计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150304
53(5):1106-1117,2016

差分隐私下一种精确直方图发布方法

张啸剑[1] 邵 超[1] 孟小峰[2]
[1](河南财经政法大学计算机与信息工程学院 郑州 450002)
[2](中国人民大学信息学院 北京 100872)
(xjzhang82@ruc.edu.cn)

Accurate Histogram Release under Differential Privacy

Zhang Xiaojian[1], Shao Chao[1], and Meng Xiaofeng[2]
[1](*College of Computer & Information Engineering, Henan University of Economics and Law, Zhengzhou* 450002)
[2](*School of Information, Renmin University of China, Beijing* 100872)

Abstract Grouping-based differentially private histogram release has attracted considerable research attention in recent years. The trade-off between approximation error caused by the group's mean and Laplace error due to Laplace noise constrains the accuracy of histogram release. Most existing methods based on grouping strategy cannot efficiently accommodate the both errors. This paper proposes an efficient differentially private method, called DiffHR (differentially private histogram release) to publish histograms. In order to boost the accuracy of the released histogram, DiffHR employs Metropolis-Hastings method in MCMC (Markov chain Monte Carlo) and the exponential mechanism to propose an efficient sorting method. This method generates a differentially private histogram by sampling and exchanging two buckets to approximate the correct order. To balance Laplace error and approximation error efficiently, a utility-driven adaptive clustering method is proposed in DiffHR to partition the sorted histogram. Furthermore, the time complexity of the clustering method is $O(n)$. DiffHR is compared with existing methods such as GS, AHP on the real datasets. The experimental results show that DiffHR outperforms its competitors, and achieves the accurate results.

Key words differential privacy; histogram release; grouping; Laplace error; approximation error

摘 要 基于分组的差分隐私直方图发布得到了研究者的广泛关注,组均值造成的近似误差与噪音造成的拉普拉斯误差之间的均衡直接制约着直方图发布精度.针对现有基于分组的直方图发布方法难以有效兼顾近似误差与拉普拉斯误差的不足,提出了一种满足差分隐私的精确直方图发布方法 DiffHR (differentially private histogram release);通过分析直方图桶计数序列的排序有助于提升发布精度,利用 Markov 链蒙特卡洛(Markov chain Monte Carlo, MCMC)方法中的 Metropolis-Hastings 技术与指数机制,提出了一种有效排序方法,通过不断置换 2 个随机选取的桶以逐渐逼近正确排序;基于抽样排序后的直方图,提出了一种基于懒散分组下界的自适应贪心聚类方法,该方法的时间复杂度为 $O(n)$,并

收稿日期:2015-04-17;**修回日期**:2015-10-19
基金项目:国家自然科学基金项目(61502146,61379050,U1404605,61202285);国家"八六三"高技术研究发展计划基金项目(2013AA013204);河南省科技厅基础与前沿技术研究项目(152300410091);河南省教育厅高等学校重点科研项目(16A520002);河南财经政法大学校重大研究课题(201426)
This work was supported by the National Natural Science Foundation of China (61502146, 61379050, U1404605, 61202285), the National High Technology Research and Development Program of China (863 Program) (2013AA013204), the Basic Research Program of Henan Science and Technology Department (152300410091), the Key Research Program of the Higher Education of Henan Educational Committee (16A520002), and the Key Research Program of Henan University of Economics and Law (201426).

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

且可有效均衡近似误差与拉普拉斯误差.DiffHR,GS,AHP 方法在真实数据上的实验结果表明,其发布精度上优于同类算法.

关键词 差分隐私;直方图发布;分组;拉普拉斯误差;近似误差

中图法分类号 TP392

快速而又准确地获取数据分布的梗概是数据分析与查询的主要任务.直方图是近似估计数据分布的主要技术之一,该技术使用分箱技术近似描述数据分布信息,将数据集按照某种属性划分成不相交的桶,每个桶由频度或者计数表示其特征.直方图的发布通常用来支持聚集查询、范围计数查询以及数据挖掘等应用.然而,如果直接发布直方图而不给予隐私保护,桶的真实计数会泄露个人的敏感信息.图 1 为 HIV 疾病监控中心确诊患者的年龄分布,其中 40 岁患者为 20 人.如果攻击者知道了除 Alice 之外其他 19 人的年龄情况,攻击者利用图 1 中的直方图可以推理出 Alice 患了 HIV 疾病,进而泄露了 Alice 的个人隐私.

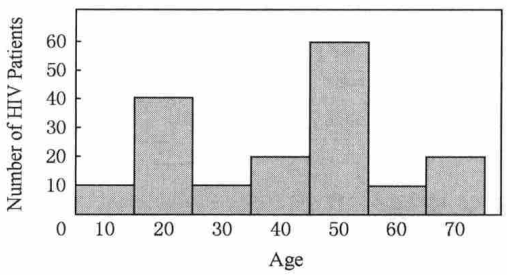


Fig. 1 Original histogram under age.

图 1 年龄分布下的原始直方图

为了阻止直方图统计信息带来的隐私泄露,在发布之前,需要对其进行隐私保护处理.目前差分隐私[1-4]已经成为一种新的隐私保护模型,基于该模型出现了多种直方图发布方法.其中,基于数据相关(data dependent)的分组方法是差分隐私下直方图发布的主要技术,该技术利用组平均值近似估计每个原始计数,例如图 2 给出图 1 的 3 个分组结果 $\{C_1,C_2,C_3\}$,其中 C_1 的 $\{10,40,10\}$ 计数被其均值 20 代替.分组操作的优点在于能够比较准确地响应范围查询.然而,由于均值与拉普拉斯机制实现隐私保护,不可避免地产生均值造成的近似误差与拉普拉斯噪音造成的拉普拉斯误差.文献[5-9]描述了目前基于分组的差分隐私直方图发布常用方法,这些方法通常在固定分组个数的基础上利用有损压缩、V-优化直方图以及等宽划分技术减少误差对发布结果的影响,然而这些方法存在 3 个问题.

1) 分组时只是考虑局部近邻的桶计数(例如文献[5-7]),而无法度量全局计数的近邻关系,这种类型的分组无法均衡近似误差与拉普拉斯误差,导致所发布直方图的精度比较低.例如,在图 1 直方图中,尽管桶 10 与桶 30 的计数值一样,它们不能被单独放在一个分组中,除非桶 20 也被放在该分组中.这通常会导致很大的近似误差.

2) 分组时大都未考虑桶计数的全局顺序性,进而无法聚集全局相似的计数.例如,如果对图 1 直方图进行排序,则桶 10,30,60 由于计数值相同而被分到同一组,进而可以减少近似误差带来的影响.虽然文献[8]采用了排序,但是其采取固定分组策略却很难均衡近似误差与拉普拉斯误差;同样文献[9]采用简单的噪音排序也会导致不准确的结果.

3) 文献[5-9]中分组方法的时间复杂度大都是 $O(kn^2)$ 或者 $O(n^2)$,例如 AHP,GS 方法.然而当直方图维度很大时,这些方法的分组效率比较低.

总而言之,目前还没有一个行之有效的方法兼顾直方图发布的近似误差与拉普拉斯误差.为此,本文提出了一种融合抽样排序与全局化分组的方法 DiffHR(differentially private histogram release),在满足差分隐私的情况下,采用 Metropolis-Hastings 技术对直方图进行抽样处理,并实施自适应贪心聚类,实现近似误差与拉普拉斯之间的均衡.

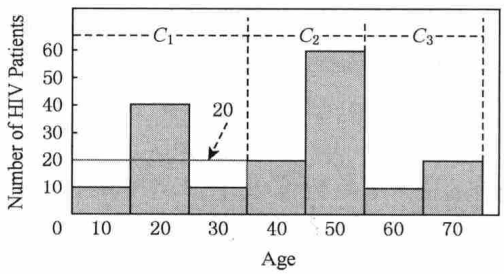


Fig. 2 Histogram after grouping.

