\boldsymbol{W}$ 的查询结果进行估计. 式(34)为该机制的表示形式.

$$\boldsymbol{A}(\boldsymbol{W},\boldsymbol{D})=\boldsymbol{WD}+(\Delta_A/\varepsilon)^m\boldsymbol{WA}^+\tilde{\boldsymbol{b}} \tag{34}$$

其中，$\boldsymbol{A}^+=(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t$，$\boldsymbol{A}^t$ 为 $\boldsymbol{A}$ 的转置.

由式(34)可以看出，矩阵机制的查询敏感性为 Δ_A，由于 $\Delta_A<\Delta_W$，则该机制所需的噪音量比较少. 此外，该机制考虑了线性查询之间的关联性.

文献[15,53]通过矩阵分解对该机制产生的噪音误差进行优化，而文献[78]却通过奇异值分析给出该噪音误差的最低下界. 虽然矩阵机制从理论角度优于拉普拉斯机制，然而，该机制仅适合小规模的数据集和查询负载. 并且该机制通常产生次优化的查询计划，其返回结果的准确性远不及直接在 $\boldsymbol{D}$ 上添加噪音产生的结果.

7.2　基于低秩机制的批量查询处理

文献[53]结合 $\boldsymbol{W}$ 为低秩矩阵(Low-rank Matrix)的特点，提出了低秩机制来改善矩阵机制的不足. 该机制首先分解 $\boldsymbol{W}_{m\times n}=\boldsymbol{B}_{m\times r}\boldsymbol{L}_{r\times n}$，$r\leq\min(m,n)$，其中 $\boldsymbol{B}$ 表示 $m\times r$ 的列满秩矩阵，$\boldsymbol{L}$ 表示 $r\times n$ 的行满秩矩阵. 通过 $\boldsymbol{L}_{r\times n}$ 确定 $\boldsymbol{W}$ 的查询敏感性 Δ_L. 因此，$Q(\boldsymbol{D})=\boldsymbol{WD}$，可以表示成 $Q(\boldsymbol{D})=\boldsymbol{BLD}$. 与拉普拉斯机制和矩阵机制不同，该机制通过对中间结果 $\boldsymbol{LD}$ 添加噪音完成隐私保护.

$$\boldsymbol{A}(\boldsymbol{W},\boldsymbol{D})=\boldsymbol{B}(\boldsymbol{LD}+Lap(\Delta_L/\varepsilon)^r) \tag{35}$$

为了避免矩阵机制产生次优化结果的弊端，低秩机制基于二次规划提出了一种 $\boldsymbol{W}$ 最优分解策略. 该策略具有线性的收敛速度.

综上，上述的矩阵机制和低秩机制在处理批量查询时均存在各自的不足，矩阵机制虽然理论优于拉普拉斯机制，然而实际应用中易产生次优化结果；低秩机制只考虑了负载矩阵的关联性，而没有顾及数据本身的关联性. 因此，如何从数据本身的实际相关性出发，设计出通用的批量处理机制是未来的一个研究方向.

8　相关应用系统

目前基于差分隐私的数据发布与分析大都着眼于理论与方法，相应的原型系统比较少. 目前仅有微软的 PINQ[19]、美国伯克利大学的 GUPT[52] 以及德克萨斯大学的 Airavat[73] 以及新加坡 ADSC 研究所的 Pioneer[76] 系统. 下面对这 4 种原型应用系统的特点进行阐述.

8.1　基于交互式的数据分析系统

PINQ[19] 是最早结合差分隐私实现交互式任务

分析的原型系统.该系统为分析者提供 API 接口,并且独立地响应聚集查询、join 查询以及聚类分析等任务.然而,由于 PINQ 让不可信任的数据分析者管理隐私预算,这样会导致预算分配的不合理性甚至造成浪费.Pioneer[76]主要是利用查询优化器来避免隐私预算的浪费.系统结合已经响应的查询与当前的查询制定出不同的查询执行计划,利用最小的预算从中挑选出能够比较精确响应当前查询的执行计划.然而,Pioneer 系统只考虑如何从当前的查询来节省隐私预算,而没有对将来查询的预算做出预测.

