软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.005364]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

本地化差分隐私研究综述*

叶青青[1], 孟小峰[1], 朱敏杰[1], 霍 峥[2]

[1](中国人民大学 信息学院,北京 100872)

[2](河北经贸大学 信息技术学院,河北 石家庄 050061)

通讯作者: 孟小峰, E-mail: xfmeng@ruc.edu.cn

摘 要: 大数据时代信息技术不断发展,个人信息的隐私问题越来越受到关注,如何在数据发布和分析的同时保证其中的个人敏感信息不被泄露是当前面临的重大挑战.中心化差分隐私保护技术建立在可信第三方数据收集者的假设基础上,然而该假设在现实中不一定成立.基于此提出的本地化差分隐私作为一种新的隐私保护模型,具有强隐私保护性,不仅可以抵御具有任意背景知识的攻击者,而且能够防止来自不可信第三方的隐私攻击,对敏感信息提供了更全面的保护.介绍了本地化差分隐私的原理与特性,总结和归纳了该技术的当前研究工作,重点阐述了该技术的研究热点:本地化差分隐私下的频数统计、均值统计以及满足本地化差分隐私的扰动机制设计.在对已有技术深入对比分析的基础上,指出了本地化差分隐私保护技术的未来研究挑战.

关键词: 隐私保护;本地化;中心化;差分隐私

中图法分类号: TP311

中文引用格式: 叶青青,孟小峰,朱敏杰,霍峥.本地化差分隐私研究综述.软件学报,2018,29(7). http://www.jos.org.cn/1000-9825/5364.htm

英文引用格式: Ye Q, Meng X, Zhu M, Huo Z. Survey on local differential privacy. Ruan Jian Xue Bao/Journal of Software, 2018,29(7) (in Chinese). http://www.jos.org.cn/1000-9825/5364.htm

Survey on Local Differential Privacy

YE Qing-Qing[1], MENG Xiao-Feng[1], ZHU Min-Jie[1], HUO Zheng[2]

[1](School of Information, Renmin University of China, Beijing 100872, China)

[2](School of Information Technology, Hebei University of Economics and Business, Shijiazhuang, 050061, China)

Abstract: With the development of information technology in the big data era, there has been a growing concern for privacy of personal information. Privacy preserving is a key challenge when releasing and analyzing data. Centralized differential privacy is based on the assumption of a trustworthy data collector; however, it is actually a bit difficult to realize in practice. To this end, local differential privacy has emerged as a new model for privacy preserving with strong privacy guarantees. By resisting adversaries with any background knowledge and preventing attacks from untrustworthy data collector, local differential privacy can protect private information thoroughly. Starting with an introduction to the mechanisms and properties, this paper surveys the state of the art of local differential privacy, focusing on the frequency estimation, mean value estimation and the design of perturbation model. Following a comprehensive comparision and analysis of existing techniques, further research challenges are put forward.

* 基金项目: 国家自然科学基金(91646203,61532010,61532016,61379050); 国家重点研发计划项目(2016YFB1000602, 2016YFB1000603);中国人民大学科学研究基金(11XNL010); 河北省自然科学基金(F2015207009)

Foundation item: National Natural Science Foundation of China (91646203,61532010,61532016,61379050); The National Key Research and Development Program of China(2016YFB1000602, 2016YFB1000603); The Research Funds of Renmin University (11XNL010); Natural Science Foundation of Hebei(F2015207009)

收稿时间: 2017-06-11; 修改时间: 2017-07-13; 采用时间: 2017-08-22; jos 在线出版时间: 2017-10-17

本文由面向隐私保护的新型技术与密码算法专刊特约编辑薛锐研究员推荐.

CNKI 网络优先出版: 2017-10-17 13:42:45, http://kns.cnki.net/kcms/detail/11.2560.TP.20171017.1342.010.html

Key words: privacy preserving; local differential privacy; centralized differential privacy

近年来,隐私问题成为普遍关注的热点问题.大数据时代,信息技术为人类社会带来便捷的同时,也产生了数据安全与用户隐私的问题.为保证信息技术的长足发展,保护个人数据隐私成为政府和企业的当务之急.在隐私保护问题上,欧盟走在了时代前沿.2016 年 4 月,欧盟通过了《一般数据法案》[1](General Data Protection Regulation, GDPR),规定了个人数据保护跨越国界,同时其明确了用户对个人信息的知情权和被遗忘权.我国于 2017 年 6 月 1 日起施行《中华人民共和国网络安全法》和《最高人民法院、最高人民检察院关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》[2],加强了个人信息保护,其中对于提供公民个人信息违法所得五千元以上可入罪.

对隐私问题的重视促进了隐私保护技术的研究.就隐私保护技术而言,隐私保护程度和数据可用性是最重要的衡量指标.为了平衡隐私保护程度和数据可用性,需要引入形式化定义对隐私进行量化,顺应这一发展趋势,研究者提出了差分隐私[1,2,3]技术.作为一种隐私保护模型,其严格定义了隐私保护的强度,即任意一条记录的添加或删除,都不会影响最终的查询结果.同时,该模型定义了极为严格的攻击模型,其不关心攻击者具有多少背景知识.相比于 k-匿名[4],l-多样性[5]和 t-紧密性[6]等需要特殊攻击假设和背景知识的方法,差分隐私因其独特的优势,成为当前学术界的研究热点.

传统的差分隐私技术将原始数据集中到一个数据中心,然后发布满足差分隐私的相关统计信息,我们称之为中心化差分隐私(Centralized Differential Privacy)技术.因此,中心化差分隐私对于敏感信息的保护始终基于一个前提假设:可信的第三方数据收集者,即保证第三方数据收集者不会窃取或泄露用户的敏感信息.然而,在实际应用中,即使第三方数据收集者宣称不会窃取和泄露用户的敏感信息,用户的隐私依旧得不到保障.2016 年,社交网络的数据泄露事件层出不穷[3]:美国社交网站 LinkedIn 近 1.7 亿个账户被黑客组织在黑市被公开销售;谷歌、雅虎和微软等企业超 2.7 亿电子邮箱信息被一名俄罗斯黑客盗取并流入黑市;土耳其近 5000 万公民个人信息被泄露,总统的个人信息被挂暗网平台;雅虎爆发互联网史上最大数据泄露,超 5 亿用户账户信息被黑客盗取;美国国安局网站遭入侵,其中黑客工具和数据被泄露,国安局网站因此瘫痪了近一昼夜.此类用户原始信息泄露事件近年来层见叠出,人们对个人信息的安全性十分担忧.

由此可知,在实际应用中想要找到一个真正可信的第三方数据收集平台十分困难,这极大地限制了中心化差分隐私技术的应用.鉴此,在不可信第三方数据收集者的场景下,本地化差分隐私(Local Differential Privacy)[7,8]技术应运而生,其在继承中心化差分隐私技术定量化定义隐私攻击的基础上,细化了对个人敏感信息的保护.具体来说,其将数据的隐私化处理过程转移到每个用户上,使得用户能够单独地处理和保护个人敏感信息,即进行更加彻底的隐私保护.目前,本地化差分技术在工业界已经得到运用:苹果公司将该技术应用在操作系统 iOS 10 上保护用户的设备数据[4],谷歌公司同样使用该技术从 Chrome 浏览器采集用户的行为统计数据[19].

本地化差分隐私技术继承自中心化差分隐私技术,同时扩展出了新的特性,使该技术具备两大特点:1)充分考虑任意攻击者的背景知识,并对隐私保护程度进行量化;2)本地化扰动数据,抵御来自不可信第三方数据收集者的隐私攻击.下面我们通过两个具体应用场景说明本地化差分隐私技术的上述两个特点的重要性.

(1) 众包数据采集.

众包(Crowdsourcing)[9]是一种利用群体智慧求解问题的方式,众包技术极大地促进了信息技术的发展,其中通过众包的方式进行数据采集是一种新的数据采集方式.由于移动设备功能的不断强大,数据收集者可以很容易将数据采集的任务分配给不同用户,例如,美国 Gigwalk 公司组织用户通过智能设备采集不同商品的价格;

1 https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

2 http://www.spp.gov.cn/xwfbh/wsfbt/201705/t20170509_190088.shtml

3 http://datayuan.baijia.baidu.com/article/715477

4 https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/

国内数据堂公司组织用户通过“众客堂”APP 进行图片、文本和语音的采集并标注以提供数据服务;高德地图公司组织用户通过“道路寻宝”APP 采集道路周边信息等.然而,众包数据采集一般与个人行为信息相关,因此当用户参与众包数据采集时,不可避免地存在泄露个人敏感信息的风险,例如,上传商品的价格时可能泄露个人的购物偏好信息,上传图片、语音等信息可能泄露个人身份特征信息,上传道路周边信息可能泄露用户的位置和轨迹信息等.因此,众包数据采集还需要隐私保护技术为其保驾护航.

关于众包数据采集的隐私问题研究,文献[10]曾指出隐私问题是众包技术发展的一大挑战,文献[11]提出基于 k-匿名的众包数据保护方法,以及文献[12]提出基于编码扰动的方法等.然而,上述研究并未考虑基于背景知识的攻击.现如今,各类运营商和大数据企业拥有大量的用户数据,互联网上大规模数据相互关联,各种数据集成和融合技术蓬勃发展,同时互联网本身的便捷性使得各种类型的信息触手可及,这些因素的综合作用使得攻击者可以很容易地从中获取背景知识,从而结合用户上传的数据推测用户的敏感信息.例如,某个用户频繁在工作日上传 A、B 两地之间的不同建筑物图片,若根据背景知识得知 A、B 分别为某居民区和商业区,则能够以较大概率猜测 A、B 两地分别为该用户的居住和工作地址.因此,在众包数据采集问题上迫切需要一种能够充分考虑背景知识,严格定义攻击模型的隐私保护技术.

(2) 敏感图像特征提取.

各种各样来自个人的图像中蕴含着诸多敏感信息,例如人脸图像、指纹图像和虹膜图像等生物特征数据可以唯一精确定位到个人,医学造影图像中蕴含着个人的相关疾病信息.直接对蕴含敏感信息的图像进行存储和分析,可能泄露其中的隐私.以如今与指纹相关的应用为例,移动智能终端的发展使得移动支付流行开来,其中以指纹支付最为便捷,因此,指纹图像的敏感性不言而喻.其它蕴含敏感信息的图像亦如此,对此类图像的存储和分析过程进行相应的隐私保护是必要的.

图像特征提取是图像处理中最初级的运算[13],是进行图像识别的关键.目前已有诸多研究工作围绕图像与隐私展开,文献[14]以医学图像为例说明了图像中蕴含的隐私问题;文献[15]提出人脸识别过程中的通过安全多方计算(secure multiparty computation)进行隐私保护;文献[16]基于云计算环境提出图像特征提取过程中基于加密的隐私保护方法等.然而,现有方法均建立在可信的数据收集者的基础上.在图像信息高度敏感的情形下,可信第三方数据收集者的假设难以立足,例如,苹果用户曾一度质疑苹果公司的云平台是否存储了用户的指纹图像.因此,敏感图像特征提取问题上迫切需要一个能够抵御不可信第三方数据收集者的隐私模型.而本地化差分隐私技术对于敏感图像特征提取的场景具有很好的适应性,该技术在用户端完成对图像的扰动处理,保证了无论是数据收集者或是数据传输过程中的攻击者,均无法窃取图像中的隐私信息.无独有偶,文献[17]曾指出云计算环境下本地化差分隐私技术在图像处理这一领域的巨大潜力.

目前,本地化差分隐私技术已经成为继中心化差分隐私技术之后一种强健的隐私保护模型[19,18,56].首先,用户对原始数据进行满足 ε-本地化差分隐私的扰动,然后将其传输给第三方数据收集者,数据收集者收到扰动后的数据后进行一系列的查询和求精处理,以得到有效的统计结果.对本地化差分隐私的研究和应用,主要考虑以下两个方面问题:(1)如何设计满足 ε-本地化差分隐私的数据扰动算法,以保护其中的敏感信息;(2)数据收集者如何对查询结果进行求精处理,以提高统计结果的可用性.

本文综述本地化差分隐私技术的最新研究进展和研究方向,一方面对本地化差分隐私的研究背景、基本定义、实现机制以及其与中心化差分隐私技术的区别进行阐述,另一方面,对当前本地化差分隐私的研究方向进行分析,并阐述最新研究进展,其中着重介绍本地化差分隐私下的数据扰动机制以及两种基本的数据发布形式:频数统计和均值统计.最后,针对本地化差分隐私的特性,提出本地化差分隐私保护技术未来的研究方向并进行具体分析.

本文第 1 节介绍本地化差分隐私保护技术的基础知识;第 2 节介绍其数据保护框架;第 3、4 节对本地化差分隐私保护技术的当前研究方向进行概括,并对研究方法进行对比和分析;第 5 节提出本地化差分隐私保护技术的研究挑战;最后第 6 节总结全文.