图 2 分组后的直方图

本文主要贡献如下:

1) 为了精确发布直方图,提出了一种满足 ε-差分隐私的方法 DiffHR,该方法利用基于 Markov 链蒙特卡洛(Markov chain Monte Carlo, MCMC)方法中 Metropolis-Hastings 技术与指数机制对直方

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图进行抽样处理,通过不断置换 2 个随机选取的桶以逼近直方图的正确排序;

2) 为了有效均衡近似误差与拉普拉斯误差,提出了一种自适应的贪心聚类方法,该方法在满足差分隐私的前提下,采用懒散分组下界有效处理大维度直方图发布,且其时间复杂度为 $O(n)$;

3) 理论分析了 DiffHR 方法满足 ε-差分隐私,通过真实数据集上的实验分析展示该方法在兼顾高可用性和准确性的同时,优于同类方法.

1 相关工作

差分隐私拥有极强的保护能力和依赖严谨的统计模型,这些特点使其得到广泛研究和应用. 当前学术界的应用着重关注差分隐私保护下的计数查询[10]、直方图发布[5]、子图查询[11]、高维数据发布[12]、关联数据查询发布[13]、机器学习[14]与数据挖掘[15]等. 鉴于直方图在数据分析中的重要性,研究者已经提出了大量基于差分隐私的直方图发布方法. Dwork 等人[1]提出了拉普拉斯机制,成为后续研究的基石. Xiao 等人[16]提出了名为 Privelet 的基于小波的直方图发布技术. Privelet 首先对输入数据进行小波变换,然后对变换后的频域数据添加多重对数噪声. Hay 等人[17]指出约束推理可以作为一种后置处理方法来改善直方图的查询准确性. Li 等人[10]提出了矩阵机制. 给定一组查询,矩阵机制首先生成另外一组被称为查询策略的查询,并在其之上添加拉普拉斯噪音,最后通过查询策略的加噪结果来估算给定查询的值. Rastogi 和 Nath[18]指出静态时间序列数据可以通过只保留前 k 个傅里叶系数来以较小的近似误差换取较大的拉普拉斯噪音降幅. 2012 年 Yuan 等人[19]提出了低秩机制,它通过计算给定负荷矩阵的低秩近似来回答批次查询. 上述方法大都属于数据无关(data independent)的方法,优势在于能够给出严谨的数据效用理论下限;但在实际数据上,这一系列方法均无法获得理想数据可用性和效率,这一缺点引出了数据相关方法的研究.

所有数据相关的研究中最有希望的当属基于分组的方法. Xu 等人[5]提出了 NoiseFirst 和 StructureFirst 算法. 对于给定的分组个数,NoiseFirst 先对每个桶添加独立拉普拉斯噪音,之后利用 V-优化直方图构造方法在加噪后的直方图上寻找分组;StructureFirst 则通过不断应用指数机制来确定不同组之间的分界. 文献[5]主要缺陷在于只考虑相对于非隐私的 V-优化直方图的近似误差,而忽略了拉普拉斯噪音带来的误差,此外其要求分组数目已知的假设也通常难以满足. Acs 等人[6]改进了文献[5]中要求给定分组个数的缺陷,设计了一种分级的二分法算法,利用指数机制动态确定分组,并通过大量实验证实基于分组的方法可以取得比数据无关方法更好的准确性. 应该指出,分组仅仅提供"本地化"的聚类,通常无法找到最优聚类,故而无法找到近似误差和拉普拉斯误差间的最佳平衡来最小化整体误差. Li 等人[7]提出了一种基于分组策略的直方图方法 DAWA,然而,该方法同样没有采用排序和"全局化"聚类方法. Kellaris 等人[8]考虑一个稍微不同的问题设置:一个用户可以重复出现在多个桶中. 他们首先利用采样方法对直方图进行排序,然后将所有桶分成大小相同的组. 然而,选取固定的组大小通常无法很好地平衡近似误差和拉普拉斯误差,从而直接影响最终的发布精度. Zhang 等人[9]提出了一种全局化聚类方法 AHP,该方法取得了比文献[5-8]中较好的发布结果;然而,AHP 在进行排序时,只是采用简单的噪音处理方式,而这种方式通常导致不准确的排序结果. 因此,本文针对上述数据相关方法存在的缺陷,提出了一种精确直方图方法 DiffHR 来均衡近似误差和拉普拉斯误差.

2 定义与理论基础

2.1 差分隐私

差分隐私要求数据库中任何一个用户的存在都不应显著地改变任何查询的结果,从而保证了每个用户加入该数据库不会对其隐私造成危险. 传统的隐私保护模型如 k-anonymity[20]、l-diversity[21]等通常对攻击背景知识与攻击模型给出特定的假设,而现实中这些假设并不成立. 相比于传统保护模型,差分隐私保护模型具有 2 个显著的特点:1)不依赖于攻击者的背景知识;2)具有严谨的统计学模型,能够提供可量化的隐私保证. 本文所关心的是在发布直方图时,如何使得所发布直方图满足差分隐私.

定义 1. 设 $H=\{H_1, H_2, \cdots, H_n\}$ 为原始直方图桶计数序列,其中 H_i 表示第 i 个桶的计数. $H'=\{H_1, H_2, \cdots, H_{r-1}, H_r \pm 1, H_{r+1}, \cdots, H_n\}$,$H$ 与 H' 相差一个计数,则二者互为近邻关系.

结合 H 与 H' 给出 ε-差分隐私的形式化定义,如定义 2 所示:

定义 2. 给定一个直方图发布方法 A,$Range(A)$ 为 A 的输出范围,若方法 A 在 H 与 H' 上任意

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

输出结果 $\widetilde{H}$($\widetilde{H}\in Range(A)$)满足下列不等式,则 A 满足 ε-差分隐私.

$$Pr[A(H)=\widetilde{H}]\leqslant \exp(\varepsilon)\times Pr[A(H')=\widetilde{H}], \quad (1)$$

其中,ε 表示隐私预算,其值越小则算法 A 的隐私保护程度越高.

从定义 2 可以看出 ε-差分隐私限制了任意一个记录对算法 A 输出结果的影响.实现差分隐私保护需要噪音机制的介入,拉普拉斯与指数机制是实现差分隐私的主要技术.而所需要的噪音大小与其响应查询函数 f 的全局敏感性密切相关.

定义 3. 设 f 为某一个查询,且 $f:H\rightarrow\mathbb{R}^d$,$f$ 的全局敏感性为

$$\Delta f=\max_{H,H'}\| f(H)-f(H')\|_1, \quad (2)$$

其中,$\mathbb{R}$ 为映射的实数空间,d 为 f 的查询维度.

文献[1]提出的拉普拉斯机制可以取得差分隐私保护效果,该机制利用拉普拉斯分布产生噪音,进而使得发布方法满足 ε-差异隐私,如定理 1 所示.

定理 1[1]. 设 f 某一个查询函数,且 $f:H\rightarrow\mathbb{R}^d$,若方法 A 符合式(3),则 A 满足 ε-差分隐私.

$$A(H)=f(H)+[Lap(\Delta f/\varepsilon)]^n, \quad (3)$$

其中,$Lap(\Delta f/\varepsilon)$ 为相互独立的拉普拉斯噪音变量,噪音量大小与 Δf 成正比,与 ε 成反比.因此,查询 f 的全局敏感性越大,所需的噪音越多.

文献[22]提出的指数机制主要处理抽样算法的输出为非数值型的结果.该机制的关键技术是如何设计打分函数 $u(H,H_i)$.设 A 为指数机制下的某个隐私方法,则 A 在打分函数作用下的输出结果为

$$A(H,H_i)=\left\{H_i:|Pr[H_i\in\mathcal{H}]\propto\exp\left(\frac{\varepsilon u(H,H_i)}{2\Delta u}\right)\right\}, \quad (4)$$

其中,Δu 为打分函数 $u(H,H_i)$ 的全局敏感性,$\mathcal{H}$ 为采用算法的输出域.由式(4)可知,H_i 的打分函数越高,被选择输出的概率越大.

定理 2[22]. 对于任意一个指数机制下的算法 A,若 A 满足式(4),则 A 满足 ε-差分隐私.