8.2 基于非交互式的数据分析系统

GUPT[52]主要是对文献[64]提出的抽样-聚集框架的扩展和实现.系统集成了数据分析者、数据拥有者和服务提供方三种角色.在假设数据分析者不可信的情况下为其提供数据分析和查询等服务.该系统首先借用强制访问控制框架(Mandatory Access Control,MAC)确保数据在通信时的安全性.系统根据数据拥有者所提供数据本身的时效性,自动地设置隐私预算来保护数据的隐私.之后 GUPT 隔离计算分析者提出的任务需求,并给出比较精确地结果.分析者的需求包括 k-均值聚类、逻辑回归分析和聚集查询等.然而,面对越来越多的分析任务,GUPT 要把隐私预算分割成多份,因此造成一些分析结果噪音量过大.

8.3 基于 MapReduce 的聚集分析系统

Airavat[73]结合 MapReduce 计算框架,在 MAC 与差分隐私技术的支持下为用户提供聚集分析结果.系统假设 MapReduce 框架中 Mapper 是不可信的,而信任提供计算结果的 Reducer.在 Mapper 端,利用 MAC 控制所有 key-value 对的映射过程;在 Reducer 端,利用 ε-差分隐私对计算结果添加噪音,进而实现隐私保护.目前 Airavat 系统仅提供聚集分析结果,而不支持基于数据挖掘和机器学习的任务分析.此外,该系统严格限制一定量的 key-value 对,因此不适应请求过大的任务分析.

文献[77]指出了差分隐私原型系统中可能存在的 3 种攻击:状态攻击、隐私预算攻击和定时攻击.状态攻击是指攻击者可能会修改系统中的静态变量;隐私预算攻击是指预算被耗尽;而定时攻击是指攻击者无限请求同一个分析任务.因此,我们结合这 3 种攻击比较上述 3 种原型系统的隐私保护性和安全性.表 9 给出了 3 种系统对应 3 种攻击的表现,"Yes"表示可以防御,"No"表示无法防御.从表 9 可以看出,PINQ 系统对上述 3 种攻击均无法预防,Airavat 和 Pioneer 仅能防止预算攻击,而 GUPT 却能够防止这 3 种攻击.

表 9 3 种原型系统安全性比较

原型系统	状态攻击	隐私预算攻击	计时攻击
PINQ[19]	No	No	No
GUPT[52]	Yes	Yes	Yes
Airavat[73]	No	Yes	No
Pioneer[76]	No	Yes	No

9 总结与展望

目前,差分隐私保护还是一个新的研究领域,很多挑战性的问题有待解决.虽然在第 5 节和第 6 节中,针对不同方法存在的不足,提出了许多未来研究方向,但是我们认为结合前面所介绍的方法还存在一些很具有挑战性的研究.

9.1 动态环境下的数据发布

已有的基于策略 1 或者策略 2 的发布方法大都是针对静态数据集的发布,未考虑数据动态变化时带来的挑战.而在实际应用中的数据通常随时间动态演化,例如疾病应急中心所记录 SARS 病毒携带者数据、Amazon 与 Flickr 网站动态推荐系统中的数据都是实时变化的.无论是应急中心还是服务网站都要动态地、实时地发布病人信息,或者商品销售信息.而这些信息常包含病人和客户的隐私信息,例如病人携带 HIV+,客户月薪 6000 等.动态数据的表现形式通常包括两种:一是数据流形式(例如应急中心监控病人的 RFID 数据流);二是数据以更新的形式出现(例如事务数据周期性地添加和删除记录).结合上述分析,我们总结出以下两点未来的研究工作.

(1)基于数据流的直方图发布.

虽然文献[27-28]针对静态数据集上的直方图的发布提出了几种代表性方法,我们认为,这些方法并不适合数据流式的直方图发布.其主要原因是静态的直方图不能满足连续性计数查询的需求(例如统计 11 月份~12 月份应急中心 HIV+病人数量).目前仅有文献[48-49,75]讨论了如何发布数据流的相关统计信息,这两篇工作所处理的动态对象是{0,1}组成的简单数据流,其发布的统计信息是到某个时刻之前所观测到 1 的个数.虽然这两种方法能够取得比较低的计数查询误差,但是对于数据流直方图发布来说,文献[48-49,75]所提出的方法还远远不够:一是缺乏良好的数据流采样模型来控制查询

敏感性；二是如何尽量延长隐私预算的使用寿命.