1 基础知识

本地化差分隐私保护技术是基于中心化差分隐私保护技术提出的数据采集框架,不同于中心化差分隐私对于可信第三方的假设,其针对的是不可信的第三方数据收集者.本节首先对本地化差分隐私进行形式化定义,接着阐述满足其定义的一种通用保护机制,最后对本地化和中心化差分隐私保护技术进行对比分析.

1.1 本地化差分隐私的定义

本地化差分隐私下的保护模型充分考虑了数据采集过程中数据收集者窃取或泄露用户隐私的可能性.该模型中,每个用户首先对数据进行隐私化处理,再将处理后的数据发送给数据收集者,数据收集者对采集到的数据进行统计,以得到有效的分析结果.即,在对数据进行统计分析的同时,保证个体的隐私信息不被泄露.本地化差分隐私的形式化定义如下:

定义 1. 给定 n 个用户,每个用户对应一条记录,给定一个隐私算法 M 及其定义域 $Dom(M)$ 和值域 $Ran(M)$,若算法 M 在任意两条记录 t 和 t' ($t,t' \in Dom(M)$)上得到相同输出结果 t^* ($t^* \subseteq Ran(M)$)满足下列不等式,则 M 满足 ε-本地化差分隐私.

$$\Pr[M(t)=t^*] \le e^{\varepsilon} \times \Pr[M(t')=t^*]$$

从定义 1 中可以看出,本地化差分隐私技术通过控制任意两条记录的输出结果的相似性,从而确保算法 M 满足 ε-本地化差分隐私.简言之,根据隐私算法 M 的某个输出结果,几乎无法推理出其输入数据为哪一条记录.在中心化差分隐私保护技术中,算法 M 的隐私性通过近邻数据集[2]来定义,因此其要求一个可信的第三方数据收集者来对数据分析结果进行隐私化处理.对于本地化差分隐私技术而言,每个用户能够独立地对个体数据进行处理,即,隐私化处理过程从数据收集方转移到单个用户端上,因此不再需要可信第三方的介入,同时也免除了不可信第三方数据收集者可能带来的隐私攻击.

定义 1 从理论的角度保证了算法满足 ε-本地化差分隐私,而实现 ε-本地化差分隐私保护需要数据扰动机制的介入.

1.2 扰动机制

目前,随机响应(Randomized Response)技术[18]是本地化差分隐私保护技术的主流扰动机制,本节主要对其原理进行阐述,对于现有研究中的其它扰动机制,将在 4.3 节中进行分析和比较.

1.2.1 随机响应技术

Warner 于 1965 年提出利用随机响应技术进行隐私保护,我们将其称为 W-RR,其主要思想是利用对敏感问题响应的不确定性对原始数据进行隐私保护.本节首先介绍随机响应技术的原理,而后分别阐述连续型数据和离散型数据下随机响应技术的应用.

随机响应技术主要包括两个步骤:扰动性统计和校正.

为了具体介绍随机响应技术,下面首先引入一个具体的问题场景.假设有 n 个用户,其中艾滋病患者的真实比例为 π,但我们并不知道.我们希望对其比例 $\hat{\pi}$ 进行统计.于是我们发起一个敏感的问题:“你是否为艾滋病患者?”每个用户对此进行响应,第 i 个用户的答案 X_i 为是或否,但出于隐私性考虑,用户不会直接响应真实答案.假设其借助于一枚非均匀的硬币来给出答案,其正面向上的概率为 p,反面向上的概率为 $1-p$.抛出该硬币,若正面向上则回答真实答案,反面向上则回答相反的答案.

首先,进行扰动性统计.利用上述扰动方法对 n 个用户的回答进行统计,可以得到艾滋病患者人数的统计值.假设统计结果中,回答“是”的人数为 n_1,则回答“否”的人数为 $n-n_1$.显然,按照上述统计,回答“是”和“否”的用户比例如下:

$$\Pr(X_i=\text{"是"})=\pi p+(1-\pi)(1-p)$$

$$\Pr(X_i=\text{"否"})=(1-\pi)p+\pi(1-p)$$

显然,上述统计比例并非真实比例的无偏估计,因此需要对统计结果进行校正.

接着,对统计结果进行校正.构建以下似然函数:

$$L=[\pi p+(1-\pi)(1-p)]^{n_1}[(1-\pi)p+\pi(1-p)]^{n-n_1}$$

并得到 π 的极大似然估计:

$$\hat{\pi}=\frac{p-1}{2p-1}+\frac{n_1}{(2p-1)n}$$

以下关于 $\hat{\pi}$ 的数学期望保证了 $\hat{\pi}$ 是真实分布 π 的无偏估计:

$$E(\hat{\pi})=\frac{1}{2(p-1)}\left[p-1+\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{2(p-1)}\left[p-1+\pi p+(1-\pi)(1-p)\right]=\pi$$

由此可得到校正的统计值,其中 N 表示统计得到的艾滋病人数估计值:

$$N=\hat{\pi}\times n=\frac{p-1}{2p-1}n+\frac{n_1}{2p-1}$$

综上,根据总人数 n 、回答“是”的人数 n_1 和扰动概率 p ,即可得到真实患病人数的统计值.为保证其满足 ε -本地化差分隐私,根据定义,隐私预算 ε 设定为:

$$\varepsilon=\ln\frac{p}{1-p}$$

1.2.2 离散型数据的随机响应

随机响应技术 W-RR 仅对包含两种取值的离散型数据进行响应,而对于具有超过两种取值的数据并不适用.因此,利用 W-RR 对离散型数据进行扰动有以下两种思路:1)对变量的不同取值进行编码和转化,使其满足 W-RR 技术对二值变量的要求;2)改进 W-RR 技术,使其能够直接适用于超过两种取值的变量.假设变量 x 总共包含 k 种不同取值,其取值集合为 $S=\{x_1,x_2,...,x_k\}$,我们称集合 S 为 x 的候选值集合.

(1)第一种思路的思想是:对于 k 个候选值,将每一个候选值都编码成长度为 $\lceil\log k\rceil$ 的 0/1 串,然后对 0/1 串的每一个位上的 0/1 进行随机响应.然而,由于 $\log k$ 并非刚好取整,因此存在某些 0/1 串未能匹配到相应候选值的情形,由此造成的匹配误差以及随机响应技术本身的扰动误差,将使得数据发布的可用性降低,可见该思路下的技术难点还在于属性候选值的编码和匹配策略上.基于该思路,本地化差分隐私下,离散型数据的随机响应方法包括 RAPPOR[19]和 S-Hist[20]等,详见 4.1.1 节.

(2)第二种思路,需要对 W-RR 技术中的概率分布进行改进.具体来说,W-RR 中将概率分配到变量的两种取值上,而对于 k 种取值的情况,需要保证概率的分布能够覆盖到 k 种取值中的任意一种.基于该思路,本地化差分隐私下,离散型数据的随机响应方法包括 k -RR[21]和 O-RR[22]等,详见 4.1.1 节.

1.2.3 连续型数据的随机响应

随机响应技术 W-RR 不能直接用于连续型数据的扰动,因此需要对连续型数据进行转换.其主要思想是,将连续型数据离散化,然后利用离散型数据下的随机响应方法,对数据进行扰动.目前已有的方法一般是将连续型数据离散化为某两个数值,然后对离散化后的数据利用随机响应技术 W-RR 进行扰动.

通过离散化并扰动后的值得到统计量,如变量的平均值,出于数据可用性的考虑,需保证统计结果与真实结果的无偏性.因此,面向连续型数据的随机响应技术的难点主要在两个方面:1)如何合理设置离散化的两个数值;2)如何保证统计结果的无偏性.

基于上述思路,本地化差分隐私下,连续型数据的随机响应方法包括 MeanEst[23,24]和 Harmony-mean[25]等,详见 4.2 节.

1.3 本地化与中心化差分隐私的异同点

本地化差分隐私保护技术是在中心化差分隐私保护技术的基础上提出的,其继承了中心化差分隐私保护技术上的组合特性,同时又对其进行了扩展,利用随机响应的扰动机制抵抗不可信的第三方数据收集者带来的

隐私攻击.下面从两者的异同点和应用场景出发,对两种技术进行比较.

(1) 组合特性.

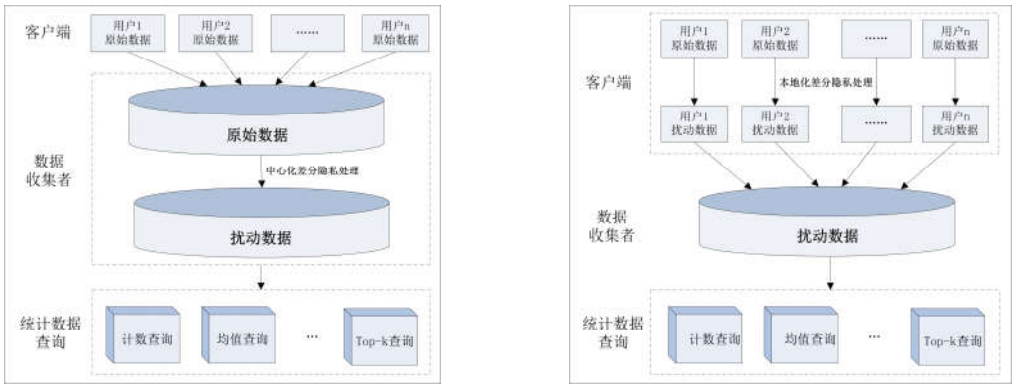
差分隐私技术具有序列组合型和并行组合性两种特性[26],序列组合性强调隐私预算 ε 可以在方法的不同步骤进行分配,而并行组合性则是保证满足差分隐私的算法在其数据集的不相交子集上的隐私性.从定义上来看,中心化差分隐私定义在近邻数据集上,而本地化差分隐私则是定义在其中的两条记录上,而隐私保证的形式并未发生变化,因此本地化差分隐私将上述两种组合特性继承下来,下面给出形式化定义.

性质 1. 给定数据集合 D 和 n 个隐私算法 $\{M_1,...,M_n\}$,且 $M_i(1\le i\le n)$ 满足 ε_i -本地化差分隐私,那么 $\{M_1,...,M_n\}$ 在 D 上的序列组合满足 ε -本地化差分隐私,其中 $\varepsilon=\sum_{i=1}^{n}\varepsilon_i$.

性质 2. 给定数据集合 D,将其划分为 n 个互不相交的子集,$D=\{D_1,...,D_n\}$,设 M 为任一满足 ε -本地化差分隐私的隐私算法,则算法 M 在 $\{D_1,...,D_n\}$ 上满足 ε -本地化差分隐私.

(2) 可信与不可信第三方.

中心化差分隐私中一个重要的假设是可信第三方数据收集者,每个用户将自己的真实数据记录发送给数据收集者,并假定其是可信的,不会泄露个人的敏感信息.而后,数据收集者利用满足需求的隐私算法对数据分析者的查询请求进行响应.本地化差分隐私中,其考虑的是第三方数据收集者不可信的的场景,因此将数据扰动的功能从数据收集方转移到了客户端,即从数据收集者处转移到了每个用户处.每个用户按照隐私算法对数据进行扰动,然后把数据上传给数据收集者,数据收集者接收数据分析者的查询请求,并进行响应.图 1(a)为中心化差分隐私保护技术的数据处理框架,图 1(b)为本地化差分隐私保护技术的数据处理框架.



(a) 中心化差分隐私处理框架 (b) 本地化差分隐私处理框架

Fig.1 Processing framework of centralized and local differential privacy

图1 中心化与本地化差分隐私的数据处理框架

(3) 噪声机制.

中心化差分隐私保护技术中,为保证所设计的算法满足 ε -差分隐私,需要噪声机制的介入,拉普拉斯机制[27]和指数机制[28]是其最常用的两种噪声机制,其中拉普拉斯机制面向连续型数据的查询,而指数机制面向离散型数据的查询.上述两种噪声机制均与查询函数的全局敏感性[27]密切相关,而全局敏感性则是定义在至多相差一条记录的近邻数据集之上,使得攻击者无法根据统计结果推测个体记录,即将个体记录隐藏在统计结果之中.在本地化差分隐私中,每个用户将各自的数据进行扰动后,再上传至数据收集者处,而任意两个用户之间并不知晓对方的数据记录,也即,本地化差分隐私中并不存在全局敏感性的概念,因此拉普拉斯机制和指数机制并不适用.目前,本地化差分隐私主要采用 1.2 节中所述的随机响应技术来确保隐私算法满足 ε -本地化差分隐私.

(4) 应用场景.