2.2 直方图发布误差

给定一个原始直方图桶计数序列 $H=\{H_1,H_2,\cdots,H_n\}$,假设通过分组之后形成 k 个分组 $C=\{C_1,C_2,\cdots,C_k\}$,每个组 $C_i=(|C_i|,\overline{C}_i)$ 覆盖 $|C_i|$ 个桶计数,$\overline{C}_i$ 表示组 C_i 的计数平均值,即:

$$\overline{C}_i=\sum_{H_j\in C_i}H_j/|C_i|.$$

然而,在采用分组技术发布满足 ε-差分隐私的直方图 $\widetilde{H}=\{\widetilde{H}_1,\widetilde{H}_2,\cdots,\widetilde{H}_n\}$ 时,2 种误差却无法避免,分别为平均值引起的近似误差 AE(approximate error)和拉普拉斯噪音引起的拉普拉斯误差 LE(Laplace error).给定一个添加噪音之后的分值 C_i,它所携带的误差为

$$err(C_i)=AE(C_i)+LE(C_i), \quad (5)$$

其中,$LE(C_i)$ 为采用拉普拉斯分布的方差度量,$AE(C_i)$ 为采用和方差度量.$LE(C_i)$ 与 $AE(C_i)$ 可表示为

$$LE(C_i)=2(\Delta f/\varepsilon)^2.$$

$$AE(C_i)=\sum_{H_j\in C_i}(H_j-\overline{C}_i)^2=\sum_{H_j\in C_i}(H_j)^2-\frac{1}{|C_i|}\left(\sum_{H_j\in C_i}H_j\right)^2.$$

因此,发布整个直方图 $\widetilde{H}$ 所期望的误差为

$$err(\widetilde{H})=E(\|H-\widetilde{H}\|_2^2)=\sum_{C_i\in C}(AE(C_i)+LE(C_i)). \quad (6)$$

从直方图发布总体误差的角度看,分组的好坏直接影响着近似误差与拉普拉斯误差之间的均衡.分组个数值越大,近似误差越小,而累积的拉普拉斯误差很大;反之,拉普拉斯误差越小,而造成近似误差很大.因此,本文需要解决的具体问题是,如何均衡近似误差与拉普拉斯误差,同时使得总体的发布误差 $err(\widetilde{H})$ 最小.

3 差分隐私下精确直方图发布方法

3.1 基于分组的原则

基于第 1 节相关工作分析,在设计新的基于分组的直方图发布方法时需要考虑的基本原则包括:

1) 针对全局分组问题,算法需要在满足差分隐私的条件下尽量把全局相似的计数值分到一组;

2) 针对近似误差与拉普拉斯误差之间的均衡问题,方法应从"全局化"自适应分组出发.

若实现原则 1,最直接有效的方法是对 H 进行排序,这样相似的计数均会紧密排列.然而,如果直接基于 H 进行排序后分组,则会破坏 ε-差分隐私[8].如何在满足差分隐私下准确排序是本文的第一研究难点.最简单的方法是对直方图每个桶添加独立的拉普拉斯噪音,根据加噪的计数值排序[8-9],然而此方法会导致不准确的排序.因此,本文采用 Metropolis-Hastings 抽样技术[23] 与指数机制使全局相似的计数值尽可能分到同一个组;针对原则 2,在不给定组个数以及不固定组大小的情况下,本文设计了一种自适应地贪心聚类分组方法.基于上述

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分析,提出了兼顾上述 2 种原则的直方图发布方法 DiffHR,如算法 1 所示.

3.2 DiffHR 方法实现

DiffHR 方法的具体描述如下:

算法 1. DiffHR 方法.

输入:直方图 $H=\{H_1,H_2,\cdots,H_n\}$、隐私预算 ε;

输出:满足差分隐私的直方图 $\widetilde{H}$.

① $\varepsilon=\varepsilon_1+\varepsilon_2$;

② $\hat{H}=Metropolis\text{-}Hastings\text{-}Sort(H,\varepsilon_1)$;
/* 对 H 进行抽样处理. 详见 3.2.1 节 */

③ $C=Greedy\text{-}Clustering(\hat{H})$; /* 对 $\hat{H}$ 进行贪心聚类分组. 详见 3.2.2 节 */

④ for $i=1$ to n do

⑤ $\quad \overline{C}_i=\sum_{H_j\in C_i,C_i\in C} H_j/|C_i|$;

⑥ end for

⑦ for $i=1$ to n do

⑧ $\quad \widetilde{H}_j=\overline{C}_i+Lap(1/\varepsilon_2)/|C_i|$; /* 对每一组的均值添加拉普拉斯噪音 */

⑨ end for

⑩ return $\widetilde{H}=\{\widetilde{H}_1,\widetilde{H}_2,\cdots,\widetilde{H}_n\}$.

该方法包括 2 个主要步骤:Metropolis-Hastings 抽样技术与贪心聚类. 给定一个原始直方图计数序列 H,给定部分隐私预算 ε_1,利用 Metropolis-Hastings 抽样技术与指数机制对 H 进行排序处理(行②);而算法的关键步骤是抽样处理后的聚类操作(行③),其目的是为了降低每个分组本身的误差 $err(C_i)$;最后,求每个簇的均值,利用剩余的隐私预算 ε_2 为各个均值添加拉普拉斯噪音(行④~⑨).

为了深入理解 DiffHR 方法在分组过程中所引入的误差,给出以下定理.

定理 3. 给定任意一个分组 C_i,则 C_i 本身具有的误差为 $err(C_i)=\sum_{H_j\in C_i}(H_j-\overline{C}_i)^2+\frac{2}{|C_i|(\varepsilon_2)^2}$,其中,$\overline{C}_i$ 为分组 C_i 的均值,$\sum_{H_j\in C_i}(H_j-\overline{C}_i)^2$ 为近似误差,$2/|C_i|(\varepsilon_2)^2$ 为拉普拉斯误差.

证明. 根据直方图发布误差可知:$\overline{C}_i=\sum_{H_j\in C_i}H_j/|C_i|$,而为 $\overline{C}_i$ 添加拉普拉斯噪音的大小为 $\frac{Lap(1/\varepsilon_2)}{|C_i|}$,则 $err(C_i)$ 的期望值可表示为

$$err(C_i)=E\Big(\sum_{H_j\in C_i}\Big(H_j-\overline{C}_i-\frac{Lap(1/\varepsilon_2)}{|C_i|}\Big)\Big)=\sum_{H_j\in C_i}(H_j-\overline{C}_i)^2-2E\Big(\sum_{H_j\in C_i}(H_j-\overline{C}_i)\frac{Lap(1/\varepsilon_2)}{|C_i|}\Big)+E\Big(\sum_{H_j\in C_i}\Big(\frac{Lap(1/\varepsilon_2)}{|C_i|}\Big)^2\Big)=\sum_{H_j\in C_i}(H_j-\overline{C}_i)^2+E\Big(\sum_{H_j\in C_i}\Big(\frac{Lap(1/\varepsilon_2)}{|C_i|}\Big)^2\Big)=\sum_{H_j\in C_i}(H_j-\overline{C}_i)^2+\frac{2}{|C_i|(\varepsilon_2)^2}.$$

证毕.

3.2.1 抽样处理方法 Metropolis-Hastings-Sort

基于排序后的直方图进行分组操作,则可以把全局性相似的桶计数分到同一个组中,如图 3 右图所示,分组 C_1 中 3 个桶{10,30,60}的计数为 10 且均值为 10,分组 C_2 中 2 个桶{40,70}的计数为 20 且均值为 20,进而使得分组 C_1 与 C_2 的近似误差为 0.

然而,如果直接基于原始直方图 H 中真实桶计数进行排序(例如图 3 右图所示),会导致个人隐私泄露. 因此,本文通过 MCMC 把直方图 H 模拟成 Markov 链,然后利用基于指数机制的 Metropolis-Hastings 抽样技术来近似逼近 H 的正确排序.