我们认为，可以采用滑动窗口(Sliding window)模型对数据流进行抽样建模. 在每个滑动窗口中，我们可以采用 V-优化直方图对窗口中的数据进行变换，由于滑动窗口大小固定，每次所采集的样本大小固定，进而使得发布 V-优化直方图的全局敏感性固定. 因此，可以利用这些特性控制查询敏感性和噪音量的大小. 然而，如果我们采用 V-优化技术发布直方图，在滑动窗向前滑动过程中，如何防止直方图自身结构披露隐私是一个具有挑战性的问题. 同时，如何利用动态规划技术设计高效的流式直方图发布方法又是一个很大的挑战.

(2) 增量更新环境下的集值型数据发布.

动态环境下另外一种数据体现是增量更新. 这类数据不具备数据流的实时性与无限性，它的典型操作是在原始数据集中添加一定量的数据，或者删除一定量的数据. 通常具有增量更新特性的数据为集值型数据，包括事务数据、搜索日志以及序列数据等. 目前，增量更新环境下还没有出现相应的集值数据发布工作. 尽管文献[34-35，40-41，43]提出了相应静态环境下集值数据发布方法，但是不能直接照搬这些方法. 在增量更新环境下直接应用基于静态数据的发布算法，虽然在某一时刻发布的数据满足差分隐私，但随着更新次数的增加，特别是无限次地更新，每次发布所需的噪音量会越来越大，所发布数据的可用性较低，累积误差比较大. 一旦隐私预算耗尽，差分隐私保护就不再起作用. 我们认为，在设计增量更新发布算法时，给出一个合理的更新边界比较合适，例如一个季度为一个更新上界. 然而，由于越靠后的更新所需的噪音越大，如何合理地分配隐私预算与有效地控制累积误差是需要研究的问题.

9.2　差分隐私下图数据的发布

已有的一些发布工作着眼于一维/多维的单个关系表上，而实际的应用中存在大量的复杂稀疏的图数据，例如，社交网络、路网以及生物化学路径等. 目前，文献[44，47] 结合边-差分隐私提出了几种支持 triangle 查询、k-stars 查询的图发布方法，然而这种查询的敏感性都非常大，有可能导致查询结果的噪音误差很大，发布结果的可用性比较低. 文献[45-46]分别基于边-差分隐私，提出了两种迭代式发布图数据的方法，而这些方法只适应于密集型的图数据. 实际上，现实的图数据大多是稀疏的，进而使得[44-47]中的方法无法适用于实际需求.

为了防止图数据本身的稀疏性导致发布数据的低可用性，我们认为，可以采用邻接矩阵与双聚类相结合的方法来发布图数据. 利用双聚类技术挖掘邻接矩阵中的所有满足阈值条件的密集区域，然后利用指数机制与拉普拉斯机制对密集区域中的所包含的边数进行扰动，使其真实计数不被披露. 然而，如何迭代地挖掘邻接矩阵中的密集区域是个很大的挑战.

9.3　分布式差分隐私保护

由于分布式环境下各个站点相互独立、数据异构的特点，通信、数据协同共享以及任务协同分析等操作会非常频繁. 而这些操作，无意间会对隐私信息造成威胁. 已有的差分隐私研究工作通常是针对集中式数据库应用需求问题，即假设数据库只有一个拥有者. 如果把数据部署到分布式系统中，例如云数据库系统，对于数据拥有者来说，他们担心不可信的云服务提供者(Untrusted cloud provider)企图扫描外包数据并出售，以及恶意攻击者的攻击. 目前已有文献[54]结合差分隐私保护技术与加密技术来解决分布式环境下的协同聚集问题(例如，多家医院把病人的医疗信息放到云端). 虽然，这些方法均可以防止不可信的云服务提供者，但是，我们认为还存在几个重要问题亟待解决：

已有的工作均假设所有的站点(数据提供者)是可信的，然而实际应用中，一些数据提供者可能与云服务端相互串通，导致密钥泄露以及聚集信息被窃取. 因此，在此应用情况下，如何利用差分隐私和加密技术防止聚集信息泄露是个迫切需要解决的问题.

此外，分布式环境下存在数据挖掘问题，例如，Heavy hitter 挖掘. 如何在该环境下，实现满足差分隐私保护的数据挖掘也是一个重要问题.