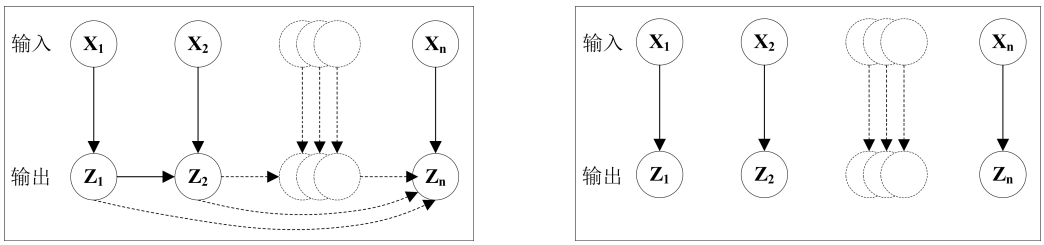
中心化差分隐私保护技术的研究主要集中在数据发布[29,30,31,32]、数据分析[33,34]和查询处理[35,36]等方面,近年来取得突出研究进展,然而,其数据的隐私化处理过程始终依靠一个可信的第三方数据收集者来完成.某种程度上来说,这一点也限制了差分隐私技术的发展,在隐私意识不断增强的背景下,如何保证数据收集者不会从中窃取用户的隐私信息将是一个重要的考量.基于此,本地化差分隐私技术应运而生,进一步细化了对个人隐私信息的保护,其摒弃了可信第三方数据收集者的假设,将数据的隐私化处理过程转移到每个用户上,这样不仅能对敏感信息进行更加彻底的保护,而且隐私化处理过程更加简洁明了.此外,由于每个用户能够掌握个人的敏感信息处理过程,这也使得用户可以根据自身需求,进行更加个性化的隐私设置[37].中心化差分隐私技术通过定义全局敏感性为查询结果添加响应噪声,再以统计的方式限制隐私信息泄露的量化边界,从而能将个体记录隐藏在统计结果中.因此,中心化差分隐私技术并不对统计数据量作特别要求.不同于此,本地化差分隐私技术对个体数据进行正向和负向的扰动,最终通过聚合大量的扰动结果来抵消添加在其中的正负向噪声,从而得到有效的统计结果.然而,由于噪声的随机性,要保证统计结果的无偏性,必然需要海量的数据集来实现满足数据可用性的统计精度.

2 基于本地化差分隐私的数据保护框架

中心化差分隐私技术对应两种数据保护框架:交互式和非交互式框架[38],其中,第三方数据收集者采集了所有用户的数据,并对用户的查询进行响应.在交互式框架中,当数据分析者提交相应的查询时,数据收集者根据查询请求,对查询结果进行相应的隐私化处理,如添加相应噪声,使其满足差分隐私要求.在非交互式框架中,数据收集者事先发布满足差分隐私的数据集相关统计信息,数据分析者提交查询后,直接从所发布的统计信息中返回相应的查询结果.

在本地化差分隐私中同样存在交互式和非交互式两种数据保护框架[23],但由于本地化差分隐私建立在不可信第三方数据收集者的基础上,因此对两种保护框架的定义不同于中心化差分隐私.如图 2 所示,其中 $X_1,...,X_n \in \chi$ 为输入序列, $Z_1,...,Z_n \in Z$ 为在 χ 上的查询 Q 对应的输出,箭头表示依赖关系,形式化表示为:

$$Q(Z_i \mid X_i = x, Z_{1:i-1} = z_{1:i-1})$$



(a) 交互式框架 (b) 非交互式框架

Fig.2 Framework of local differential privacy

图2 本地化差分隐私的数据保护框架

交互式框架,如图 2(a)所示,第 i 个输出 Z_i 依赖于第 i 个输入 X_i 以及前 i 个输出 $Z_{1:i-1}$,但与前 i 个输入 $X_{1:i-1}$ 无依赖关系,即对任意的 $i \neq j$,依赖关系的形式化表示如下:

$$\{X_i, Z_1,...,Z_{i-1}\} \to Z_i \text{ 且 } Z_i \perp X_j \mid \{X_i, Z_1,...,Z_{i-1}\}$$

因此,交互式框架下,本地化差分隐私的形式化定义为:对任意的 $x, x' \in \chi$,给定隐私预算 ε,若查询 Q 满足以下不等式,那么认为 Z_i 是 X_i 的一个满足 ε -本地化差分隐私保护的表示.

$$\sup_S \frac{Q(Z_i \in S \mid X_i = x, Z_1 = z_1,..., Z_{i-1} = z_{i-1})}{Q(Z_i \in S \mid X_i = x', Z_1 = z_1,..., Z_{i-1} = z_{i-1})} \le e^{\varepsilon}$$

非交互式框架则对交互式框架的简化,如图 2(b)所示,其中第 i 个输出 Z_i 仅依赖于第 i 个输入 X_i ,形式化表示为:

$$X_i \to Z_i \text{ 且 } Z_i \perp \{X_j, Z_j, j \neq i\} \mid X_i$$

因此,非交互式框架下的本地化差分隐私,其形式化定义则是交互式框架下的一种简化,表示为:对任意的 $x, x' \in \chi$,给定隐私预算 ε ,若查询 Q 满足以下不等式,那么认为 Z_i 是 X_i 的一个满足 ε -本地化差分隐私保护的表示.

$$\sup_S \frac{Q(Z_i \in S \mid X_i = x)}{Q(Z_i \in S \mid X_i = x')} \leq e^{\varepsilon}$$

综上所述,交互式和非交互式数据保护框架的最大区别在于输出结果之间的关联性.1)交互式框架适用于最终输出结果与前 i 个输出有依赖关系的情形,如通过家族病史数据进行疾病诊断.假设某个家族存在红绿色盲的遗传病史,家族中的某个成员有相应的性状表现 X_i ,那么在判断其是否患有红绿色盲,即判断 $Z_i = 1$ 还是 $Z_i = 0$ 时,除了需要考虑该成员本身的性状表现,还需要考虑其父辈和祖辈是否患有红绿色盲,即考虑 $Z_{1:i-1}$ 的取值情况.家族病史数据记录了家族成员对于某些疾病的患病情况,由于遗传等因素,其中通常存在前后的关联关系.对此类数据进行本地化差分隐私保护时,由于某个个体的数据会对其他个体的输出产生影响,因此需要考虑用交互式框架对其进行保护.2)非交互式框架适用于前后的输入输出之间无依赖关系的情形,如商场的购物数据分析.一般而言,不同用户的购物清单数据之间不存在相互的关联关系,因此,对该类数据进行本地化差分隐私保护时,直接应用非交互式框架即可.

基于本地化差分隐私的数据保护框架充分考虑了数据记录之间的相互关联关系,并以此为依据将其分为交互式和非交互式框架.而中心化差分隐私并未直接考虑数据的关联关系,文献[39]指出数据记录中的关联性将弱化中心化差分隐私技术的保护能力,文献[40]则进一步讨论了具有不同程度背景知识的攻击者对关联数据中的用户隐私的攻击能力,并提出基于贝叶斯理论的保护模型.对比而言,本地化差分隐私技术进一步考虑了数据中包含的关联信息,因此其考虑了更为全面的隐私保护场景.

3 主要研究方向

本地化差分隐私作为新兴的隐私保护技术,是当前的研究热点,主要应用于统计数据库领域.目前,本地化差分隐私技术的主要研究方向如表 1 所示.

Table 1 Existing research of local differential privacy

表1 本地化差分隐私保护的研究方向

研究方向	示例
本地化差分隐私的扰动机制研究	W-RR[18]、 Compression[60]、 Distortion[62]
基于本地化差分隐私的单值频数发布	RAPPOR[19]、O-RAPPOR[22]、S-Hist[20]、PCE[37]、 k -RR[18]、O-RR[22]、 k -Subset[52,53]
基于本地化差分隐私的多值频数发布	RAPPOR-unknown[54]、 Harmony-frequency[25]、 LDPMiner[56]、 LoPub[57,58]
基于本地化差分隐私的均值发布	MeanEst[23,24]、 Harmony-mean[25]

由表 1 可知,本地化差分隐私技术的研究方向包括扰动机制的研究以及统计数据的发布.本地化差分隐私的扰动机制主要包括随机响应、信息压缩和扭曲两种.随机响应技术的扰动框架简洁直观,并且其扰动程度可直接量化,因此本地化差分隐私下的研究工作大都基于随机响应技术展开,包括针对离散型数据的频数发布和针对连续型数据的均值发布.频数发布形式包括列联表、直方图等,其中根据变量的数量不同分为单值频数发布和多值频数发布.基于本地化差分隐私的单值频数发布主要是通过编码-解码技术以及概率扰动技术发布属性候选值的频数来保护数据隐私,而多值频数发布则是在此基础上进一步利用采样技术和降维技术等提高了数据的可用性.目前,针对本地化差分隐私的均值发布研究工作还较少,其主要思想一般是在无偏估计的前提下对连续值进行离散化.

4 本地化差分隐私方法对比与分析

基于随机响应技术的扰动机制是当前研究热点,其相关工作主要集中在满足 ε-本地化差分隐私保护的算法研究.根据算法的查询类型不同,可将其分为两类:1)针对离散型数据的频数统计查询;2)针对连续型数据的均值统计查询.以下 4.1 和 4.2 节根据频数统计和均值统计两种不同的查询类型对方法进行分析和比较,4.3 节补充说明除随机响应技术外的其它扰动机制,最后 4.4 节对本地化差分隐私技术所呈现的实验特性进行分析.

4.1 基于本地化差分隐私的频数统计

频数发布是指针对离散型数据,返回给定约束条件下记录的数量,即计数查询的结果.中心化差分隐私保护下相应的数据发布形式包括列联表数据发布[41,42]、针对二值型数据集的列计数数据发布[43,44]、直方图数据发布[45]、图数据发布[46,47]等,这些数据发布形式都是基于计数查询进行的.例如以下表 2 所示的医疗数据集,进行统计后可得到表 3 所示的关于疾病的频数统计,该统计结果是众多统计分析、机器学习、数据挖掘方法的基础.

Table 2 Diseases of patients
表2 病人患病信息

用户姓名	疾病名称
Tom	艾滋病
Mike	哮喘, 白血病
Sally	乳腺癌, 高血压
Abel	糖尿病
Carl	糖尿病, 抑郁症, 高血压
Helen	糖尿病, 高血压
David	抑郁症

Table 3 Distribution of diseases
表3 疾病的频数发布

疾病名称	患病人数
艾滋病	1
哮喘	1
抑郁症	2
乳腺癌	1
高血压	3
白血病	1
糖尿病	3

然而,一旦将表 3 所示的频数统计结果发布出去,攻击者就可能根据已有的背景知识推测出表 2 中的某个用户所患疾病信息,导致用户隐私泄露.本地化差分隐私保护技术首先在每个用户端对个人数据进行隐私化处理,如 Tom 将以 40%的概率提交“艾滋病”,分别以 10%的概率提交其余六种疾病,如此一来,Tom 所提交的数据就具有了一定的随机性,从而保护了敏感信息.以下将基于本地化差分隐私的频数发布方法分为单值频数统计和多值频数统计进行分析.

4.1.1 单值频数统计

单值频数统计是指每个用户只发送一个变量取值的情形,用户将数据发送给数据收集者后,数据收集者根据已有的或统计得到候选值列表,统计其中每一个候选值的频数并进行发布.

RAPPOR[19]方法是单值频数统计的代表方法,其中变量的值以字符串的形式表示.假设总共有 n 个用户,第 i 个用户 U_i 对应某个敏感取值 $x_i \in \chi$,且 $|\chi| = k$,现在希望统计取值 $x_i (1 \le i \le k)$ 的频数,RAPPOR 方法图示如图 3 所示.

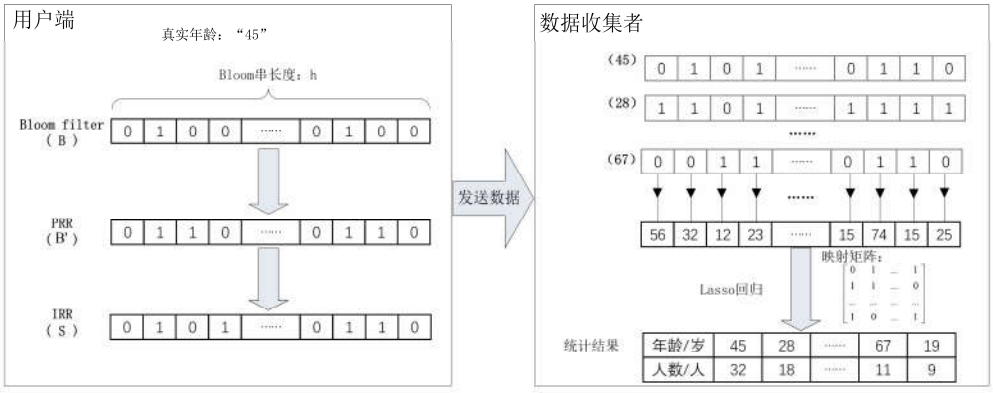


Fig.3 Process of RAPPOR

图3 RAPPOR 方法图示

χ 在图 3 中代表年龄属性,其中存在某个取值 x_i ="45",首先利用 Bloom Filter[48]技术将其表示成一个长度为 h 的向量 $B=\{0,1\}^h$,同时记录下字符串与 Bloom 串的映射关系矩阵,然后利用随机响应技术对向量 B 的每一个位进行扰动得到*永久性随机响应*(Permanent Randomized Response, PRR)结果 B',其中,扰动的方式按照以下公式进行, $f\in[0,1]$ 表示概率取值:

$$P(B_i'=x)=\begin{cases}0.5f, & x=1\\ 0.5f, & x=0\\ 1-f, & x=B_i\end{cases}$$

接着,再对向量 B' 的每一个位进行第二次扰动得到*瞬时性随机响应*(Instantaneous Randomized Response, IRR)结果 S,其中,第二次扰动的方式按照以下公式进行,其中 $p\in[0,1]$ 和 $q\in[0,1]$ 分别表示 B_i' 取值为 1 和 0 时置 S_i 为 1 的概率:

$$P(S_i=1)=\begin{cases}p, & if\ B_i'=1\\ q, & if\ B_i'=0\end{cases}$$

每个用户得到扰动结果 S 后,将其发送给第三方数据收集者,数据收集者统计每一位上 1 出现的次数并进行校正,然后结合映射矩阵通过 Lasso 回归方法[49]完成每个年龄值对应的频数统计.

