软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(2):381−393 [doi: 10.13328/j.cnki.jos.004863]

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

基于差分隐私的流式直方图发布方法*

张啸剑[1], 孟小峰[2]

[1](河南财经政法大学 计算机与信息工程学院,河南 郑州 450002)

[2](中国人民大学 信息学院,北京 100872)

通讯作者: 张啸剑, E-mail: xjzhang82@ruc.edu.cn, http://www.huel.edu.cn

摘 要: 基于差分隐私保护模型,已经存在多种静态数据集上的直方图发布方法,而目前着重考虑数据流环境下的直方图发布方法却很少.由于数据流本身潜在的复杂性,直接利用现有的满足差分隐私的直方图发布方法处理数据流存在着很多不足,例如发布直方图的可用性低、发布误差大等.基于此,提出了一种基于滑动窗分割的流式直方图发布方法 SHP(streaming histogram publication).该方法通过连续分割每个滑动窗中的桶计数,使其构成不同的分组.根据不同的范围计数查询敏感性,提出了 3 种拉普拉斯噪音添加机制以实现差分隐私保护,分别是滑动窗机制、时间点机制以及自适应抽样机制.在自适应抽样机制中,SHP 算法基于当前的滑动窗,依赖于一种自适应抽样方法对下一时刻的计数进行预测,若预测值与真实值的差异小于给定的阈值则发布预测值,否则发布噪音值.该抽样方法可以有效地节省整体的隐私预算.在真实数据集上对 SHP 算法的可用性进行度量,结果显示,基于抽样的 SHP 算法的可用性高于另外两种方式.

关键词: 差分隐私;数据流;直方图发布;近似误差;拉普拉斯误差

中图法分类号: TP311

中文引用格式: 张啸剑,孟小峰.基于差分隐私的流式直方图发布方法.软件学报,2016,27(2):381−393. http://www.jos.org.cn/1000-9825/4863.htm

英文引用格式: Zhang XJ, Meng XF. Streaming histogram publication method with differential privacy. Ruan Jian Xue Bao/ Journal of Software, 2016,27(2):381−393 (in Chinese). http://www.jos.org.cn/1000-9825/4863.htm

Streaming Histogram Publication Method with Differential Privacy

ZHANG Xiao-Jian[1], MENG Xiao-Feng[2]

[1](School of Computer and Information Engineering, He'nan University of Economics and Law, Zhengzhou 450002, China)

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

Abstract: Various approaches have been proposed to release histogram on static datasets with differential privacy, while little work exist that handle dynamic datasets. Those existing static approaches are ill-suited for the practical applications on data stream due to the inherent complexity of publishing streaming histograms. With this consideration, this paper addresses the challenge by proposing a partitioning-based method, called SHP (streaming histogram publication), which partitions the count values of each sliding window into different groups for releasing the final histogram. In view of different global sensitivity of queries adopted by this paper, three incremental utility-based mechanisms for adding Laplace noise are proposed to achieve differential privacy. The three mechanisms are sliding window

* 基金项目: 国家自然科学基金 (61502146, 61303017, 61202285); 国家高技术研究发展计划(863)(2013AA013204); 高等学校博士学科点专项科研基金(20130004130001); 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)

Foundation item: National Natural Science Foundation of China (61502146, 61303017, 61202285); National High-Tech R&D Program of China (863) (2013AA013204); Research Fund for the Doctoral Program of Higher Education of China (20130004130001); Basic Research Program of He'nan Science and Technology Department (152300410091); Research Program of the Higher Education of He'nan Educational Committee (16A520002)

收稿时间: 2014-03-30; 采用时间: 2015-05-25

mechanism, time point mechanism, and adaptive sampling mechanism, respectively. In the third mechanism, SHP relies on the adaptive sampling method to predict the next arriving count value at non-sampling time points. If the difference between the predicted value and the true value is less than a user-defined threshold, then it releases the predicted value, otherwise, releases the true value. This mechanism can save privacy budget in terms of sampling interval updates. Experimental results or real datasets show that the utility of SHP based on sampling is better than the other two mechanisms.

Key words: differential privacy; data stream; histogram publication; approximate error; Laplace error

作为一种实时动态数据类型,数据流存在于多种应用领域之中,例如疾病应急控制、Amazon 与 Flickr 网站的推荐系统、传感器网络数据处理等.快速而又准确地发布数据流上的聚集统计信息,不但能够提供商业机会,而且可以提供有价值的社会服务.数据流本身的动态性、连续性等特点给实时分析带来很多挑战.目前,概要、抽样以及直方图是分析数据流的常用统计技术.本文利用直方图对该类型数据的统计信息进行概要处理.直方图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.然而,由于数据流中通常蕴含着敏感的个人信息,直接发布直方图统计信息可能会披露个人隐私.