9.4　差分隐私下的大数据分析

目前，越来越多的应用涉及到大数据，例如，社交网络、微博、医疗信息、生命科学以及定位系统服务等. 虽然通过数据挖掘和机器学习技术对数据进行聚类、分割、孤立点分析以及回归分析等，可以抽取出有价值的知识以及数据内部的规律. 然而，大数据分析的最大障碍是数据隐私问题. 在某种程度上，隐私不可怕，可怕的是用户的行为可以通过大数据分析被预测出来. 例如，Facebook 就曾因跟踪用户的数据，并通过分析这些数据来评估 Facebook 的广告效果，而引发了隐私维权机构的质疑. 又比如，大数据下的个性化推荐系统，个性化推荐是电子商务网站根据用户的兴趣特点和购买行为，向用户推荐

感兴趣的信息和商品.然而,用户的商品购买信息以及行为模式很有可能被商务网站挖掘出来,进而导致隐私信息泄露.

我们认为,根据大数据的流式特点,利用采样技术、直方图以及概要技术,结合差分隐私可以做大数据的回归分析、模式挖掘、个性化推荐以及概要数据发布等研究.

9.5 差分隐私下其它研究点

首先,差分隐私保护的精髓是隐私预算 ε 带来的不确定性(Uncertainty).然而,对于数据发布者来说,当给定隐私预算 ε 的情况下,定量地度量差分隐私保护强度是非常困难的.而对于普通的数据用户,如何设置一个合适的预算 ε 值,来保护敏感数据并最大化所产生数据的可用性是个很大挑战.

其次,差分隐私保护成立的前提是假设数据集中的每条记录之间是相互独立的.然而,现实中的许多数据集中的记录之间是存在相互关联的.例如,某个数据 D 存储 Bob 家庭中 10 成员信息,假设 Bob 感染了传染性疾病,由于 10 个成员同属于一个家庭,具有彼此关联性,则我们可以推理出整个家庭可能都感染了传染性病.差分隐私对此类情况数据的保护程度很低,因此,如何扩展差分隐私保护框架,使其能够保护具有关联性的数据集是一个很大的挑战.

参 考 文 献

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

[2] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity//Proceedings of the 22nd International Conference on Data Engineering (ICDE). Atlanta, Georgia, USA, 2006: 24-35

[3] Li N, Li T. t-closeness: Privacy beyond k-anonymity and l-diversity//Proceedings of the 23rd International Conference on Data Engineering (ICDE). Istanbul, Turkey, 2007: 106-115

[4] Wong R C W, Li J, Fu A W, Wang K. (α,k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Philadelphia, Pennsylvania, USA, 2006: 754-759

[5] Ganta S R, Kasiviswanathan S P, Smith A. Composition attacks and auxiliary information in data privacy//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). New York, USA, 2008: 265-273

[6] Wong R C W, Fu A, Wang K, et al. Can the utility of anonymized data be used for privacy breaches. ACM Transactions on Knowledge Discovery from Data, 2011, 5(3): 16

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

[8] Dwork C. Differential privacy: A survey of results//Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC). Xi'an, China, 2008: 1-19

[9] Dwork C. Lei J. Differential privacy and robust statistics//Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC). Bethesda, MD, USA, 2009: 371-380

[10] Dwork C, Naor M, Reingold O, et al. On the complexity of differentially private data release: Efficient algorithms and hardness results//Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC). Bethesda, MD, USA, 2009: 381-390

[11] Dwork C. The differential privacy frontier (extended abstract)//Proceedings of the 6th Theory of Cryptography Conference (TCC). San Francisco, CA, USA, 2009: 496-502

[12] Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao Xiao-Kui. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009, 32(5): 847-861 (in Chinese)
(周水庚, 李丰, 陶宇飞, 肖小奎. 面向数据库应用的隐私保护研究综述. 计算机学报, 2009, 32(5): 847-861)

[13] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis//Proceedings of the 3th Theory of Cryptography Conference (TCC). New York, USA, 2006: 363-385

[14] McSherry F, Talwar K. Mechanism design via differential privacy//Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS). Providence, RI, USA, 2007: 94-103

[15] Li C, Hay M, Rastogi V, et al. Optimizing linear counting queries under differential privacy//Proceedings of the 41st Annual ACM Symposium on Theory of Computing (PODS). Bethesda, MD, USA, 2010: 123-134