RAPPOR 方法的统计误差来自两个方面:1)采用 Bloom Filter 技术进行编码,在解码过程中可能存在属性候选值冲突的问题,对于这一点,RAPPOR 通过调整相应的参数,有效地降低了候选值冲突带来的误差;2)随机响应技术对数据进行扰动产生的误差问题,其中数据扰动带来的渐近误差边界为 $O(\frac{k}{\varepsilon\sqrt{n}})$,这是算法本身固有的误差.

RAPPOR 方法的主要存在两个方面的缺陷:1)用户和数据收集者之间的传输代价比较高,即每个用户需要传输长度为 h 的向量给数据收集者;2)数据收集者需预先采集候选字符串列表,以进行频数统计.

针对 RAPPOR 方法第一个方面的缺陷,即通信代价高的问题,同样在单值情形下,S-Hist[20]方法中每个用户对字符串进行编码后,随机选择其中的一个比特位,利用随机响应技术进行扰动后,将其发送给数据收集者,因此大大降低了传输代价.同时,S-Hist 方法假设列表中字符串的候选值个数 k 超过了用户的数量 n,于是利用随机投影(Random Projection, RP)技术[50]将每个字符串表示成 $m=O(n)$ 维二值变量,即生成一个随机投影矩阵

$\Phi_{m\times k}$,其中每个元素的取值集合为$\left\{\frac{1}{\sqrt{m}},-\frac{1}{\sqrt{m}}\right\}$,以使得矩阵中每个列向量的内积为 1,而任意不同的两个列向量的内积为 0.矩阵如下所示:

$$\Phi_{m\times k}=\begin{bmatrix}\frac{1}{\sqrt{m}} & \frac{-1}{\sqrt{m}} & \cdots & \frac{1}{\sqrt{m}}\\ \frac{-1}{\sqrt{m}} & \frac{1}{\sqrt{m}} & \cdots & \frac{-1}{\sqrt{m}}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{1}{\sqrt{m}} & \frac{-1}{\sqrt{m}} & \cdots & \frac{1}{\sqrt{m}}\end{bmatrix}$$

S-Hist 方法对于频数统计的查询,其渐近误差边界为$O(\frac{\sqrt{\log k}}{\varepsilon\sqrt{n}})$,S-Hist 方法的主要优点在于其极大地降低通信代价,后续不少研究基于 S-Hist 方法进行改进并进行相关应用[56,37].文献[56]中将其与采样技术相结合用于 Heavy hitter[51]查询;文献[37]对 S-Hist 中隐私预算进行个性化设置,针对位置隐私的场景提出 PCE 方法,在保护用户位置信息的前提下统计不同区域中的用户数,其方法本身也是属于频数发布的范畴.

针对 RAPPOR 方法第二个方面的缺陷,即预先采集候选字符串列表的问题,Kairouz 等人基于变量取值未知的情形提出了 O-RAPPOR 方法[22].O-RAPPOR 基于 RAPPOR 方法的编码和解码方式,同时在此基础之上又引入了哈希映射和分组(Cohort)操作.对于每个字符串首先利用哈希函数进行一次值的映射,后续的扰动步骤直接对哈希值进行处理,而不关注字符串本身,如此便无需预知候选字符串的列表.其中,为了减小哈希映射中哈希值冲突带来数据可用性下降的问题,对字符串的集合进行了分组操作.O-RAPPOR 方法中,首先对不同分组内的字符串用一个不同的哈希函数进行映射,生成 Bloom Filter[48]串 $BLOOM_c^{(k)}$,然后通过 RAPPOR 方法对 $BLOOM_c^{(k)}$ 进行数据扰动.具体扰动方式与 RAPPOR 方法相同,此处不再赘述.

上述方法主要是考虑将变量取值编码成二值的形式,使其满足随机响应技术 W-RR[18]的要求,再利用该技术进行数据扰动.除此之外,Kairouz 等人提出一种梯度响应技术 k-RR[21],其主要克服了随机响应技术针对二值变量这一问题,对于变量中含有 $k(k>2)$ 个候选值的情况,可以直接进行随机响应.对于任意的输入 $R\in\chi$,其响应输出 $R'\in\chi$ 的方式如下公式所示:

$$P(R'\mid R)=\frac{1}{k-1+e^{\varepsilon}}\begin{cases}e^{\varepsilon}, & if\ R'=R\\ 1, & if\ R'\neq R\end{cases}$$

即,以 $\frac{e^{\varepsilon}}{k-1+e^{\varepsilon}}$ 的概率响应真实的结果,以 $\frac{1}{k-1+e^{\varepsilon}}$ 的概率响应其余 $k-1$ 个结果中的任意一种,使其满足 ε-本地化差分隐私.注意到,当 $k=2$ 时,上式即与随机响应 W-RR 的形式相同,因此,这是一种更为泛化的定义形式.

目前,RAPPOR 和 k-RR 为单值频数发布下的经典方法,以下从三个方面对二者进行简单比较:1)以 RAPPOR 为代表的方法中,其对变量的某一个取值进行编码、随机响应并解码,而 k-RR 则直接在变量的多个取值之间进行随机响应,其整体框架更加简洁;2)就不同的隐私预算而言,RAPPOR 和 k-RR 两种方法的性能表现出一定的差异性[22]:以隐私预算 $\varepsilon=\ln k$ 为界,RAPPOR 适用于隐私预算较高的情形,而 k-RR 在隐私预算较低的情形下则表现更优;3)从方法的理论误差来看,k-RR 的渐近误差边界为 $O(\frac{\sqrt{k^3}}{\varepsilon\sqrt{n}})$,因此 RAPPOR 方法具有更优的误差边界.

基于 k-RR 方法,Kairouz 等人针对变量取值未知的情形提出了 O-RR 方法[22].O-RR 方法是对 k-RR 方法的一个改进,在 k-RR 的基础上同样引入哈希映射和分组操作,其中通过哈希映射,使得方法不再关注字符串本身,

从而无需预先采集候选字符串列表,而通过分组操作则可进一步降低哈希映射值冲突的概率.直观地,对第 i 个字符串 $S_i \in \chi$ 指定一个分组 $c_i = \{1,2,...,C\}$,同一组内的字符串使用相同的哈希函数进行映射并分成 k 个子集,按照以下公式得到映射值,其中保证不同组的哈希函数之间无关联:

$$x_i = HASH_{c_i}(S_i) \bmod k = HASH_{c_i}^{(k)}(S_i)$$

这样便降低了哈希值冲突的概率,极端情况下,对于处在不同分组中的相同字符串,其映射结果发生冲突的概率仅为 $\frac{1}{k}$,这样有效提高数据的可用性.

得到映射的哈希值后,再利用 k -RR 方法的扰动方式进行响应,对于任意的字符串 S_i ,其响应输出 R' 的方式如下公式所示:

$$P(R' \mid R) = \frac{1}{C(k-1+e^{\varepsilon})} \begin{cases} e^{\varepsilon}, & if\ HASH_{c_i}^{(k)}(S_i) = R' \\ 1, & if\ HASH_{c_i}^{(k)}(S_i) \neq R' \end{cases}$$

注意到上式中较 k -RR 多了因子 C ,这是因为对每个字符串都指定了 C 个分组中的其中一个.方法 O-RR 与 k -RR 有相同的渐近误差边界 $O(\frac{\sqrt{k^3}}{\varepsilon\sqrt{n}})$,但运行时间上增加了分组和映射的开销.

上述方法均考虑扰动输出为单个取值的情形,即对任意的输入变量,进行数据扰动后仅输出一个取值,我们称其为一对一扰动.考虑一对多的情形,如字符串的模糊匹配,对于指定的输入,输出结果是一个集合.此时上述方法不再适用.基于此, k -Subset[52,53]方法被提出,具体如下.

对任意的输入 $R \in \chi$,输出 $R' \subseteq \chi$,满足 $|\chi| = k$, $|R'| = c$, k -Subset 方法的扰动输出如下所示:

$$P(R' \mid R) = \frac{1}{(ce^{\varepsilon} + k - c) \cdot \binom{k}{c}} \begin{cases} ke^{\varepsilon}, & if\ |R'| = c\ and\ R \in R' \\ k, & if\ |R'| = c\ and\ R \notin R' \end{cases}$$

其中,取值的组合总共包括 $\binom{k}{c} = \binom{k\text{-}1}{c\text{-}1} + \binom{k\text{-}1}{c}$ 种,第一项 $\binom{k\text{-}1}{c\text{-}1}$ 为满足 $R \in R'$ 的情形,第二项 $\binom{k\text{-}1}{c}$ 则为 $R \notin R'$ 的情形.通过上述扰动方式可以得到输出集合 R' 在不同取值下的频数值,进一步地,还可以通过该取值得到集合中每个元素的频数值,即每一个候选值的频数值.

显然,当输出 R' 中仅包含一个元素,即 d =1 时, k -Subset 方法将退化为 k -RR,即 k -RR 方法为 k -Subset 的一种特殊情形,相较于 k -RR, k -Subset 方法对扰动输出的定义是一种更为泛化的形式.

k -Subset 方法将输出扩展为集合的形式,这种一对多的扰动方式相较于 k -RR 的一对一扰动方式,有效降低了扰动过程中输入输出之间的匹配误差,相对于 k -RR 方法 $O(\frac{\sqrt{k^3}}{\varepsilon\sqrt{n}})$ 的渐近误差边界, k -Subset 的渐近误差边界为 $O(\frac{k}{\varepsilon\sqrt{n}})$,可见 k -Subset 提高了数据的发布精度.

以上介绍了本地化差分隐私下的频数统计方法.表 4 对上述方法的主要优缺点、通信代价以及渐近误差边界和计算开销等进行了对比分析,其中通信代价是指从每个用户处到数据收集方的数据传输的开销,这里我们近似认为通信代价与数据量成正比;渐近误差边界中, n 指总用户数, k 指属性候选值个数, h 表示 Bloom Filter 串的长度, c 表示 k -Subset 方法中输出集合的大小;计算开销是指数据收集者对用户数据进行统计时的计

算代价,分为“高”、“中”、“低”三个级别.

Table 4 Existing methods of single-valued frequency estimation under LDP

表4 本地化差分隐私下的单值频数统计方法对比分析

方法名称	主要优点	主要缺点	通信代价	渐近误差边界	计算开销	是否要求属性候选值已知
RAPPOR[19]	发布误差较小,数据可用性高	需考虑 Bloom filter 参数的设置问题	$O(h)$	$O(\frac{k}{\varepsilon\sqrt{n}})$	高;额外的回归计算	是
O-RAPPOR[22]	哈希技术隐藏了属性候选值列表	需考虑 Bloom filter 参数的设置问题	$O(h)$	$O(\frac{k}{\varepsilon\sqrt{n}})$	高;额外的回归计算	否
S-Hist[20]	采样技术降低了通信代价	查询精度不稳定,适用于属性候选值较多的情形	$O(1)$	$O(\frac{\sqrt{\log k}}{\varepsilon\sqrt{n}})$	中;额外的编码与字符串匹配	是
PCE[37]	个性化设置隐私参数	查询精度不稳定,适用于属性候选值较多的情形	$O(1)$	$O(\frac{\sqrt{\log k}}{\varepsilon\sqrt{n}})$	中;额外的编码与字符串匹配	是
k -RR[18]	无须编码和解码过程,简化数据扰动过程	隐私预算较低时数据可用性不高	$O(1)$	$O(\frac{\sqrt{k^3}}{\varepsilon\sqrt{n}})$	低; 仅涉及频数统计	是
O-RR[22]	哈希技术隐藏了属性候选值列表	隐私预算较低时数据可用性不高	$O(1)$	$O(\frac{\sqrt{k^3}}{\varepsilon\sqrt{n}})$	低; 仅涉及频数统计	否
k -Subset[52,53]	同时适应于列联表等统计查询	统计单值频数带来额外误差	$O(1)$	$O(\frac{k}{\varepsilon\sqrt{n}})$	中;额外根据集合频数计算元素频数	是

从表 4 可以看出,基于 RAPPOR 的方法可用性较高,但同时也带来较高的计算开销;基于 S-Hist 的方法极大降低了通信代价,但其采样过程却带来了一定的精度损失;基于 k -RR 的方法简化了数据扰动过程,同时也牺牲了一定的发布精度.