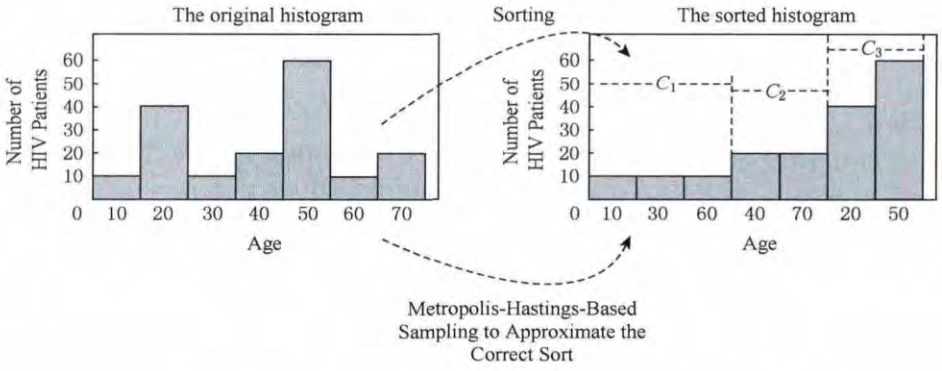


Fig. 3 Sorted-based histogram after grouping.

图 3 基于排序分组后的直方图

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1) 基于指数机制的 Metropolis-Hastings 抽样技术

根据指数机制的定义可知,直方图中任意一个桶被抽样的概率可以表示为

$$\pi(H_i)=\frac{\exp\left(\frac{\varepsilon_1 u(H,H_i)}{2\Delta u}\right)}{\sum_{H_j\in\mathcal{H}}\exp\left(\frac{\varepsilon_1 u(H,H_j)}{2\Delta u}\right)}. \tag{7}$$

为了采用 MCMC 中的 Metropolis-Hastings 抽样技术,首先给出相关描述信息.

定义 4. 给定一个 Markov 链$\{X_t:t=0,1,\cdots\}$,则其定义为

$$P(X_{t+1}=H_j\mid X_t=H_i,X_{t-1}=H_{i-1},\cdots)=P(X_{t+1}=H_j\mid X_t=H_i).$$

该定义说明下一个状态只取决于当前的状态.

设 X 为一个随机变量,其在状态空间 $\mathcal{H}=\{H_1,H_2,\cdots,H_{|\mathcal{H}|}\}$ 中进行取值. Metropolis-Hastings 方法根据 Markov 链定义在空间 $\mathcal{H}$ 中抽样出一个不可约的、非周期的 Markov 链,并使其服从平稳概率分布 $\pi(X)$. 基于 Metropolis-Hastings 的 Markov 链构造过程,实际上是状态空间的转换过程,而状态转换由转换概率 P 决定.

$$P(H_i,H_j)=P(X_{t+1}=H_j\mid X_t=H_i)=Q(H_i,H_j)\alpha_{H_iH_j}, \tag{8}$$

其中,$Q(H_i,H_j)$为建议概率分布,$\alpha_{H_iH_j}$表示接受概率,即表示状态 H_i 以概率 $\alpha_{H_iH_j}$ 跳转到状态 H_j,以$(1-\alpha_{H_iH_j})$的概率保持原来状态.

$$\alpha_{H_iH_j}=\min\left\{1,\frac{\pi(H_j)Q(H_j,H_i)}{\pi(H_i)Q(H_i,H_j)}\right\}. \tag{9}$$

采用式(7)中的 $\pi(H_i)$代替接收概率式(9)中的 $\pi(H_i)$,则 $\alpha_{H_iH_j}$ 重新表述为

$$\alpha_{H_iH_j}=\min\left\{1,\frac{\exp\left(\frac{\varepsilon_1 u(H,H_j)}{2n\Delta u}\right)Q(H_j,H_i)}{\exp\left(\frac{\varepsilon_1 u(H,H_i)}{2n\Delta u}\right)Q(H_i,H_j)}\right\}. \tag{10}$$

根据式(10)可知,剩余的 2 个问题分别是如何设计打分函数 $u(H,H_j)$与建议概率分布 $Q(H_i,H_j)$.

打分函数 $u(H,H_j)$在指数机制中至关重要,我们希望那些打分高的桶计数被抽样出来. 给定直方图 H 与当前状态 H_i,采用抽样状态 H_j 与当前状态 H_i 的距离作为打分函数,如式(11)所示:

$$u(H,H_j)=-|H_j-H_i|. \tag{11}$$

根据式(7)与式(11)可知,距离当前状态 H_i 越近的桶计数越容易被 Metropolis-Hastings 方法抽样. $u(H,H_j)$的敏感度为 1,因为改变 H 的桶中任意一个计数,最多影响 $u(H,H_j)$的变化为 1.

同样,构造一个合理的建议概率分布 $Q(H_i,H_j)$对 Metropolis-Hastings 抽样效果也至关重要. 因此,根据不同状态之间的距离来设计 $Q(H_i,H_j)$. 给定一个距离阈值 θ(实验中设置 $\theta=50$),可以求出与当前状态 H_i 近邻的状态集合 $S_1(H_i)=\{H_j:||H_j-H_i|\leqslant\theta\}$. 假设 $S_1(H_i)$中的状态是被等概率抽样,则 $Q(H_i,H_j)=1/S_1(H_i)$. 若 H_i 的距离不满足 θ 的状态称之为 H_i 的非近邻状态,该集合为 $S_2(H_i)=\{H_j:||H_j-H_i|>\theta\}$. 同理,$Q(H_i,H_j)=1/S_2(H_i)$.

2) Metropolis-Hastings-Sort 算法

基于上述分析,给出 Metropolis-Hastings-Sort 方法的具体实现细节.

算法 2. Metropolis-Hastings-Sort 算法.

输入:直方图 $H=\{H_1,H_2,\cdots,H_n\}$、隐私预算 ε_1;

输出:逼近正确排序的直方图 $\hat{H}$.

① for $i=1$ to n do

② 随机选择一个桶计数作为种子计数;

③ 时刻 i 的 Markov 链的状态为 H_i,且 $\hat{H}=\hat{H}\cup H_i$,抽取样本 H_j;

④ 根据阈值 θ,计算从当前桶计数 H_i 到近邻桶 H_j 的跳转概率 $Q(H_i,H_j)$;

⑤ 以接收概率 $\alpha_{H_iH_j}$ 接收桶 H_j,其中 $\alpha_{H_iH_j}$ 如式(10)所示;

⑥ if 抽样 Markov 链收敛到 $\pi(X)$ then

⑦ $\hat{H}=\hat{H}\cup H_j$

⑧ 比较 H_i 与 H_j 之间的大小,并在 $\hat{H}$ 中置换位置;

⑨ end if

⑩ end for

⑪ return $\hat{H}=\{\hat{H}_1,\hat{H}_2,\cdots,\hat{H}_n\}$.

Metropolis-Hastings-Sort 算法根据转移概率 $Q(H_i,H_j)$与跳转概率 $\alpha_{H_iH_j}$ 决定当前状态 H_i 是否跳转到状态 H_j. 一旦抽样 Markov 链收敛到目标分布 $\pi(X)$,即可获得逼近正确排序的直方图 $\hat{H}$. 本文利用 Gelman-Rubin 提出的方差比率方法[24]验证 Markov 链是否收敛. 该方法根据不同的初值,同时产生 m 条 Markov 链,每条链长度为直方图维度 n,则 Markov 链间的方差比率可以表示为 $R=\sqrt{\frac{n-1}{n}W+\frac{1}{n}B/W}$. 其中,$W$ 为 Markov 链内的方差均值,B 为链间的方差均值. 如果 R 趋近 1,则说明

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

基于 Metropolis-Hastings 的 MCMC 收敛. 本文实验中 $m=30$,基于 3 种数据集,R 大都 200 次迭代后趋近 1.

根据 Metropolis-Hastings-Sort 方法得到 $\hat{H}$ 后,如何对其进行分值并具有较高的效率具有很大挑战.

3.2.2 基于懒散分组下界的方法 Greedy-Clustering

根据排序后的直方图 $\hat{H}=\{\hat{H}_1,\hat{H}_2,\cdots,\hat{H}_n\}$,在不给定分组个数 k 值的情况下,Greedy-Clustering 方法从左至右贪婪自适应地划分 $\hat{H}$. 在贪婪划分过程中,判断每个组的形成条件至关重要. 假设已经形成的当前某个组 $C_i=\{\hat{H}_l,\hat{H}_{l+1},\cdots,\hat{H}_{j-1}\}$ 且误差为 $err(C_i)$,而对于下一个桶 $\hat{H}_j$ 是否要加入组 C_i,应给出相应的判断条件. 如果 $\hat{H}_j$ 加入组 C_i 使得 $err(C_i)$ 降低,则 $\hat{H}_j$ 与组 C_i 合并在一起,否则创建一个新组. 因此,分 2 种情况判断是否形成新的分组.