[16] Zhang J, Zhang Z, Xiao X, et al. Functional mechanism: Regression analysis under differential privacy//Proceedings of the 38th Conference of Very Large Databases (VLDB). Istanbul, Turkey, 2012: 1364-1375

[17] Ghosh A, Roughgarden, Sundararajan M. Universally utility-maximizing privacy mechanism//Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC). Bethesda, MD, USA, 2009: 351-360

[18] Dwork C, Naor M, Vadhan S P. The privacy of the analyst and the power of the state//Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS). New Brunswick, NJ, USA, 2012: 400-409

[19] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Providence, Rhode Island, USA, 2009: 19-30

[20] Shen E, Yu T. Mining frequent graph patterns with differential privacy//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Chicago, USA, 2013

[21] Xiao X, Bender G, Hay M, Gehrke J. iReduct: Differential privacy with reduced relative errors//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Athens, Greece, 2011: 229-240

[22] Li Y, Zhang Z, Winslett M, Yang Y. Compressive mechanism: utilizing sparse representation in differential privacy//Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society (WPES). Chicago, USA, 2011: 177-182

[23] Xiao X, Xiong L, Yuan C. Differential privacy via wavelet transforms. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2011, 23(8): 1200-1214

[24] Hardt M, Talwar K. On the geometry of differential privacy//Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC). Cambridge, Massachusetts, USA, 2010: 705-714

[25] Chen R, Acs G, Castelluccia C. Differentially private sequential data publication via variable-length N-grams//Proceedings of the ACM Conference on Computer and Communications Security(CCS). Raleigh, NC, USA, 2012: 638-649

[26] Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency//Proceedings of the 36th Conference of Very Large Databases (VLDB). Istanbul, Turkey, 2010: 1021-1032

[27] Xu J, Zhang Z, Xiao X, et al. Differentially private histogram publication//Proceedings of IEEE 28th International Conference on Data Engineering (ICDE). Washington, DC, USA, 2012: 32-43

[28] Xu J, Zhang Z, Xiao X, et al. Differential private histogram publication. International Journal of Very Large Database (VLDBJ), 2013, 22(6): 797-822

[29] Jagadish H V, Koudas N, Muthukrishnan S, et al. Optimal histograms with quality guarantees//Proceedings of the 36th Conference of Very Large Databases (VLDB). New York, USA, 2010: 275-286

[30] Xiao Y, Xiong L, Fan L, Goryczka S. DPCube: Differentially private histogram release through multidimensional partitioning. CoRR Abs/1202.5358, 2012

[31] Acs G, Chen R. Differentially private histogram publishing through lossy compression//Proceedings of the 11th IEEE International Conference on Data Mining (ICDM). Brussels, Belgium, 2012: 84-95

[32] Rastogi V, Nath S. Differentially private aggregation of distributed time-series with transformation and encryption//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Indianapolis, Indiana, USA, 2010: 735-746

[33] Cormode G, Procopiuc C M, Srivastava D, et al. Differentially private spatial decompositions//Proceedings of IEEE 28th International Conference on Data Engineering (ICDE). Washington, USA, 2012: 20-31

[34] Chen R, Mohammed N, Fung B C M, et al. Publishing set-valued data via differential privacy//Proceedings of the 37th Conference of Very Large Databases (VLDB). Seattle, USA, 2011: 1087-1098

[35] Chen R, Fung B C M, Desai B C, Sossou N M. Differentially private transit data publication: A case study on the montreal transportation system//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Beijing, China, 2012: 493-502

[36] Inan A, Kantarcioglu M, Ghinita G, Bertino E. Private record matching using differential privacy//Proceedings of the 13th International Conference on Extending Database Technology. Lausanne, Switzerland(EDBT), 2010: 123-134

[37] He Y, Naughton J F. Anonymization of set-valued data via top-down, local generalization//Proceedings of the 35th Conference of Very Large Databases (VLDB). Lyon, France, 2009: 934-945

[38] Qardaji W H, Yang W, Li N. Differentially private grids for geospatial data//Proceedings of IEEE 29th International Conference on Data Engineering(ICDE). Brisbane, Australia, 2013: 757-768