4.1.2 多值频数统计

多值频数统计是指每个用户发送多个变量取值的情形,用户将数据发送给数据收集者后,数据收集者根据已有的或统计得到候选值列表,统计其中每一个候选值的频数并进行发布.不同于单值频数统计问题,多值的情形需要考虑隐私预算的分割问题.直观地,可以将单值频数统计方法重复地用在多值情形中的每一个变量上,但如此将导致以下两个问题:1)需要根据变量的数量分割隐私预算,当变量较多时,数据可用性急剧降低;2)忽略了变量之间的关联关系,损失部分信息.

对于第一个问题,假设共有 d 个变量,S-Hist 方法对每个变量进行相同的处理,在分割隐私预算后,每个变量分配到的隐私预算为 $\frac{\varepsilon}{d}$,这将直接导致渐近误差边界和结果的方差增长 d 倍.以 RAPPOR 和 S-Hist 方法为例,对于 RAPPOR 方法,渐近误差边界由 $O(\frac{k}{\varepsilon\sqrt{n}})$ 上升至 $O(\frac{dk}{\varepsilon\sqrt{n}})$;对于 S-Hist 方法,其渐近误差边界由 $O(\frac{\sqrt{\log k}}{\varepsilon\sqrt{n}})$ 上升至 $O(\frac{d\sqrt{\log k}}{\varepsilon\sqrt{n}})$.同理,通信代价也将增长为原来的 d 倍.因此,直接将单值频数发布方法重复 d 次作为多值情形的频数发布方法,其在数据可用性和传输代价上均不可行.而对于第二个问题,变量之间的关联关系也是一种重要的信息,可以通过信息熵、互信息等指标来体现,独立处理每一个变量时,忽略了变量之间的关联关系,将损失其中蕴含的某些信息.

针对 RAPPOR 方法第二个方面的缺陷,即需要预先采集候选字符串列表,文献[54]提出一种改进方法,用于多值频数统计,我们称该方法为 RAPPOR-unknown.方法是基于 RAPPOR 实现的,每个用户对数据的扰动处理与 RAPPOR 一致,但针对 RAPPOR 中数据收集者需要预先采集候选字符串列表这一缺陷,RAPPOR-unknown 中基

于 n-gram[1]思想,利用图 3 中用户端的数据扰动方式得到扰动结果后,从字符串中抽取 r 个长度相同的子串,然后将扰动结果和子串的相关信息一起发送给第三方数据收集者,其中,数据的传输格式如下所示:

$$\left\{X', G_1', ..., G_r', g_1, ..., g_r\right\}$$

其中, X' 表示 RAPPOR 方法对字符串 x 的扰动结果,即 $X' = RAPPOR(x)$, $G_1', ..., G_r'$ 表示原字符串的 r 个子串, $g_1, ..., g_r$ 表示 $G_1', ..., G_r'$ 在原字符串中的位置.因此,关于隐私预算 ε 的分配,RAPPOR-unknown 将其 $(r+1)$ 等分,分别分配给扰动结果和 r 个子串的计算过程.

采集数据后,数据收集者通过共现(Co-occurrence)技术还原字符串,以得到字符串列表.针对多值的情形,RAPPOR-unknown 利用期望最大化(Expectation Maximization, EM)算法[55],估计多个变量的联合概率分布,以进行列联表的查询.

RAPPOR-unknown 是基于 RAPPOR 方法的一个改进,其渐近误差边界为 $O(\frac{k}{\varepsilon\sqrt{n}})$,但其通信代价却更高,为 $O(h)+O(r)$,这主要是因为除了传输扰动后的数据,还需要将子串及其位置信息一起传输给数据收集者用于统计候选值列表,比 RAPPOR 方法多出的 $O(r)$ 即表示子串及其位置信息的传输代价.显然,RAPPOR-unknown 不适合子串较多的情形,否则,不仅通信代价高,而且导致数据可用性降低.

虽然 S-Hist 方法极大地降低了通信代价,但在实际应用中,S-Hist 方法的准确性并不稳定,这主要是由于随机投影矩阵 $\Phi_{m\times k}$ 中每个元素的正负取值具有随机性,要使得任意两个列向量的内积为 0,其前提是字符串的候选值 k 足够大.鉴此,Nguyên 等人针对关系型数据提出 Harmony[25]方法,可用于连续型数据的均值统计和离散型数据的频数统计两种查询.本节介绍针对频数统计问题的方法,称其为 Harmony-frequency,5.2 节将介绍针对均值统计的方法,称其为 Harmony-mean 方法.Harmony-frequency 对随机投影矩阵进行了改进,用迭代的方式生成一个 $k\times k$ 的随机投影矩阵 $\Phi_{k\times k}$,其中保证了矩阵中任意两个列向量正交,即内积为 0.包含 d 个变量的情形下, Harmony-frequency 方法的渐近误差边界为 $O(\frac{\sqrt{d\log k}}{\varepsilon\sqrt{n}})$,相对于 S-Hist 方法 $O(\frac{d\sqrt{\log k}}{\varepsilon\sqrt{n}})$ 的渐近误差边界,Harmony-frequency 有更高的发布精度.

LDPMiner[56]是集值数据下的频数发布方法,其针对 Heavy hitter 查询.假设有 n 个用户,每个用户均包含 d 个项中的 l 个项,Heavy hitter 集合大小为 k' .LDPMiner 方法包括两个阶段:1)数据收集者采集数据,确定 Heavy hitter 集合并将其返回给用户;2)用户发送集合中 k' 个项所对应的数据.方法的整体框架如图 4 所示.

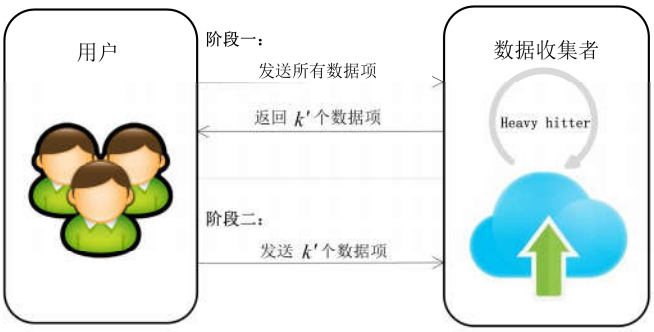


Fig.4 Framework of LDPMiner

图4 方法 LDPMiner 的整体框架

[1] n-gram 即字符串的长度为 n 的子串

LDPMiner 是基于 RAPPOR 和 S-Hist 方法实现的一个组合方法,考虑到每个用户需要发送多个数据项,通信代价较高,因此利用随机采样技术令每个用户只发送其中的一个数据项,然后利用 RAPPOR 方法和 S-Hist 方法来进行数据扰动,其中把采样技术和 RAPPOR 方法的组合叫做 sampling RAPPOR[56],而把采样技术和 S-Hist 方法的组合叫做 sampling SH[56].如图 4 所示,LDPMiner 方法包含以下两个阶段.

阶段一:sampling SH 方法.用户通过 sampling SH 方法将扰动后的数据发送给数据收集者,数据收集者统计每个数据项的频数,确定频数最高的 k' 个项的集合,并将该集合返回给用户.

阶段二:sampling RAPPOR 方法.用户通过 sampling RAPPOR 方法将阶段一返回的数据项所对应的数据再次发送给数据收集者,数据收集者统计每个数据项的频数,得到 k' 个数据项的频数.

LDPMiner 方法主要通过两个方面的数据处理提高数据可用性:1)通过采样技术避免了隐私预算在不同数据项上的分割;2)频繁项集合的确定和频数统计两个操作分离,缩小了需要扰动的数据项集合,从而增大了 k' 个数据项所分配到的隐私预算.

LoPub[57,58]是结合了 RAPPOR 方法和概率图模型的多值频数统计方法.整体思路包括三个步骤:1)用户端的数据扰动;2)数据收集者估计联合概率分布并进行数据降维;3)合成数据集.首先,每个用户通过 RAPPOR 方法对元组进行扰动后,将其发送给数据收集者,即利用 Bloom Filter 技术对数据集的每个属性的每个候选值进行编码,然后将元组中 d 个属性的候选值的编码串接在一起,利用随机响应技术对该串进行扰动,把得到的永久性随机响应结果发送给数据收集者.随后,类似于文献[59]中基于中心化差分隐私的高维数据发布方法,数据收集者对采集到的数据进行频数统计进而构建出马尔科夫网络(Markov network),利用属性之间的关联性得到极大团(Maximal clique),并将属性的联合概率分布以极大团的形式来表示,以此达到数据降维的目的.最后,通过联合概率分布重新合成一个数据集进行数据发布.

以上介绍了本地化差分隐私下的频数统计方法.表 5 对上述方法的主要优缺点、通信代价、渐近误差边界、隐私预算的分配等进行了对比分析,其中 d 表示变量的个数,h 表示 Bloom Filter 串的长度,r 表示 RAPPOR-unknown 方法中子串的个数,m 表示 S-Hist 方法中字符串的编码长度,l 表示 LoPub 方法中字符串的平均编码长度.

Table 5 Existing methods of multiple-valued frequency estimation under LDP

表5 本地化差分隐私下的多值频数统计方法对比分析

方法名称	主要优点	主要缺点	通信代价	渐近误差边界	计算开销	是否要求属性候选值已知	隐私预算 ε 的分配策略
RAPPOR-unknown[54]	无需预知属性候选值列表	需考虑 Bloom filter 参数的设置问题	$O(h)+O(r)$	$O(\frac{dk}{\varepsilon\sqrt{n}})$	高;额外的字符串匹配开销	否	根据子串个数分割 ε
Harmony-frequency[25]	发布误差较小,数据可用性高	采样技术带来精度下降	$O(m)$	$O(\frac{\sqrt{d\log k}}{\varepsilon\sqrt{n}})$	中;额外的编码与字符串匹配开销	是	通过采样技术,ε 全部分配给采样变量
LDPMiner[56]	发布误差小,数据可用性高	仅适用于 Heavy hitter 查询	$O(h)+O(m)$	$O(\frac{dk}{\varepsilon\sqrt{n}})$	高;两轮的字符串解码开销	是	ε 被划分到两个阶段
LoPub[57,58]	弱化高维度对精度的影响	高昂的通信代价和计算开销	$O(d*l)$	$O(\frac{dk}{\varepsilon\sqrt{n}})$	高;额外的概率图模型构建和推理开销	是	按属性分割 ε

4.2 基于本地化差分隐私的均值统计

本地化差分隐私下的均值统计,其主要思想是对个体值添加正向和负向的噪声,最终通过聚合大量的扰动结果以抵消其中的正负向噪声,从而使统计结果满足一定的可用性要求.

本地化差分隐私下的均值统计方法最早由 Duchi 等人提出[23,24],我们称该方法为 MeanEst,其主要思想是

将包含 n 个元组的 d 维数据集中的第 i 个元组 $t_i \in [-1,1]^d$ 按照一定的概率分布,并结合随机响应技术,转变成一个仅含二值变量的元组 $t_i^* \in \{-B,B\}^d$,同时保证最终的统计结果是一个无偏估计量.其中, B 的计算仅与隐私预算 ε 和数据维度 d 有关:

$$B=\begin{cases}\dfrac{2^d+C_d\cdot(e^\varepsilon-1)}{\binom{d-1}{(d-1)/2}\cdot(e^\varepsilon-1)}, & d\text{为奇数}\\ \dfrac{2^d+C_d\cdot(e^\varepsilon-1)}{\binom{d-1}{d/2}\cdot(e^\varepsilon-1)} & d\text{为偶数}\end{cases}$$

其中,

$$B=\begin{cases}2^{d-1}, & d\text{为奇数}\\ 2^{d-1}-\dfrac{1}{2}\binom{d}{d/2}, & d\text{为偶数}\end{cases}$$

首先,MeanEst 方法根据以下公式计算一个采样元组 $v=\{-1,1\}^d$,

$$P(v[A_j]=x)=\begin{cases}\dfrac{1}{2}+\dfrac{1}{2}t_i[A_j], & x=1\\ \dfrac{1}{2}-\dfrac{1}{2}t_i[A_j], & x=-1\end{cases}$$

其中, A_j 表示元组 v 中的第 j 个变量.

然后以 $\dfrac{e^\varepsilon}{1+e^\varepsilon}$ 的概率返回一个元组 $t_i^* \in [-B,B]^d$,使得 $t_i^*\cdot v>0$,以 $\dfrac{1}{1+e^\varepsilon}$ 的概率返回一个元组 $t_i^* \in [-B,B]^d$,使得 $t_i^*\cdot v<0$.

元组 t_i 的转换结果 t_i^* 被发送给数据收集者,数据收集者对 t_i^* 按不同维度进行均值统计,统计结果的渐近误差边界为 $O(\dfrac{\sqrt{d\log d}}{\varepsilon\sqrt{n}})$.同时,由 B 的计算公式可知,其值与数据集的维度 d 呈指数关系,当数据集的维度较高时,所需的时间代价和空间代价都比较高,使得 B 值的计算受到限制.因此该方法不适用于维度较高的数据集.