1) 如果 $\hat{H}_j$ 与 C_i 合并,对于新组 $\hat{H}_j\cup C_i$ 所形成的误差为

$$err(\hat{H}_j\cup C_i)=AE(\hat{H}_j\cup C_i)+\frac{2}{(|C_i|+1)(\varepsilon_2)^2}.$$

2) 如果 $\hat{H}_j$ 与 C_i 不合并在一起,则二者携带的误差为 $err(C_i)+err(\hat{H}_j)$,其中分组 C_i 携带的误差可以表示为 $err(C_i)=AE(C_i)+2/|C_i|(\varepsilon_2)^2$. 关键是如何快速地计算出 $err(\hat{H}_j)$.

如果 $\hat{H}_j$ 不与 C_i 合并,则 $\hat{H}_j$ 有可能与其后面的 $n-j+1$ 个分组合并在一起. 然而,此时无法判断 $\hat{H}_j$ 究竟与 $n-j+1$ 个分组中的哪一个合并在一起,进而无法快速计算出 $err(\hat{H}_j)$. 虽然无法计算出 $err(\hat{H}_j)$ 的具体值,但可以估算出它的懒散误差下界,该下界采用 $err^*(\hat{H}_j)$ 表示.

$err^*(\hat{H}_j)$ 同样也是由近似误差 $AE(\hat{H}_j)$ 与拉普拉斯误差 $LE(\hat{H}_j)$ 组成. 若能快速计算出这 2 种误差的下界,则可以表示出 $err^*(\hat{H}_j)$. 若 $\hat{H}_j$ 不与 C_i 合并,用 2 种极端的情况分布估算 $AE(\hat{H}_j)$ 与 $LE(\hat{H}_j)$.

1) 如果 $\hat{H}_j$ 单独成为一个组,则此时 $AE(\hat{H}_j)=0$,$LE(\hat{H}_j)=2/(\varepsilon_2)^2$.

2) 如果 $\hat{H}_j$ 与其后面的 $n-j+1$ 个桶计数合并成一个新组,$AE(\hat{H}_j)=(\hat{H}_j-\bar{C}_{j,n-j+1})^2$,其中 $\bar{C}_{j,n-j+1}$ 表示 $\{\hat{H}_l,\hat{H}_{l+1},\cdots,\hat{H}_{n-j+1}\}$ 的平均值. 而此时,$\hat{H}_j$ 所引起的拉普拉斯误差 $LE(\hat{H}_j)=2/(n-j+1)^2(\varepsilon_2)^2$.

根据上述 2 种极端情况下的 $AE(\hat{H}_j)$ 与 $LE(\hat{H}_j)$ 取值情况,可以确定 $err(\hat{H}_j)$ 的懒散下界 $err^*(\hat{H}_j)$,进而得到定理 4.

定理 4. $err^*(\hat{H}_j)\leqslant\frac{2}{(n-j+1)^2(\varepsilon_2)^2}$.

证明. $err^*(\hat{H}_j)=\min(AE(\hat{H}_j)+LE(\hat{H}_j))\leqslant$

$\{\min(AE(\hat{H}_j))+\min(LE(\hat{H}_j))\}=$

$\left\{\min(0,(\hat{H}_j-\bar{C}_{j,n-j+1})^2)+\min\left(\frac{2}{(\varepsilon_2)^2},\right.\right.$

$\left.\left.\frac{2}{(n-j+1)^2(\varepsilon_2)^2}\right)\right\}=\frac{2}{(n-j+1)^2(\varepsilon_2)^2}$. 证毕.

根据上述分析,给出 Greedy-Clustering 方法的具体实现细节,如算法 3 所示.

算法 3. Greedy-Clustering 算法.

输入:直方图 $\hat{H}=\{\hat{H}_1,\hat{H}_2,\cdots,\hat{H}_n\}$;

输出:聚类后所得分组集合 C.

① $C=\varnothing$;

② $C_i=\{\hat{H}_1\}$;

③ while $j<n$ do

④ 计算 $err(C_i\cup\hat{H}_j)$; /* 该误差表示分组 C_i 与桶计数 $\hat{H}_j$ 合并带来的误差 */

⑤ 计算 $err(C_i)+\frac{2}{(n-j+1)^2(\varepsilon_2)^2}$; /* 该误差表示 C_i 与桶计数 $\hat{H}_j$ 分开带来的误差 */

⑥ if $err(C_i\cup\hat{H}_j)<err(C_i)+\frac{2}{(n-j+1)^2(\varepsilon_2)^2}$ then

⑦ $C_i=C_i\cup\hat{H}_j$;

⑧ else

⑨ $C=\cup C_i$;

⑩ $C_i=\{\hat{H}_j\}$;

⑪ end if

⑫ end while

⑬ return C.

假设已经形成的当前某个组为 $C_i=[\hat{H}_l,\hat{H}_{l+1},\cdots,\hat{H}_{j-1}]$ 且误差为 $err(C_i)$,如果 $\hat{H}_j$ 加入 C_i 使得 $err(C_i)$

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

降低,则 $\hat{H}_j$ 与 C_i 合并在一起(行⑥~⑦),否则创建一个新组(行⑨~⑩). Greedy-Clustering 方法的时间复杂度为 $O(n)$.

例如:直方图 $H=\{10,10,10,20,20,40,60\}$ 经过算法 2 处理得到 $\hat{H}=\{9,10,10,18,21,42,54\}$,利用 Greedy-Clustering 算法 $\hat{H}$ 进行分组. 设 $\varepsilon_2=0.5$,得到的临时组 $C_1=\{9,10,10\}$,判断 $\hat{H}_4=18$ 是否与组 C_1 合并. $err(C_1\cup\hat{H}_4)=11/2$, $err(C_1)=11/3$, $err^*(\hat{H}_4)=1/2$. 由于 $err(C_1\cup\hat{H}_4)>err(C_1)+err^*(\hat{H}_4)$,则 $\hat{H}_4$ 不与 C_1 合并, $C_1=\{9,10,10\}$, $C_2=\{18\}$.

4 隐私性与可用性分析

算法的隐私性主要从 ε-差分隐私的概念和性质的角度来证明,而可用性主要是利用最终的发布误差进行度量. 本节主要论证 DiffHR 是否满足 ε-差分隐私,并在算法 3 的基础上与 AHP 以及 GS 算法进行可用性对比分析.

4.1 隐私分析

定理 5. DiffHR 满足 ε-差分隐私.

证明. 在 DiffHR 方法中,行②利用指数机制对原始直方图 H 进行 Metropolis-Hastings 抽样. 相当于 n 个查询,这 n 个查询记为 Q_1. 根据定义 3 可知,删除原始 H 一条记录,最多影响 H 中一个桶的计数,则 $\Delta Q_1=1$,根据定理 2 可知行②满足 ε_1-差分隐私. 行⑦利用拉普拉斯机制计算每个组的噪音均值. 由于每个组的大小已知,相当于查询每个组的计数总和,即是查询 SUM 值,该类查询记为 Q_2. 同理 $\Delta Q_2=1$. 根据定理 1 可知,行⑧满足 ε_2-差分隐私. 根据定理 6 差分隐私的顺序组合性质可知,DiffHR 满足 $(\varepsilon_1+\varepsilon_2)$-差分隐私. 由于 $\varepsilon=\varepsilon_1+\varepsilon_2$,则 DiffHR 满足 ε-差分隐私. 证毕.

定理 6[25]. 设 D 为隐私数据集, $A_1,A_2,\cdots,A_n$ 为 n 个随机算法,且 $A_i(1\leqslant i\leqslant n)$ 满足 ε_i-差分隐私. $\{A_1,A_2,\cdots,A_n\}$ 在 D 上操作的顺序组合满足 ε-差分隐私,且 $\varepsilon=\sum_{i=1}^{n}\varepsilon_i$.