[39] Xiao Y, Xiong L, Yuan C. Differentially private data release through multidimensional partitioning//Proceedings of the 7th VLDB Workshop on Secure Data Management (SDM). Singapore, 2010: 150-168

[40] Cormode G, Procopiuc C M, Srivastava D, Tran T T L. Differentially private summaries for sparse data//Proceedings of the 15th International Conference on Database Theory (ICDT). Berlin, Germany, 2012: 299-311

[41] Goetz M, Machanavajjhala A, Wang G, et al. Publishing search logs—A comparative study of privacy guarantees. IEEE Transactions on Knowledge and Data Engineering (TKDE). 2012, 24(3): 520-532

[42] Goetz M, Machanavajjhala A, Wang G, et al. Privacy in search logs. CoRR abs/0904.0682, 2009

[43] Korolova A, Kenthapadi K, Mishra N, Ntoulas A. Releasing search queries and clicks privately//Proceedings of 18th International Conference World Wide Web (WWW). Madrid, Spain, 2009: 171-180

[44] Karwa V, Raskhodnikova S, Smith A, Yaroslavtsev G. Private analysis of graph structure. Proceedings of the VLDB Endowment (PVLDB), 2011, 4(11): 1146-1157

[45] Sala A, Zhao X, Wilson C, et al. Sharing graphs using differentially private graph models//Proceedings of the 11th ACM SIGCOMM Conference on Internet Measurement (IMC). Berlin, Germany, 2011: 81-98

[46] Gupta A, Roth A, Ullman J. Iterative constructions and private data release//Proceedings of the 9th Theory of Cryptography Conference (TCC). Taormina, Italy, 2012: 339-356

[47] Rastogi V, Hay M, Miklau G, Suciu D. Relationship privacy: Output perturbation for queries with joins//Proceedings of the 28th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS). Providence, Rhode Island, USA, 2009: 273-282

[48] Dwork C, Naor M, Pitassi T, Rothblum G N. Differential privacy under continual observation//Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC). Cambridge, Massachusetts, USA, 2010: 715-724

[49] Chan T-H H, Shi E, Song D. Private and continual release of statistics. ACM Transactions on Information and System Security (ATIS), 2011, 14(3): 26

[50] Li C, Miklau G. An adaptive mechanism for accurate query answering under differential privacy//Proceedings of the 38th Conference of Very Large Databases (VLDB). Istanbul, Turkey, 2012: 514-525

[51] Cormode G, Procopiuc C M, Srivastava D, Yaroslavtsev G. Accurate and efficient private release of datacubes and contingency tables//Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE). Brisbane, Australia, 2013: 745-756

[52] Mohan P, Thakurta A, Shi E, et al. GUPT: Privacy preserving data analysis made easy//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Scottsdale, USA, 2012: 349-360

[53] Yuan G, Zhang Z, Winslett M, et al. Low-rank mechanism: Optimizing batch queries under differential privacy. Proceedings of the VLDB Endowment (PVLDB), 2012, 5(11): 1352-1363

[54] Shi E, Chan T-H H, Rieffel E G, et al. Privacy-preserving aggregation of time-series data//Proceedings of the Network and Distributed System Security Symposium (NDSS). San Diego, USA, 2011

[55] Fung B C M, Wang K, Yu P S. Top-down specialization for information and privacy preservation//Proceedings of the 21st International Conference on Data Engineering (ICDE). Tokyo, Japan, 2005: 205-216

[56] Yu S, Fung G, Rosales R, et al. Privacy-preserving cox regression for survival analysis//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Las Vegas, USA, 2008: 1034-1042

[57] Bhaskar R, Laxman S, Simth A, Thakruta A. Discovering frequent patterns in sensitive data//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Washington, USA, 2010: 503-512

[58] Li N, Qardaji W, Su D, Cao J. PrivBasis: Frequent itemset mining with differential privacy//Proceedings of the 38th Conference of Very Large Databases (VLDB). Istanbul, Turkey, 2012: 1340-1351

[59] Zeng C, Naughton J F, Cai J. On differentially private frequent itemset mining//Proceedings of the 39th Conference of Very Large Databases (VLDB). Trento, Italy, 2013: 1087-1098

[60] Blum A, Dwork C, McSherry F, Nissim K. Practical privacy: The SuLQ framework//Proceedings of the 24th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS). Baltimore, USA, 2005: 128-138