在此基础上,Harmony-mean[25]方法进一步简化了 MeanEst 方法,通过采样技术降低了数据的传输代价.Harmony-mean 方法中,对于第 i 个输入元组 $t_i \in [-1,1]^d$,其输出元组为 $t_i^* \in [-\dfrac{e^\varepsilon+1}{e^\varepsilon-1}d,\ 0,\ \dfrac{e^\varepsilon+1}{e^\varepsilon-1}d]^d$.具体来说,首先初始化元组 $t_i^* =< 0,0,...,0 >$,然后从 d 个数据维度中随机抽取一个维度 j,以一定的概率令其取值为 $\dfrac{e^\varepsilon+1}{e^\varepsilon-1}d$ 或 $-\dfrac{e^\varepsilon+1}{e^\varepsilon-1}d$,公式如下:

$$P(t_i^*[A_j]=x)=\begin{cases}\dfrac{t_i[A_j]\cdot(e^\varepsilon-1)+e^\varepsilon+1}{2e^\varepsilon+2}, & x=\dfrac{e^\varepsilon+1}{e^\varepsilon-1}d\\ \dfrac{e^\varepsilon+1-t_i[A_j]\cdot(e^\varepsilon-1)}{2e^\varepsilon+2}, & x=-\dfrac{e^\varepsilon+1}{e^\varepsilon-1}d\end{cases}$$

即,输出的元组中仅一个维度上的变量有相应的取值,而其它维度上变量的取值为 0.因此,Harmony-mean 方法

中通信代价为 MeanEst 方法的 $\frac{1}{d}$,而其渐近误差边界为 $O(\frac{\sqrt{d\log d}}{\varepsilon\sqrt{n}})$,两者具有相近的发布精度.

以上介绍了本地化差分隐私下的均值统计方法.表 6 对上述方法的主要优缺点、通信代价、渐近误差边界以及隐私预算的分配进行了对比分析,如下所示.

Table 6 Existing methods of mean value estimation under LDP

表6 本地化差分隐私下的均值统计方法对比分析

方法名称	主要优点	主要缺点	通信代价	渐近误差边界	计算开销	隐私预算 ε 的分配
MeanEst[23,24]	发布误差小,数据可用性高	时空复杂度高,仅适用于低维数据,且个体数据偏离原始数据的程度大	$O(d)$	$O(\frac{\sqrt{d\log d}}{\varepsilon\sqrt{n}})$	高; 需要遍历变量的所有组合	根据变量个数平均分配 ε
Harmony-mean[25]	时间复杂度低,且发布误差小,数据可用性高	个体数据偏离原始数据的程度大	$O(1)$	$O(\frac{\sqrt{d\log d}}{\varepsilon\sqrt{n}})$	低;仅涉及均值计算	通过采样技术, ε 全部分配给采样变量

4.3 基于信息压缩和扭曲的扰动机制

目前,随机响应是本地化差分隐私的主流扰动机制,上述频数统计和均值统计方法均基于此实现.此外,还有基于信息压缩和扭曲的扰动机制.为完整阐述本地化差分隐私的研究现状,本节补充说明基于信息压缩和扭曲的扰动机制.

文献[60]提出基于压缩输入域的扰动机制,我们称之为 Compression.每个用户对应一个 d 维的元组 $X=(x_1,x_2,...,x_d)$,其中 $X\in\chi$,称 χ 为输入域.对输入域进行压缩得到 $\hat{\chi}$,$|\hat{\chi}|\le|\chi|$,记 $\hat{\chi}$ 中的元组为 $\hat{X}=(\hat{x}_1,\hat{x}_2,...,\hat{x}_d)$.压缩过程保证元组 X 与 $\hat{X}$ 的对应关系满足本地化差分隐私保护,其中压缩率通过以下公式定义:

$$\rho=\frac{\log_2|\hat{\chi}|}{\log_2|\chi|}$$

对输入域进行的压缩处理会使得 X 与 $\hat{X}$ 之间产生一定的偏差,该偏差由以下公式定义:

$$L(X,\hat{X})=\frac{1}{d}\sum_{i=1}^{d}L(x_i,\hat{x}_i)$$

为了对 X 与 $\hat{X}$ 之间的偏差程度进行限制,引入了信息扭曲度 δ,对上述偏差的期望值进行约束,即:

$$\max_{P\in\mathcal{P}}\mathbb{E}_{P\times Q}[L(X,\hat{X})]\le\delta$$

其中,$\mathbb{E}_{P\times Q}[L(X,\hat{X})]=\sum_{x_i,x_i'}P(x_i)Q(\hat{x}_i\mid x_i)L(x_i,\hat{x}_i)$ 为偏差的数学期望,P 和 Q 表示 X 与 $\hat{X}$ 的分布.显然,最终把问题转化为了给定扭曲度 δ 和压缩率 ρ 下求解最小 ε 的凸优化问题,可通过二分法求解[61].其中,ε 表示隐私的保护程度,δ 表示数据统计结果的可用性,通过压缩率 ρ 将两者直接联系起来.

文献[62]提出基于信息扭曲的扰动机制,我们称之为 Distortion.该方法通过扰动函数 $Q(\hat{x}\mid x)$ 对 d 维元组 $X=(x_1,x_2,...,x_d)$ 进行扰动得到 $\hat{X}=(\hat{x}_1,\hat{x}_2,...,\hat{x}_d)$.与 Compression 方法相同,Distortion 首先定义信息扭曲度 δ 用于约束 X 与 $\hat{X}$ 的偏差,保证数据具有一定的可用性.

$$\max_{P\in\mathcal{P}}\mathbb{E}_{P\times Q}[L(X,\hat{X})]\le\delta$$

扰动函数则根据信息扭曲度 δ 的值来确定:

$$Q(\hat{x}\mid x)=\begin{cases}1-\delta, & \hat{x}=x\\ \dfrac{\delta}{k-1}, & \hat{x}\neq x\end{cases}$$

其中 $k=|x|$ 表示 x 的取值个数.该模型保证扰动过程满足 ε -本地化差分隐私,且 $\varepsilon=\log(k-1)+\log\dfrac{1-\delta}{\delta}$.

上述两种基于信息压缩和扭曲的扰动机制主要从信息损失的角度考虑输入和输出的关系,而基于随机响应技术的扰动机制则主要以一定的概率分布衡量输入和输出的关系.表 7 对几种扰动机制进行了比较.

Table. 7 Purtubation mechanisms under LDP

表7 本地化差分隐私下的扰动机制对比分析

扰动机制	主要优点	主要缺点	计算复杂度
W-RR[18]	扰动框架简洁直观,可扩展性强	仅适用于二值变量的扰动	低;仅涉及统计计算
Compression[60]	对信息损失的定量化	时空复杂度高,且通信代价高	高;额外的信息压缩率和扭曲度计算
Distortion[62]	对信息损失的定量化	仅适用于属性候选值个数较少的情形	高;额外的信息扭曲度计算

从表 7 看出,W-RR 模型由于其高度可扩展性成为了本地化差分隐私下主流的扰动机制.Compression 模型虽然对输入域进行了压缩,但是每个用户的传输代价并未减少,因此其通信代价很高.此外,Compression 仅适用于低维的情况,主要是因为输入与输出之间的匹配需计算笛卡尔乘积,当维度较高时,得到的匹配结果将呈指数式增长,将使得相应的时空复杂度过高.类似地,Distortion 模型中,隐私预算的设置依赖于属性候选值的个数,且随候选值个数增加而增加,因此当属性候选值个数较多时,隐私保护程度将下降.

4.4 本地化差分隐私技术的实验特性分析

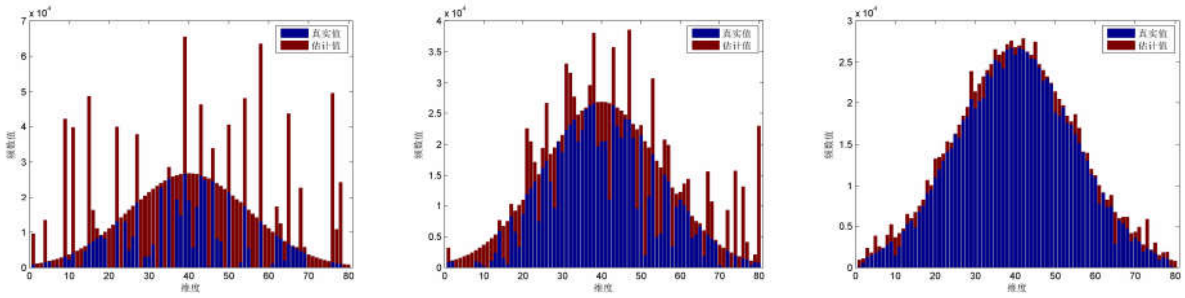
从定义来看,本地化差分隐私技术对数据的保护程度主要依赖于隐私预算 ε 的设定,因此 ε 的取值决定了隐私化处理后数据的可用性高低.但同时,本地化差分隐私保护技术对所处理数据集的数据量有一定要求,数据量也是影响数据可用性的一个重要因素.

本节针对本地化差分隐私技术的实验特性进行比较分析,主要包括两个方面:隐私预算 ε 对隐私保护程度的影响和数据量 N 对数据可用性的影响.下面结合上述经典方法对本地化差分隐私下的频数统计和均值统计进行分析,其中,利用 RAPPOR[19]方法进行频数统计,利用 Harmony-mean[25]方法进行均值统计.

4.4.1 隐私预算对数据可用性的影响

对隐私保护而言,隐私保护程度与数据可用性呈负相关,隐私保护程度高则数据可用性低,隐私保护程度低则数据可用性高.本地化差分隐私中,隐私保护的程度同样由参数 ε 决定,其通过控制随机响应技术输出真实值的概率值来控制数据的偏离程度,进而保护隐私.以下实验利用上述两种方法,比较不同 ε 值对统计结果造成的影响.其中,ε 包含以下三种取值:0.1, 0.5 以及 2.0.

首先,对于频数统计任务,我们模拟 10^6 个用户,每个用户发送一个字符串,字符串取值集合为 $\{V_1,V_2,...,V_{80}\}$. 频数统计的任务是根据数据收集者所采集的数据,统计每个字符串出现的频率,而真实的字符串频数值满足正态分布.不同隐私预算下的统计结果如图 5 所示,蓝色柱子高度表示真实频数值,红色柱子高度表示隐私化处理后的统计值.

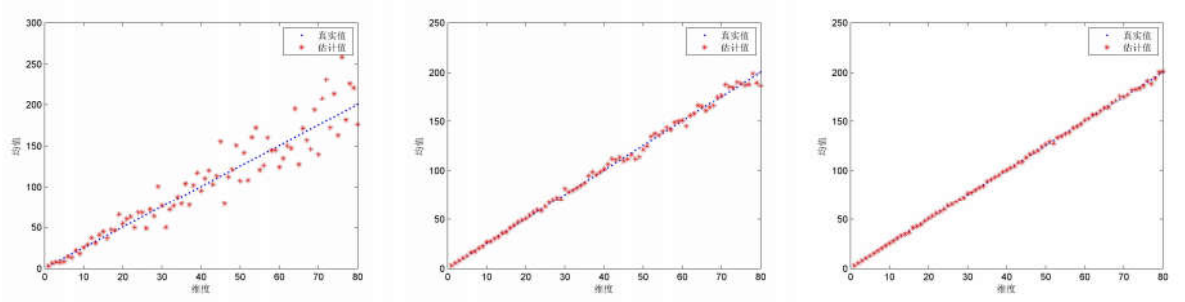


(a) ε=0.1 (b) ε=0.5 (c) ε=2.0

Fig.5 Frequency estimation under different privacy budget
图5 不同隐私预算下的频数统计结果

由图 5 可知,当给定较少的隐私预算,如 ε=0.1 时,统计结果偏离真实值的程度较大,而当给定较多的隐私预算,如 ε=2.0 时,统计结果则比较接近真实值.

对于均值统计任务,我们则是模拟多属性的情形.同样包括 10^6 个用户,每个用户对应一个 80 维的元组,其中每个元素都是从 [1,400] 上随机选取的整数,表示用户在该维度上的取值,同时使得不同维度上的均值满足线性分布.不同隐私预算下的统计结果如图 6 所示,蓝色点表示给定维度的真实均值,红色点表示隐私化处理后的估计值.



(a) ε=0.1 (b) ε=0.5 (c) ε=2.0

Fig.6 Mean value estimation under different privacy budget
图6 不同隐私预算下的均值统计结果

由图 6 可知,当给定较少的隐私预算时,均值统计结果偏离真实值的程度较大,而当给定较多的隐私预算时,均值的统计结果十分接近真实值.