下面给出一个例子说明直接发布流式直方图会导致隐私泄露.例如,流感应急控制中心给出的数据流如图 1 所示.图 2 中的等宽直方图表示了图 1 中在某一时刻已知年龄属性取值后流感疾病的分布情况.假设攻击者知道了 Alice 的年龄为 52 岁,但不知道她是否感染流感.如果该攻击者获得了桶[50,60]中除 Alice 之外其他人的病况(例如感染了流感的病人计数为 1),通过直方图的桶[50,60]计数 2,能够推理出 Alice 感染了流感.因此,在发布数据流的直方图时,要对相应的统计信息进行保护.

姓名	年龄	流感
Alice	20	Yes
Carol	30	Yes
Ellen	40	Yes
Frank	20	Yes
Grace	50	No
Dave	60	Yes
Bob	70	No
…	…	…

Fig.1 Sensitive data stream

图 1 敏感数据流

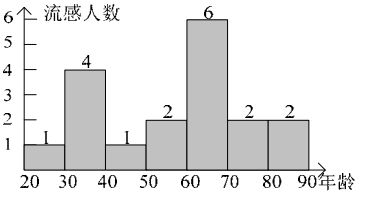


Fig.2 Original histogram

图 2 原始直方图

目前,静态环境中已经存在许多满足差分隐私的直方图发布方法[1−3].这些方法通常采用分组的方法来提高发布精度.分组的主要思想是对原始直方图的桶计数按照近邻关系重新分组,每个分组的数字特征由该组的均值表示.例如,图 3 中的 3 个分组$\{G_1,G_2,G_3\}$即为图 2 中的直方图划分结果.然而,现有的这些基于分组的静态发布方法无法适应于数据流环境,其主要原因包括 3 点:(1) 数据流的动态性要求所发布的直方图要连续更新,否则无法概要全部统计信息;(2) 数据流的实时性要求直方图应及时发布,一旦新的计数值到达,需要立即对其保护处理,然后发布;(3) 最后一个主要原因是,现有的静态方法无法一次性把所有的数据载入内存进行处理.此外,这些方法也无法较好地均衡数据流中直方图的近似误差与拉普拉斯误差.

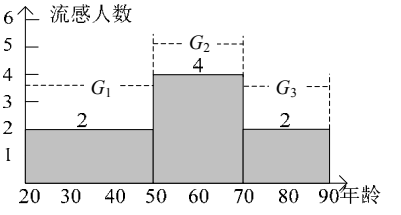


Fig.3 Histogram after partitioning

图 3 划分后的直方图

总而言之,目前还没有一个行之有效的直方图发布方法来同时兼顾数据流统计信息的隐私性与可用性.为

了满足上述 3 种要求,本文提出了一种有效的数据流直方图发布方法.

本文主要贡献如下:

1) 提出了一种有效的流式直方图发布方法 SHP.该方法利用了一种增强动态规划技术来划分每个滑动窗中的直方图.该划分方法能够有效均衡近似误差与拉普拉斯误差.
2) 基于不同的查询敏感性,提出了 3 种拉普拉斯噪音添加机制.为了节省整体的隐私预算,在第 3 种噪音添加机制中提出了一种自适应的抽样方法,在每个滑动窗中实现自适应抽样操作.利用添加噪音的抽样点对新到来的计数值进行估计,若估计值与真实值的差值满足设定阈值,则直接发布估计值,进而可以节省隐私预算.
3) 在 3 种真实数据集上进行了可用性评估实验,实验结果显示,第 3 种机制下的 SHP 方法均优于其他两种机制.