[61] Friedman A, Schuster A. Data mining with differential privacy//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Washington, USA, 2010: 493-502

[62] Mohammed N, Chen R, Fung B C M, Yu P S. Differentially private data release for data mining//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). San Diego, USA, 2011: 493-501

[63] Quinlan J R. Induction of decision tress. Machine Learning, 1989, 1(1): 81-106

[64] Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis//Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC). San Diego, USA, 2007: 75-84

[65] Dwork C. A firm foundation for private data analysis. Communications of the ACM, 2011, 54(1): 86-95

[66] Smith A. Privacy-preserving statistical estimation with optimal convergence rate//Proceedings on the 43th Annual ACM Symposium on Theory of Computing (STOC). 2011: 813-822

[67] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression//Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS). Vancouver, British Columbia, Canada, 2008: 289-296

[68] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization. Journal of Machine Learning Research, 2011, 12: 1069-1109

[69] Lei J. Differentially private *m*-estimators//Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS). Granada, Spain, 2011: 361-369

[70] Gupta A, Ligett K, McSherry F, et al. Differentially private combinatorial optimization//Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). Austin, USA, 2010: 1106-1125

[71] Peng S, Yang Y, Zhang Z, et al. DP-tree: Indexing multi-dimensional data under differential privacy//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Scottsdale, USA, 2012: 864

[72] Barak B, Chaudhuri K, Dwork C, et al. Privacy, accuracy, and consistency too: A holistic solution to contingency table release//Proceedings of the 26th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS). Beijing, China, 2007: 273-282

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

[74] Kellaris G, Papadopoulos S. Practical differential privacy via grouping and smoothing//Proceedings of the 39th Conference of Very Large Databases (VLDB). Trento, Italy, 2013: 301-312

[75] Bolot J, Fawaz N, Muthukrishnan S, et al. Private decayed predicate sums on streams//Proceedings of the 15th International Conference on Database Theory (ICDT). Genoa, Italy, 2013: 284-295

[76] Peng S, Yang Y, Zhang Z, et al. Query optimization for differentially private data management systems//Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE). Brisbane, Australia, 2013: 1093-1104

[77] Haeberlen A, Pierce B C, Narayan A. Differential privacy under fire//Proceedings of the 20th USENIX Security Symposium. San Francisco, USA, 2011

[78] Li C, Miklau G. Efficient batch query answering under differential privacy, 2011: CoRR abs/1103

ZHANG Xiao-Jian, born in 1980, Ph. D. candidate. His main research interests include differential privacy and data mining.

MENG Xiao-Feng, born in 1964, professor, Ph. D. supervisor. His main research interests include cloud data management, Web data management, native XML databases, and flash-based databases, privacy protection, etc.

Background

Digital techniques have enabled many organizations to easily collect large amounts of personal information, such as transaction data, web search queries, etc. Publication and analysis on such data can potentially provide enormous opportunities for those organizations such as marketing, and advertising. However, such data often involves sensitive information that could breach individual privacy. Most of the existing studies mainly depend on specific assumptions about the background knowledge of the adversary, which may lead to rather limited privacy protection.

Differential privacy has emerged as a new model that provides strong privacy guarantees independent of an adversary's background knowledge, in general, which requires the outcome of any analysis should not excessively rely on a single user record. Therefore, no matter how much the adversary knows about the other records in a database, the adversary will be unable to guess whether the user opts-in or opts-out the database.

In this paper, The authors give an overview of the state-of-the-art methods for data publication and analysis with differential privacy, including histogram publication, indexing tree publication, grid publication, frequent pattern mining, regression analysis and classification, etc. The authors organize existing works in terms of the interactive and non-interactive frameworks, and give an in depth analysis to the representative methods. Finally, according to the inherent requirements of data publication and analysis based on differential privacy, they discuss data releases in terms of histogram and partition techniques, analysis based on regression skills, and identify the challenges and opportunities of future research in this field.

Differential privacy is still young research field that have received a lot of concerns recent years. Before this paper, they have studied this direction since 2012, and solved several interesting problems. Related research papers are published in DASFAA, Journal of Computer Research and Development, WAIM.

This research is partially supported by the grants from the Natural Science Foundation of China (Nos. 61379050 and 91224008); the National 863 High-Tech Program (No. 2013AA013204); Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130004130001).