4.2 可用性分析

可用性是度量隐私算法的另外一个标准,采用发布误差进行度量. DiffHR 方法发布直方图的最终误差由总体的近似误差与总体的拉普拉斯误差构成. 接下来比较 DiffHR 与 AHP 以及 GS 所发布直方图的可用性和准确性. DiffHR 方法主要是基于懒散分组下界的聚类划分方法进行与其他 2 种方法进行比较.

1) DiffHR 方法与 AHP 方法比较

与 DiffHR 方法相同,AHP 方法同样是利用隐私预算 ε_2 产生每个组的噪音均值. 对于组 C_i,AHP 方法下的 $err_{AHP}(C_i)=AE(C_i)+\frac{2}{|C_i|(\varepsilon_2)^2}$. 因此,聚类分组质量的好坏直接决定着直方图发布误差的大小. 由于 DiffHR 采用 Metropolis-Hastings 逼近直方图正确排序,而 AHP 方法只是基于拉普拉斯噪音的直方图进行排序. 因此,DiffHR 分组的质量好于 AHP,则 DiffHR 的最终发布准确性优于 AHP.

2) DiffHR 方法与 GS 方法比较

尽管 GS 算法也是采用排序与全局化分组策略,但是该方法采用固定宽度划分可能会导致发布准确性比较低. 定理 7 可以解释固定宽度划分存在的缺陷. 设 GS 等宽划分宽度为 w.

定理 7[8]. 设 $C_i=\{\hat{H}_1,\hat{H}_2,\cdots,\hat{H}_w\}$,其中 $\hat{H}_1\leqslant\hat{H}_2\leqslant\cdots\leqslant\hat{H}_w$. 如果 $\hat{H}_{j+1}-\hat{H}_j\geqslant\sqrt{2(w^2-w+1)}/\varepsilon(w-1)$, $1\leqslant j\leqslant w$,则 $err(C_i)\geqslant err(C_i^1)+err(C_i^2)$,其中, $C_i^1=\{\hat{H}_1,\hat{H}_2,\cdots,\hat{H}_j\}$, $C_i^2=\{\hat{H}_{j+1},\hat{H}_{j+2},\cdots,\hat{H}_w\}$.

因此,根据定理 7 可知,GS 方法在固定宽度划分下所形成的组,总能够分裂成 2 个非均匀的误差较小的子组. 相反,DiffHR 却能够自适应地找出这些大小不一的子组,并使得最终发布误差最小.

5 实验结果与分析

实验所采用的 3 个数据集 Waitakere[8], Nettrace[10], Search_log[5] 被广泛使用于直方图发布. 其中 Waitakere 是新西兰 2006 年的人口街区普查数据,抽取 7725 街区作为直方图的桶; Search_log 合成了 2004~2010 年 Google Trends 与 American Online 搜索关键字为"奥巴马"的搜索日志,该数据集包含 32 768 条记录; Nettrace 源自于一所高校内联网的 IP 层网络轨迹数据,包含 65 535 条记录桶,每个桶记录校外主机链接校内某个主机个数.

结合上述 3 种数据集,采用均方误差(MSE)度量 DiffHR, GS 以及 AHP 方法发布直方图的准确性与可用性. 设置隐私预算参数 ε 分别为 0.01, 0.10, 1.00.

均方误差通常用来度量发布直方图 $\widetilde{H}$ 响应一组

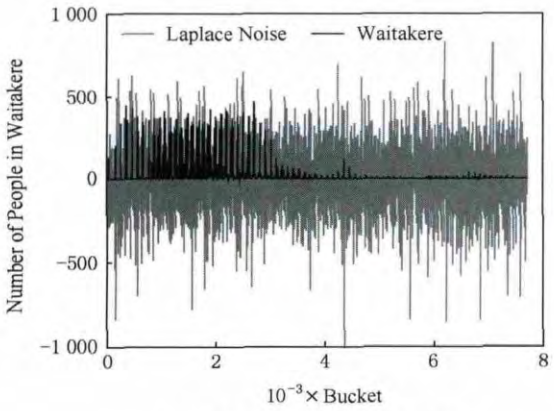
?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

范围计数查询时的质量，假设 $Q=\{Q_1,Q_2,\cdots,Q_l\}$ 为一组范围计数查询，则均方误差可表示为

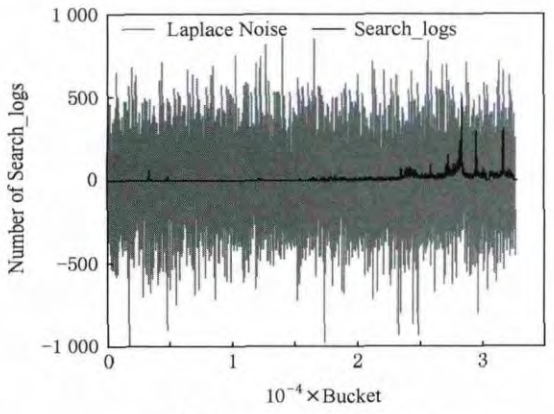
$$MSE(H,\widetilde{H},Q)=\frac{\sum_{i=1}^{l}(Q_i(H)-Q_i(\widetilde{H}))^2}{l}.$$

1) 直接添加拉普拉斯噪音与 DiffHR 方法处理对比

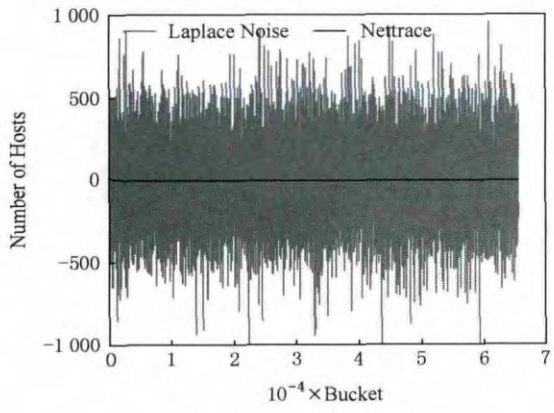
针对上述 3 个数据集，设 $\varepsilon=0.01$，直接对 3 种数据集添加拉普拉斯噪音，结果如图 4 所示. 其中灰色表示拉普拉斯噪音，黑色表示 3 种数据集中直方图的真实计数. 可以看出，拉普拉斯噪音几乎覆盖了真实的计数值，因此，直接在直方图上添加拉普拉斯噪音，其可用性很低，并且不能较好地响应范围计数查询. 图 5 中灰色为基于 DiffHR 方法产生的噪音，黑色同样为 3 种数据集上的真实计数. 可以看出 DiffHR 方法对真实直方图的扰动远低于拉普拉斯机制，特别是在 Search_logs 与 Nettrace 有序的数据集上，DiffHR 方法的噪音影响更小.



(a) Waitakere and Laplace noise



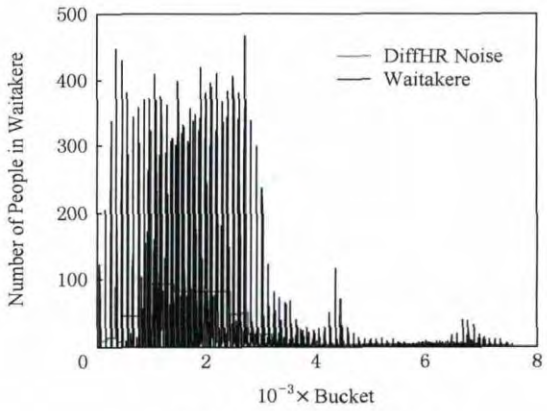
(b) Search_logs and Laplace noise



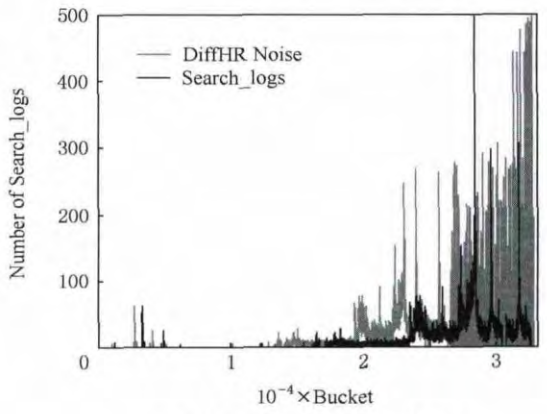
(c) Nettrace and Laplace noise

Fig. 4 Original histogram and Laplace noise.