对于本地化差分隐私技术的保护机制而言,不同的隐私预算 ε 直接决定了随机响应技术中用于响应真实结果的概率 p , ε 越大则 p 越大,即用户以更高的概率响应真实结果,因此无论对于频数统计还是均值统计,都提高了数据的可用性.即,对于同一个数据集,隐私预算的大小直接决定了数据可用性的高低,隐私预算大则数据可用性高,隐私预算小则数据可用性低.直观地,以 W-RR[18]为例, ε 与 p 的函数关系如图 7 所示.

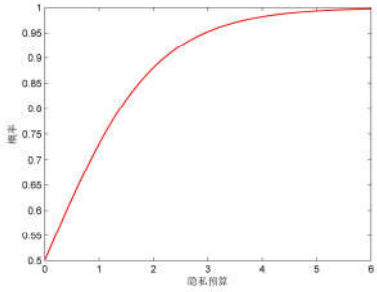


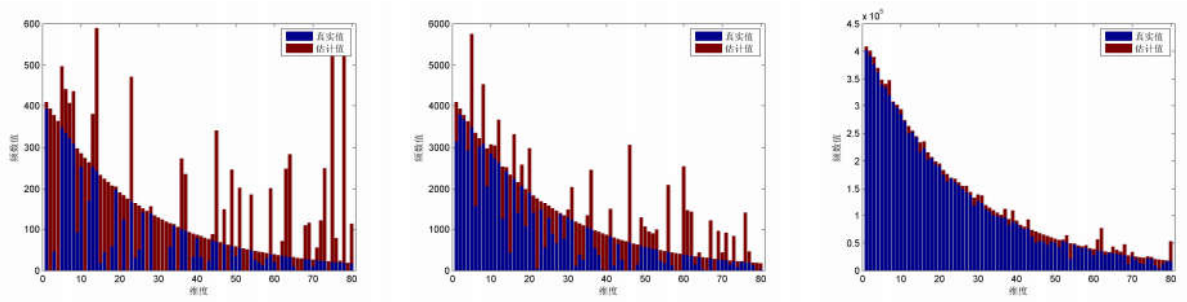
Fig.7 Functional relation between ε and p

图7 ε 与 p 的函数关系

4.4.2 数据量对数据可用性的影响

本地化差分隐私技术通过把数据中的正向扰动噪声和负向扰动噪声相互抵消来保证数据可用性,这就必然要求统计数据量足够大.本节就数据量大小对统计结果可用性高低的影响进行分析,其中,设置隐私预算 ε=1.0 ,数据量 N 设置三个不同的数量级: 10^4 , 10^5 和 10^6 .

对于频数统计任务,为避免不同数据分布可能带来的影响,本节中模拟生成字符串频数满足指数分布的数据集,同样模拟 10^6 个用户,每个用户发送一个字符串,字符串取值集合为 $\{V_1,V_2,...,V_{80}\}$.数据量按照上述 N 的三个不同取值,生成三个模拟数据集,其候选值的频数值满足指数分布.不同数量下的频数统计结果如图 8 所示,蓝色柱子高度表示真实频数值,红色柱子高度表示隐私化处理后的估计值.



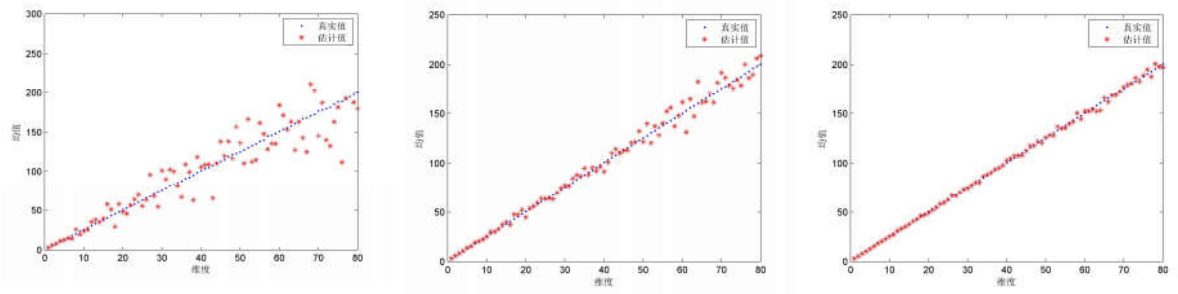
(a) $N=10^4$ (b) $N=10^5$ (c) $N=10^6$

Fig.8 Frequency estimation under different data volume

图8 不同数据量下的频数统计结果

由图 8 可知,在给定相同的隐私预算(ε=1.0)下,即对于相同的隐私保护程度,当所统计的数据量较小,如 $N=10^4$ 时,统计结果与真实值的偏差较大,而当所统计的数据量较大,如 $N=10^6$ 时,统计结果则比较接近真实值.

对于均值统计任务,生成四个模拟数据集,分别包括 10^4 、 10^5 和 10^6 个元组,数据集其它特征与 4.4.1 节中均值统计任务所描述相同.不同数据量下的统计结果如图 9 所示, 蓝色点表示给定维度的真实均值,红色点表示隐私化处理后的统计均值.



(a) $N=10^4$ (b) $N=10^5$ (c) $N=10^6$

Fig.9 Mean value estimation under different data volume

图9 不同数据量下的均值统计结果

由图 9 可知,同样在给定相同的隐私预算(ε=1.0)下,当所统计的数据量较小时,统计结果与真实值的偏差较大,而当所统计的数据量较大时,统计结果十分接近真实值.

在本地化差分隐私技术中,无论是频数统计还是均值统计,当给定相同隐私预算时,用于统计的数据量大小决定了数据可用性高低,数据量大则统计结果的可用性高,数据量小则统计结果的可用性低.这一特性主要从大数定律[63]的角度来解释.大数定律阐释了在相同条件下重复多次试验,其随机事件的频率近似于它的概率.本地化差分隐私中,每个用户分别以一定的概率来响应真实值,然后数据收集者基于相同的概率对结果进行统计分析.由于响应具有随机性,统计结果往往偏离真实值.当数据量较小时,由于事件的频率与概率之间差值较大,统计结果和真实值之间的偏离程度更为明显,而当数据量较大时,事件的频率接近于概率,因此统计结果和真实值之间的差异性就变得很小.

5 未来研究挑战

目前,本地化差分隐私还是一个新的研究领域,现有研究主要集中在两个方面:1)理论上,设计满足本地化差分隐私的保护机制;2)方法上,对频数和均值两种统计结果进行保护.然而,大数据时代下,数据的复杂性、多样性和大规模性等将带来新的大数据隐私风险[64].因此,结合现有研究,我们认为本地化差分隐私保护技术还有很多挑战性问题亟待解决,以下从三个方面进行阐述.

5.1 复杂数据类型的本地化差分隐私保护

信息技术时代带来了数据规模的爆炸式增长,而数据之间的关联性也使得数据类型更加复杂.显然,在隐私保护问题上,这为隐私保护技术带来极大的挑战.目前,本地化差分隐私保护技术的研究主要还是针对简单的数据类型,例如对包含一个或多个属性的关系数据和集值数据,进行频数统计或均值统计.然而,简单的数据类型对于当下空前的数据分析需求而言,还远远不够.下面以键值对数据(Key-value database)和图数据(Graph database)为例,分析其隐私风险并说明该数据类型对本地化差分隐私技术的挑战.

键值对是一种常用的数据类型,中心化差分隐私技术对键值对数据[65]的隐私保护,其思想一般是将相同“键”所对应的“值”进行统计,并添加噪声.然而实际上,我们认为,在本地化差分隐私场景下,除了“值”,“键”本身也具有一定的敏感性,因为数据收集者可能通过“键”重新定位某个用户,从而获取敏感信息.因此,对于键值对数据,需要同时对“键”和“值”进行隐私化处理,且仍需保证“键”和“值”之间的对应性.

图数据则是一种更为复杂的数据类型,其根据图结构定义了结点、边和属性,存储相应信息,例如,社交网络数据、路网数据等.中心化差分隐私技术针对图数据的隐私保护,包括子图计数[66]、结点数据发布[67]等,其主要难点在于图的结构特点使得查询的全局敏感性极高,从而使得噪声过大.基于本地化差分隐私技术的图数据发布,虽不存在敏感性过大的问题,但由于每个用户对数据的扰动过程相互独立,数据收集者如何根据扰动后的数据构建可用性高的图结构,即如何保证原始数据之间的关联性是一大挑战.况且,现实中很多图数据是稀疏的,

这进一步加大了还原数据关联性的难度.

5.2 不同查询和分析任务的本地化差分隐私保护

执行不同的查询任务,进行相应的数据分析,是为了从数据中抽取有价值的规则和知识.目前,本地化差分隐私下的研究工作所针对的查询任务主要包括两类简单的聚集查询:离散型数据下的计数查询和连续型数据下的均值查询,并且数据的扰动方式一般根据查询类型而定,二者紧密耦合,使得相应扰动方式仅能支持特定的查询任务.然而,面对大数据下各种复杂的查询和分析任务,仅上述两类基本的聚集查询还远远不够.基于本地化差分隐私保护的数据分析,目前也仅有文献[25]讨论了线性回归、Logistic 回归和 SVM 分类三种分析任务.

以数据分析中的聚类分析和频繁模式挖掘为例,聚类分析是为了把性质相近的数据归入一类,典型应用如客户群体发现、社区发现等;频繁模式挖掘的目的则是找出数据集中的频繁项集,进而发现模式规律,典型应用如搜索日志分析、购买行为分析等.聚类分析和频繁模式挖掘都是数据分析中的常用方法,其中涉及的查询包括计数查询、均值查询、范围查询和最值查询等.本地化差分隐私下,不仅要求能够支持不同的查询类型,还要求扰动后的数据能够同时支持多种不同的查询.与中心化差分隐私不同,本地化差分隐私通过抵消添加在数据中的正向和负向噪声来得到比较准确的统计结果,但就单条数据记录而言,通常扰动前后数据的偏差较大,这也进一步加大了针对不同查询的难度.

因此,我们认为,针对不同查询和分析任务的本地化差分隐私保护技术主要考虑以下三个方面的问题:1)提供对除计数查询和均值查询外的多种查询方式的支持;2)数据扰动方式与查询类型的解绑,使得扰动后的数据能够同时支持多种查询;3)提高数据分析结果的可用性.

5.3 基于本地化差分隐私的高维数据发布

隐私数据发布问题下,当数据维度不高时,现有大多数方法均可以取得较好的统计结果,如中心化差分隐私下基于数据立方体的方法[68],基于多权重机制的方法[69,70]和基于学习模型的方法[71]等.然而,当属性个数较多时,即对于数据维度较高的数据集,现有的方法将遇到瓶颈[72,73],因此,数据的高维度给差分隐私技术带来极大挑战.对于本地化差分隐私保护技术而言,高维数据不仅带来数据规模变大和信噪比降低两个方面的影响,而且还将增加通信代价,并且通信代价根据不同的扰动机制随数据维度的增加呈线性增长或指数增长,而通信代价的增加将直接为本地化差分隐私技术的应用带来限制.

目前,针对高维数据的发布,主要是利用属性划分的思想,在满足本地化差分隐私的基础上,将高维数据的联合概率分布分解为多个低维的边缘概率分布的形式,以多个边缘概率通过某种推理机制近似估计联合概率分布.其中的关键步骤是对两两属性之间的关联性判断.数据中存在 d 个属性时,对应的关联性存在 $\binom{d}{2}$ 种,这就意味着需要把有限的隐私预算进行 $\binom{d}{2}$ 次分割,高维数据下,这势必带来很大噪声,使得推理结果的准确性大大降低.因此,还需要辅以其它的方式进行降维,如对属性进行聚类或分组等.目前仅有文献[57,58]讨论了本地化差分隐私下的高维数据发布问题,但其只考虑了数据收集者如何依据属性之间的相关性进行降维处理,因此高维数据下通信代价问题依旧存在.为了降低高维数据带来的通信代价,现有方法一般是在不同维度上利用采样技术进行降维[20,56],然而,采样技术依然不可避免地导致数据可用性下降.因此在该类问题上,还需考虑通信代价和数据可用性之间的平衡关系.

综上所述,我们认为,本地化差分隐私下的高维数据发布主要考虑三个方面的问题:1)如何在一定隐私预算内衡量属性之间的关联性,从而进行降维处理;2)如何设计推理模型,最小化边缘分布到联合分布的近似误差,提高数据可用性;3)如何控制高维数据在用户和数据收集者之间的通信代价.

6 结束语

大数据时代个人数据高度敏感,如何防止隐私信息泄露是当前面临的重大挑战.本地化差分隐私是继中心化差分隐私后新兴的隐私保护模型,其打破了中心化差分隐私中关于可信第三方数据收集者的的假设,在用户端对数据进行隐私化处理.目前,本地化差分隐私保护技术是隐私保护领域的研究热点,本文对其研究成果进行总结和分析,综述了本地化差分隐私保护技术的研究现状,总结该技术在频数统计和均值统计中的应用,并进行实验特性分析.最后,本文就现有研究工作和现实需求进行探讨,结合二者提出未来研究挑战.总之,本地化差分隐私保护技术还是一个新兴研究领域,仍有诸多关键问题需要进行深入而细致的研究.