1 相关工作

本文分别从静态环境与数据流环境入手,分析总结与直方图发布相关的研究工作.文献[4]最早提出了一种简单的静态直方图发布方法,该方法直接针对直方图的每个桶计数添加拉普拉斯噪音.然而,该方法在响应较长范围的计数查询时,噪音的累积造成返回结果的误差很高.为了能够比较精确地响应长范围查询,出现了基于层次模型[5,6]与基于划分模型[1−3]的发布方法.文献[5]利用层次树重新组织直方图计数,层次树中的任何范围查询可以由其对应的孩子结点响应,并且采用保序回归技术对发布结果进行求精处理.文献[6]利用哈尔小波变换层次性地分割原始计数,然后再利用含噪音的小波系数重构直方图.该研究展示了小波变换下的直方图能够比较准确地响应较长范围查询.文献[3]提出了两种基于划分的发布方法,分别基于动态规划技术对噪音前和噪音后的计数值进行划分.然而,这两种方法在分组时只考虑了非隐私 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.文献[1]利用有损聚类技术划分直方图,进而弥补了文献[3]的不足.文献[7,8]提出了一种矩阵机制处理直方图.该机制把一批相关的查询作为一个负载进行响应,并通过矩阵分解优化噪音结果.然而,该机制仅适用于小规模数据集,通常产生次优查询结果.为了避免矩阵机制的弊端,文献[9]提出了一种低秩机制,利用二次规划技术产生最优噪音结果.然而,上述基于层次模型与划分模型的方法仅适合静态数据集,而无法适应于复杂的数据流环境.

文献[10]首先提出了一种支持二进制数据流下的持续性计数查询方法,同时,在每个时间戳上给出一个误差下界.文献[11]利用二分机制处理了与文献[10]相同的问题,并且能够灵活地响应计数查询.然而,这两种方法均采用基于事件的差分隐私保护技术[10].文献[12]通过卡尔曼滤波预测数据流中的部分计数值来降低噪音,然而该方法在实际的应用中无法获得理想的准确性.文献[13]研究如何准确地响应一组流式滑动窗查询,利用最小查询子集的线性组合来近似最终查询结果.然而,该方法却不能直接用于我们的问题.文献[14]研究如何在多个事件数据流上回答一个滑动窗查询.其核心思想是:如何在当前滑动窗口中有效地分配隐私预算,并给出相应的范围查询结果.文献[15]考虑分布式数据流中的保护隐私的监控问题,但该方法仅检测数据流的聚合统计值是否超过预设阈值,而不实际发布具体统计值,进而无法解决本文的研究问题.总的来说,现有的差分隐私下数据流处理技术均无法直接满足直方图发布的需求.

2 预备知识

2.1 滑动窗模型

首先介绍本文中的数据流模型和流式直方图发布要求.一个长度为 T 的数据流由定义在 T 个时间戳上的数据点构成 $D=\{x_1,x_2,\ldots,x_T\}$,其中每个数据点 x_i 包含第 $i1$ 和第 i 个时间戳之间的动态信息.与文献[10,11]处理数据流的形式不同,x_i 可以包含复杂信息,而不仅仅是 0/1 字符流或计数信息.为了方便讨论,假设 x_i 为一个直方图.采用滑动窗口模型对数据流 D 进行建模.每个滑动窗口 w_i 由 3 个参数定义:开始时间、当前时间和窗口大小(所包

含数据点的个数).通常,所有滑动窗口均使用固定窗口大小|w|.例如,在图 4 中给出了一个数据流和两个大小为 4 的连续滑动窗口.数据流直方图发布要求在每一个时间戳发布当前滑动窗口所对应的直方图.例如在疾病监控的例子中,要求在每个时间戳发布当前滑动窗口中确诊患者的年龄分布.因此,依据滑动窗模型窗口每向前滑动一个时间戳,就会在当前窗口中发布一个满足差分隐私的直方图.例如,在图 4 所示的疾病监控的例子中,滑动窗 w_i 与 w_{i+1} 中的隐私直方图为 $\tilde{\boldsymbol{H}}(w_i)$ 和 $\tilde{\boldsymbol{H}}(w_{i+1})$, 分别对应着确诊患者的年龄分布.$i$+1

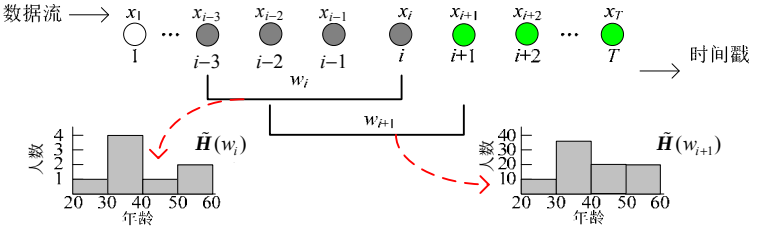


Fig.4 Sliding window model

图 4 滑动窗模型

2.2 差分隐私保护模型

差分隐私技术可以确保在某一数据库中插入或者删除一条记录的操作不会影响任何查询的输出结果,从而保证了每条记录删除或者加入该数据库不会对其隐私造成威胁.为了定义数据流上差分隐私技术,首先给出两个数据流的近邻关系.