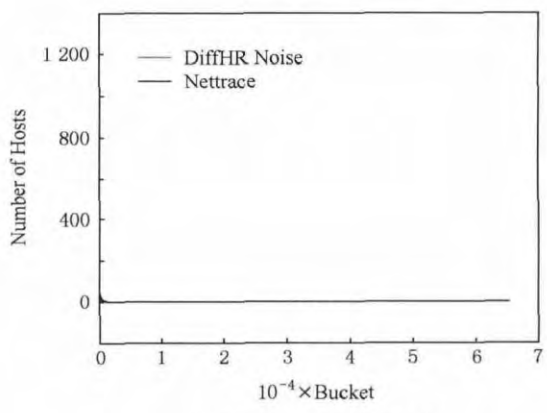
图 4 原始直方图与拉普拉斯噪音



(a) Waitakere and DiffHP noise



(b) Search_logs and DiffHP noise



(c) Nettrace and DiffHP noise

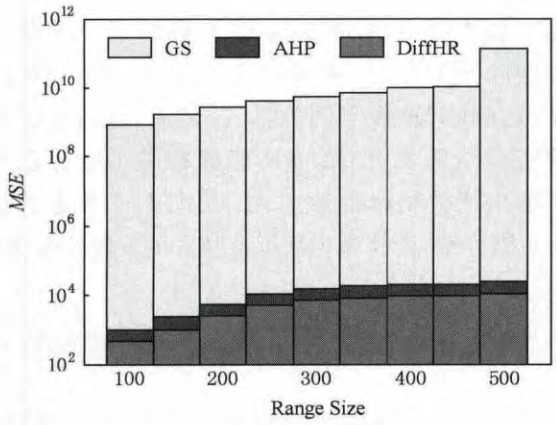
Fig. 5 Original histogram and DiffHR noise.

图 5 原始直方图与 DiffHR 噪音

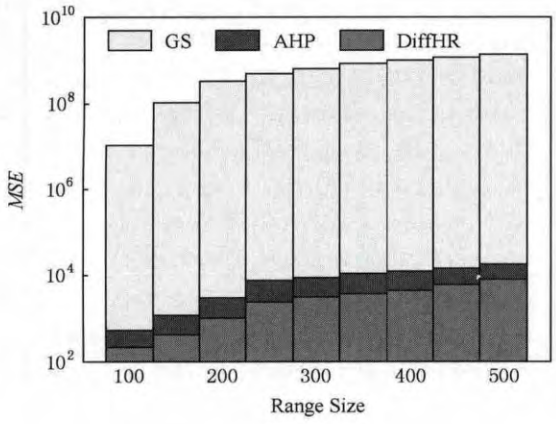
?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2) DiffHR 与 GS 以及 AHP 的均方差(mean squared error, *MSE*)值对比

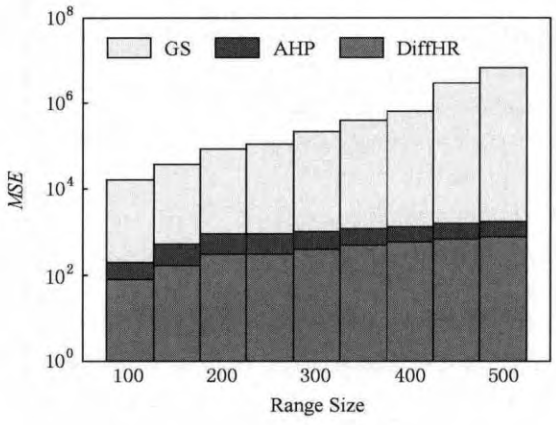
通过改变范围查询的大小以及结合不同的隐私预算参数 ε 值来对比 DiffHR,GS 以及 AHP 方法响应范围查询的结果准确性. 进而比较这 3 种算法所发布直方图的可用性. 由图 6～8 可以发现,范围查询的 *MSE* 值越小表示发布直方图的可用性越高. 随着范围查询大小的增加,以及参数 ε 从 0.01 变化到 1.00,DiffHR,GS 以及 AHP 方法的 *MSE* 值也在相应增加,其原因是大的查询范围以及小的 ε 值造成拉普拉斯噪音的累积,进而增加拉普拉斯误差. 对比着其他 2 种算法,DiffHR 响应 3 种数据集



(a) $\varepsilon=0.01$



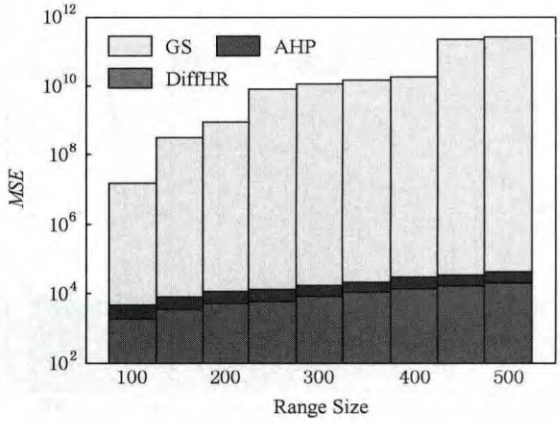
(b) $\varepsilon=0.10$



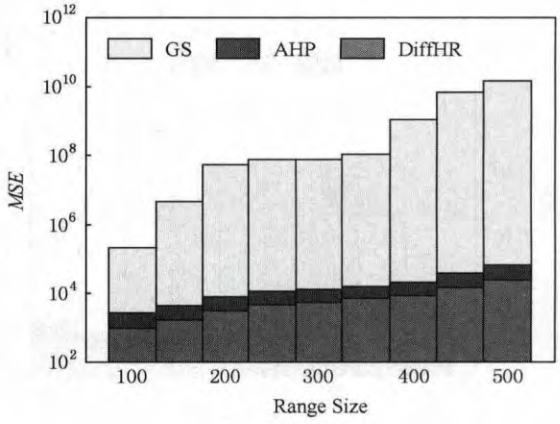
(c) $\varepsilon=1.00$

Fig. 6 *MSE* on Waitakere under different ε values (y-axis is in log-scale).

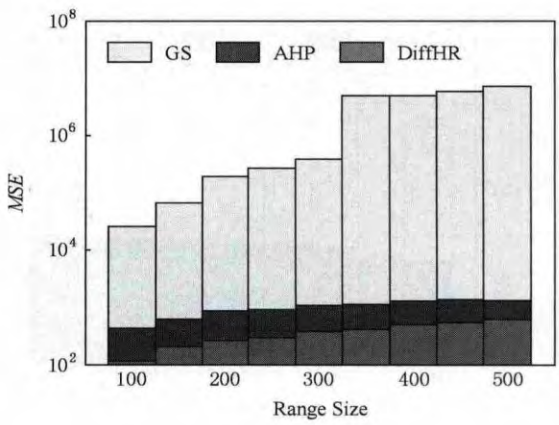
图 6 Waitakere 下 ε 变化时 *MSE* 的变化



(a) $\varepsilon=0.01$



(b) $\varepsilon=0.10$



(c) $\varepsilon=1.00$

Fig. 7 *MSE* on Search_logs under different ε values (y-axis is in log-scale).

图 7 Search_logs 下 ε 变化时 *MSE* 的变化

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

上的范围查询时所造成的 MSE 比较小. 特别是在 Waitakere 与 Search_log 数据集上,DiffHR 所取得的范围查询响应精度是 GS 方法的将近 10 倍,是 AHP 方法的将近 3 倍. 虽然在 Nettrace 数据集上,DiffHR 方法与 AHP 方法取得近似的均方误差,但是 DiffHR 方法仍然优于 AHP 方法,原因是 DiffHR 方法采用了比 AHP 方法更准确的排序策略.