References:

[1] Dwork C: Differential Privacy. ICALP (2) 2006: 1-12.

[2] Dwork C, Lei J. Differential privacy and robust statistics. Proceedings of the forty-first annual ACM symposium on Theory of computing. ACM, 2009: 371-380. [doi: 10.1145/1536414.1536466]

[3] Smith A. Privacy-preserving statistical estimation with optimal convergence rates. Proceedings of the forty-third annual ACM symposium on Theory of computing. ACM, 2011: 813-822. [doi: 10.1145/1993636.1993743]

[4] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. PODS. 1998, 98: 188.

[5] Machanavajjhala A, Kifer D, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 3. [doi: 10.1109/ICDE.2006.1]

[6] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007: 106-115. [doi: 10.1109/ICDE.2007.367856]

[7] Kasiviswanathan S P, Lee H K, Nissim K, Raskhodnikova S, Smith A. What can we learn privately? . Foundations of Computer Science (FOCS), 2008 49th Annual IEEE Symposium on. IEEE, 2008: 531-540.

[8] Duchi J C, Jordan M I, Wainwright M J. Local privacy and statistical minimax rates. Foundations of Computer Science (FOCS), 2013 54th Annual IEEE Symposium on. IEEE, 2013: 429-438. [doi: 10.1109/FOCS.2013.53]

[9] Howe J. Crowdsourcing: How the power of the crowd is driving the future of business[M]. Random House, 2008.

[10] Li G, Wang J, Zheng Y, Franklin MJ. Crowdsourced data management: A survey. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(9): 2296-2319. [doi: 10.1109/TKDE.2016.2535242]

[11] Wu S, Wang X, Wang S, Zhang Z, Tung AK. K-anonymity for crowdsourcing database. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2207-2221. [doi: 10.1109/TKDE.2013.93]

[12] Varshney L R, Vempaty A, Varshney P K. Assuring privacy and reliability in crowdsourcing with coding. Information Theory and Applications Workshop (ITA), 2014. IEEE, 2014: 1-6. [doi: 10.1109/ITA.2014.6804213]

[13] Mao J, Jain A K. Artificial neural networks for feature extraction and multivariate data projection. IEEE Transactions on Neural Networks, 1995, 6(2): 296-317. [doi: 10.1109/72.363467]

[14] Finn R L, Wright D, Friedewald M. Seven types of privacy[M]//European data protection: coming of age. Springer Netherlands, 2013: 3-32.

[15] Erkin Z, Franz M, Guajardo J, Katzenbeisser S, Lagendijk I, Toft T. Privacy-preserving face recognition. International Symposium on Privacy Enhancing Technologies Symposium. Springer Berlin Heidelberg, 2009: 235-253.

[16] Qin Z, Yan J, Ren K, Chen C W, Wang C. Towards efficient privacy-preserving image feature extraction in cloud computing. Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 497-506. [doi: 10.1145/2647868.2654941]

[17] Ren K. Privacy-preserving image processing in cloud computing. Chinese Journal of Network and Information Security, 2016,(01):12-17.

[18] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965, 60(309): 63-69.

[19] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. ACM, 2014: 1054-1067. [doi: 10.1145/2660267.2660348]

[20] Bassily R, Smith A. Local, private, efficient protocols for succinct histograms. Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing. ACM, 2015: 127-135. [doi: 10.1145/2746539.2746632]

[21] Kairouz P, Oh S, Viswanath P. Extremal mechanisms for local differential privacy. Advances in neural information processing systems. 2014: 2879-2887.

[22] Kairouz P, Bonawitz K, Ramage D. Discrete distribution estimation under local privacy. Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA. 2016:2436-2444.

[23] Duchi J C, Jordan M I, Wainwright M J. Local privacy, data processing inequalities, and statistical minimax rates. arXiv preprint arXiv:1302.3203, 2013.

[24] Wainwright M J, Jordan M I, Duchi J C. Privacy aware learning. Advances in Neural Information Processing Systems. 2012: 1430-1438.

[25] Nguyên T T, Xiao X, Yang Y, Hui S C, Shin H, Shin J. Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy. arXiv preprint arXiv:1606.05053, 2016.

[26] McSherry F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009: 19-30. [doi: 10.1145/1559845.1559850]

[27] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference. Springer Berlin Heidelberg, 2006: 265-284.

[28] McSherry F, Talwar K. Mechanism design via differential privacy. Foundations of Computer Science(FOCS), 2007 48th Annual IEEE Symposium on. IEEE, 2007: 94-103. [doi: 10.1109/FOCS.2007.66]

[29] Wang Q, Zhang Y, Lu X, Wang Z, Qin Z, Ren, K. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy. IEEE Transactions on Dependable and Secure Computing, 2016. [doi: 10.1109/TDSC.2016.2599873]

[30] Zhang X, Chen R, Xu J, Meng X, Xie Y. Towards accurate histogram publication under differential privacy. Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2014: 587-595.

[31] Su S, Tang P, Cheng X, Chen R, Wu Z. Differentially private multi-party high-dimensional data publishing. Data Engineering (ICDE), 2016 IEEE 32nd International Conference on. IEEE, 2016: 205-216. [doi: 10.1109/ICDE.2016.7498241]

[32] Yaroslavtsev G, Cormode G, Procopiuc C M, Srivastava D. Accurate and efficient private release of datacubes and contingency tables. Data Engineering (ICDE), 2013 IEEE 29th International Conference on. IEEE, 2013: 745-756. [doi: 10.1109/ICDE.2013.6544871]

[33] Zhang T, Zhu Q. Dynamic Differential Privacy for ADMM-Based Distributed Classification Learning. IEEE Transactions on Information Forensics and Security, 2017, 12(1): 172-187. [doi: 10.1109/TIFS.2016.2607691]

[34] Abadi M, Chu A, Goodfellow I, McMahan H B, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016: 308-318.[doi: 10.1145/2976749.2978318]

[35] Bun M, Steinke T, Ullman J. Make up your mind: The price of online queries in differential privacy. Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2017: 1306-1325.

[36] Yuan G, Yang Y, Zhang Z, Hao Z. Convex Optimization for Linear Query Processing under Approximate Differential Privacy. arXiv preprint arXiv:1602.04302, 2016.

[37] Chen R, Li H, Qin A K, Kasiviswanathan S P, Jin H. Private spatial data aggregation in the local setting. Data Engineering (ICDE), 2016 IEEE 32nd International Conference on. IEEE, 2016: 289-300. [doi: 10.1109/ICDE.2016.7498248]

[38] Zhang X, Meng X. Differential Privacy in Data Publication and Analysis. Chinese Journal of Computers, 2014, (04):927-949.

[39] Kifer D, Machanavajjhala A. No free lunch in data privacy. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011: 193-204. [doi: 10.1145/1989323.1989345]

[40] Yang B, Sato I, Nakagawa H. Bayesian differential privacy on correlated data. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015: 747-762. [doi: 10.1145/2723372.2747643]

[41] Zhang J, Cormode G, Procopiuc C M, Srivastava D, Xiao X. Privbayes: Private data release via bayesian networks. Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014: 1423-1434. [doi: 10.1145/2588555.2588573]

[42] Qardaji W, Yang W, Li N. Priview: practical differentially private release of marginal contingency tables. Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014: 1435-1446. [doi: 10.1145/2588555.2588575]

[43] Kellaris G, Papadopoulos S. Practical differential privacy via grouping and smoothing. Proceedings of the VLDB Endowment. VLDB Endowment, 2013, 6(5): 301-312. [doi: 10.14778/2535573.2488337]

[44] Day W Y, Li N. Differentially private publishing of high-dimensional data using sensitivity control. Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security. ACM, 2015: 451-462. [doi: 10.1145/2714576.2714621]

[45] Xu J, Zhang Z, Xiao X, Yang Y, Yu G, Winslett M. Differentially private histogram publication. The VLDB Journal, 2013, 22(6): 797-822.

[46] Zhang J, Cormode G, Procopiuc C M, Srivastava D, Xiao, X. Private release of graph statistics using ladder functions. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015: 731-745.

[47] Day W Y, Li N, Lyu M. Publishing graph degree distribution with node differential privacy. Proceedings of the 2016 International Conference on Management of Data. ACM, 2016: 123-138. [doi: 10.1145/2723372.2737785]

[48] Bloom B H. Space/time trade-offs in hash coding with allowable errors. Communications of the ACM, 1970, 13(7): 422-426.[doi: 10.1145/362686.362692]

[49] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 1996: 267-288.

[50] Blum A, Ligett K, Roth A. A learning theory approach to noninteractive database privacy. Journal of the ACM (JACM), 2013, 60(2): 12. [doi: 10.1145/2450142.2450148]

[51] Hsu J, Khanna S, Roth A. Distributed private heavy hitters. Automata, Languages, and Programming, 2012: 461-472.

[52] Wang S, Huang L, Wang P, Nie Y, Xu H, Yang W, Li X, Qiao C. Mutual Information Optimally Local Private Discrete Distribution Estimation. arXiv preprint arXiv:1607.08025, 2016.

[53] Ye M, Barg A. Optimal Schemes for Discrete Distribution Estimation under Locally Differential Privacy. arXiv preprint arXiv:1702.00610, 2017.

[54] Fanti G, Pihur V, Erlingsson Ú. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. Proceedings on Privacy Enhancing Technologies, 2016, 2016(3): 41-61.

[55] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1977: 1-38.

[56] Qin Z, Yang Y, Yu T, Khalil I, Xiao X, Ren K. Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016: 192-203. [doi: 10.1145/2976749.2978409]

[57] Ren X, Yu C M, Yu W, Yang S, Yang X, McCann J A, Yu P S. LoPub: High-Dimensional Crowdsourced Data Publication with Local Differential Privacy. arXiv preprint arXiv:1612.04350, 2016.

[58] Ren X, Yu C M, Yu W, Yang S, Yang X, McCann J. High-Dimensional Crowdsourced Data Distribution Estimation with Local Privacy. Computer and Information Technology (CIT), 2016 IEEE International Conference on. IEEE, 2016: 226-233. [doi: 10.1109/CIT.2016.57]

[59] Chen R, Xiao Q, Zhang Y, Xu J. Differentially private high-dimensional data publication via sampling-based inference. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 129-138. [doi: 10.1145/2783258.2783379]

[60] Xiong S, Sarwate A D, Mandayam N B. Randomized requantization with local differential privacy. Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016: 2189-2193. [doi: 10.1109/ICASSP.2016.7472065]

[61] Boyd S, Vandenberghe L. Convex optimization[M]. Cambridge university press, 2004.

[62] Sarwate A D, Sankar L. A rate-distortion perspective on local differential privacy. Allerton. 2014: 903-908. [doi: 10.1109/ALLERTON.2014.7028550]

[63] Newey W K, McFadden D. Large sample estimation and hypothesis testing. Handbook of econometrics, 1994, 4: 2111-2245.

[64] Meng X, Zhang X. Big data privacy management. Journal of Computer Research and Development, 2016, 52(2):265-281.

[65] Viswanath B, Kiciman E, Saroiu S. Keeping information safe from social networking apps. Proceedings of the 2012 ACM workshop on Workshop on online social networks. ACM, 2012: 49-54. [doi: 10.1145/2342549.2342561]

[66] Karwa V, Raskhodnikova S, Smith A, Yaroslavtsev G. Private analysis of graph structure. ACM Transactions on Database Systems (TODS), 2014, 39(3): 22. [doi: 10.1145/2611523]

[67] Hay M, Li C, Miklau G, Jensen D. Accurate estimation of the degree distribution of private networks. Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on. IEEE, 2009: 169-178. [doi: 10.1109/ICDM.2009.11]

[68] Ding B, Winslett M, Han J, Li Z. Differentially private data cubes: optimizing noise sources and consistency. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011: 217-228. [doi: 10.1145/1989323.1989347]

[69] Hardt M, Rothblum G N. A multiplicative weights mechanism for privacy-preserving data analysis. Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on. IEEE, 2010: 61-70. [doi: 10.1109/FOCS.2010.85]

[70] Hardt M, Ligett K, McSherry F. A simple and practical algorithm for differentially private data release. Advances in Neural Information Processing Systems. 2012: 2339-2347.

[71] Thaler J, Ullman J, Vadhan S. Faster algorithms for privately releasing marginals. International Colloquium on Automata, Languages, and Programming. Springer Berlin Heidelberg, 2012: 810-821.

[72] Aggarwal C C. On randomization, public information and the curse of dimensionality. Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007: 136-145. [doi: 10.1109/ICDE.2007.367859]

[73] Aggarwal C C. Privacy and the dimensionality curse. Privacy-Preserving Data Mining, 2008: 433-460.

附中文参考文献:

[17] 任奎. 云计算中图像数据处理的隐私保护.网络与信息安全学报,2016,(01):12-17.

[38] 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护.计算机报,2014,(04):927-949.

[64] 孟小峰, 张啸剑. 大数据隐私管理.计算机研究与发展,2016,52(2):265-281.