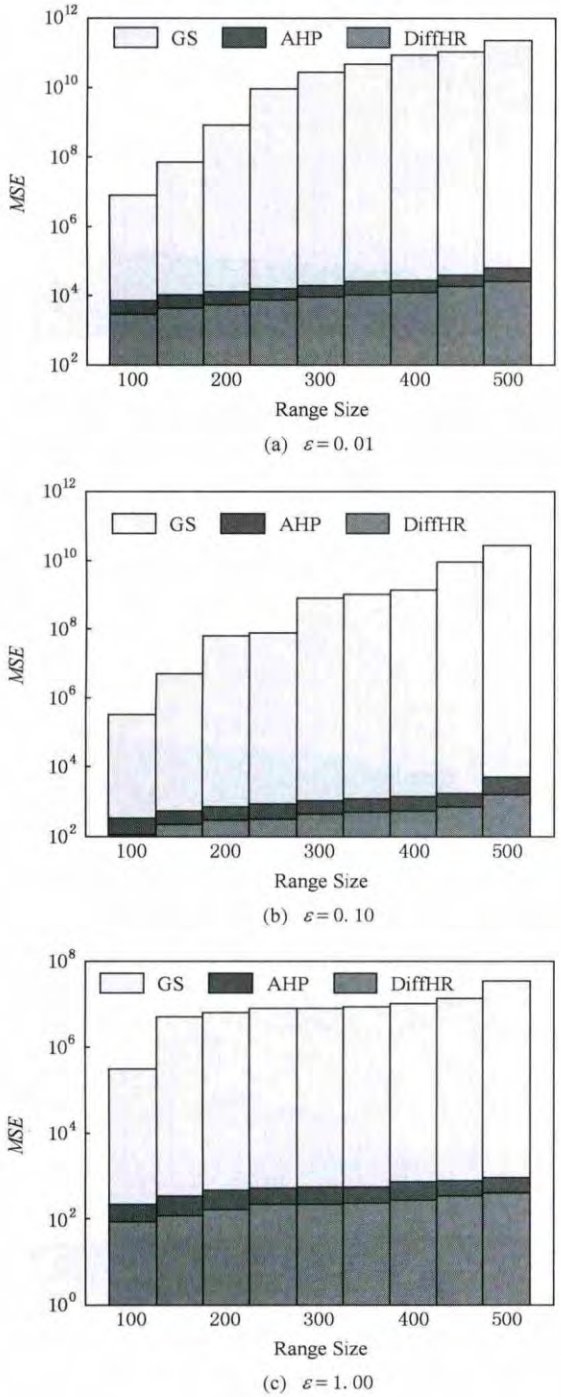


Fig. 8 *MSE* on Nettrace under different ε values (*y*-axis is in log-scale).

图 8 Nettrace 下 ε 变化时 *MSE* 的变化

6 结束语

针对差分隐私行下基于分组的直方图发布问题,本文首先对现有基于分组发布进行分析,并在此基础上提出基于聚类分组算法 DiffHR,引入满足差分隐私的 Metropolis-Hastings 抽样技术进行逼近直方图正确排序. 在排序的基础上提出了一种基于懒散分组的聚类方法. 从理论角度进行对比分析的结果显示,DiffHR 优于同类算法. 最后,通过真实数据集的实验结果表明 DiffHR 能够在满足差分隐私的同时输出比较精确的直方图. 未来工作考虑数据流下发布直方图,设计满足数据流特点的分组发布策略.

参 考 文 献

[1] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C] //Proc of the 3rd Theory of Cryptography Conf (TCC 2006). Berlin: Springer, 2006: 363-385

[2] Dwork C. Differential privacy [C] //Proc of the 33rd Int Colloquium on Automata, Languages and Programming (ICALP 2009). Berlin: Springer, 2006: 1-12

[3] Dwork C, Lei J. Differential privacy and robust statistics [C] //Proc of the 41st Annual ACM Symp on Theory of Computing (STOC 2009). New York: ACM, 2009: 371-380

[4] Dwork C, Naor M, Reingold O, et al. On the complexity of differentially private data release: Efficient algorithms and hardness results [C] //Proc of the 41st Annual ACM Symp on Theory of Computing (STOC 2009). New York: ACM, 2009: 381-390

[5] Xu J, Zhang Z, Xiao X, et al. Differential private histogram publication [J]. The VLDB Journal, 2013, 22(6): 797-822

[6] Acs G, Chen R. Differentially private histogram publishing through lossy compression [C] //Proc of the 11th IEEE Int Conf on Data Mining (ICDM 2012). Piscataway, NJ: IEEE, 2012: 84-95

[7] Li C, Hay M, Miklau G, et al. A data- and workload-aware algorithm for range queries under differential privacy [J]. Proceedings of the Very Large Database Endowment, 2014, 7(5): 341-352

[8] Kellaris G, Papadopoulos S. Practical differential privacy via grouping and smoothing [J]. Proceedings of the Very Large Database Endowment, 2013, 6(5): 301-312

[9] Zhang X, Chen R, Xu J, et al. Towards accurate histogram publication under differential privacy [C] //Proc of the 14th SIAM Int Conf on Data Mining (SDM 2014). Philadelphia, PA: SIAM, 2014: 587-595

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[10] Li C, Hay M, Rastogi V, et al. Optimizing linear counting queries under differential privacy [C] //Proc of the 41st Annual ACM Symp on Theory of Computing (PODS 2010). New York: ACM, 2010: 123-134

[11] Chen S, Zhou S. Recursive mechanism: Towards node differential privacy and unrestricted joins [C] //Proc of the 2014 ACM SIGMOD Int Conf on Management of Data (SIGMOD 2014). New York: ACM, 2013: 653-664

[12] Zhang J, Cormode G, Procopiuc C M, et al. PrivBayes: Private data release via Bayesian networks [C] //Proc of the 2014 ACM SIGMOD Int Conf on Management of Data (SIGMOD 2014). New York: ACM, 2014: 1423-1434

[13] Zhu T, Xiong P, Li G, et al. Correlated differential privacy: Hiding information in non-iid data set [J]. IEEE Trans on Information Forensics and Security, 2015, 10(2): 229-242

[14] Kusner M J, Jacob R G, Roman G, et al. Differentially drivate Bayesian optimization [C] //Proc of the 32nd Int Conf on Machine Learning (ICML). Brookline, MA: Microtome, 2015: 918-927

[15] Su S, Xu S, Cheng X, et al. Differentially private frequent itemset mining via transaction splitting [J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27(7): 1875-1891

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

[17] Hay M, Rastogi V, Miklau G, et al. Boosting the accuracy of differentially private histograms through consistency [C] // Proc of the 36th Int Conf on Very Large Data Bases (VLDB 2010). New York: ACM, 2010: 1021-1032

[18] Rastogi V, Nath S. Differentially private aggregation of distributed time-series with transformation and encryption [C] //Proc of the 2010 ACM SIGMOD Int Conf on Management of Data (SIGMOD 2010). New York: ACM, 2010: 735-746

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

[20] Samarati P. Protecting respondents' identities in microdata release [J]. IEEE Trans on Knowledge and Data Engineering, 2001, 13(6): 1010-1027

[21] Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity: Privacy beyond *k*-anonymity [J]. ACM Trans on Knowledge Discovery from Data, 2007, 1(1): 1-36

[22] McSherry F, Talwar K. Mechanism design via differential privacy [C] //Proc of the 48th Annual IEEE Symp on Foundations of Computer Science (FOCS 2007). Piscataway, NJ: IEEE, 2007: 94-103

[23] Rubinstein R, Kroese D. Simulation and the Monte Carlo Method [M]. New York: Wiley Heyden Ltd, 2008

[24] Gelman A, Rubin D. Inference from iterative simulation using multiplesequences [J]. Statistical Science 1992, 1(4): 451-511

[25] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data qnalysis [C] //Proc of the 2009 ACM SIGMOD Int Conf on Management of Data (SIGMOD 2009). New York: ACM, 2009: 19-30

Zhang Xiaojian, born in 1980. PhD, lecturer in the College of Computer and Information Engineering, Henan University of Economics and Law. His main research interests include differential privacy, data mining, and graph data management.

Shao Chao, born in 1977. PhD, associate professor and master supervisor in the College of Computer and Information Engineering, Henan University of Economics and Law. His research interests include machine learning, data mining and visualization.

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. Executive director of China Computer Federation. His main research interests include cloud data management, Web data management, native XML databases, and flash-based databases, privacy-preserving, and etc.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net