定义 1(近邻关系). 给定数据流 D 和 D',如果二者之间最多相差 1 个用户,则 D 和 D'为近邻数据流.

定义 2(ε-差异隐私). 给定数据流 D 和 D',二者互为近邻关系.给定一种隐私算法 A,若算法 A 在 D 和 D'上任意输出的结果 O 满足下列不等式,则 A 为基于用户的ε-差分隐私.

$$\Pr[A(D)=O]\leq \exp(\varepsilon)\times \Pr[A(D')=O] \tag{1}$$

不等式中,Pr[·]控制着算法 A 的随机性,参数ε用来控制隐私保护程度.从不等式可知:参数ε的值越小,$A(D)=O$ 和 $A(D')=O$ 的概率值越接近,说明算法 A 的隐私保护程度越高.

不同于文献[10,11]中的差分隐私保护方法,本文在发布流式直方图时,ε-差分隐私保护技术允许某个用户的记录出现在数据流中的任意时刻.

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为拉普拉斯机制[4].而基于该机制且满足差分隐私的算法所需噪音大小与全局敏感性密切相关.

定义 3(全局敏感性). 对于任意一个函数 $f:D\to R^d$,函数 f 的全局敏感性为

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

其中,R 表示所映射的实数空间,d 表示函数 f 的查询维度.

文献[4]提出的拉普拉斯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,如定理 1 所示.

定理 1[1]. 对于任意一个函数 $f:D\to R^d$,若算法 A 的输出结果满足下列等式,则 A 满足ε-差分隐私:

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

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$\Delta f$ 成正比,与ε成反比,f 的全局敏感性越大,所需噪音就越大.

2.3 直方图分割

给定滑动窗 w_i,设 $\boldsymbol{H}=\{x_{i-w+1},x_{i-w+2},\ldots,x_i\}$,其包含 w 个数据点.为了能够比较精确地响应 w_i 中的范围计数查询,$\boldsymbol{H}$ 通常被分割成大小不同的分组.例如,给定 $\boldsymbol{H}$ 与分组个数 k,设 $\boldsymbol{G}$ 为某个分组策略,则 $\boldsymbol{H}$ 被分割成 $\boldsymbol{G}=\{G_1, G_2,\ldots,G_k\}$,其中,$G_i$ 包含|G_i|个计数值.设 $\overline{G}_i$ 为 G_i 中所有计数的均值.由于利用 $\overline{G}_i$ 替代 G_i 中每个计数值,不可避免

地引入近似误差,而该误差通常采用和方差(sum squared error,简称 SSE)度量,表达式如公式(4)所示.

$$SSE(G_i)=\sum_{x_j\in G_i}(x_j-\bar{G}_i)^2 \tag{4}$$

其中, $\bar{G}_i=\sum_{x_j\in G_i}\frac{x_j}{|G_i|}$.

给出一个例子说明该分割的思想.设 $\boldsymbol{H}$={1,1,4,2,6,2,2},如图 2 中所示,基于 $\boldsymbol{H}$ 相应的分组如图 3 所示,$\boldsymbol{G}$={G_1=(1,1),G_2=(4,4,4),G_3=(2,2)},则 $SSE(G_2)=(4-4)^2+(4-2)^2+(4-6)^2=8$.

2.4 问题定义

在最终发布滑动窗中直方图时,需要采用拉普拉斯噪音对直方图的每个计数进行扰动,才能到达隐私保护的效果.而所添加的噪音会降低窗口中范围计数查询的精度.$\boldsymbol{H}_i$ 与 $\tilde{\boldsymbol{H}}_i$ 分别对应当前滑动窗 w_i 中的原始直方图和隐私直方图,本文的问题是如何使每个滑动窗中的直方图发布误差最小.采用公式(5)度量 $\boldsymbol{H}_i$ 的发布误差.

$$Err(\tilde{\boldsymbol{H}}_i)=E(\|\boldsymbol{H}_i-\tilde{\boldsymbol{H}}_i\|_2^2)=E\left(\sum_{j=i-w+1}^{i}(x_j-\tilde{x}_j)^2\right) \tag{5}$$

其中, $\tilde{x}_j$ 为 x_j 的差分隐私保护之后的形式.

3 SHP 流式直方图发布方法

本节主要介绍 SHP 算法的概述以及该算法的具体实现细节,其中包括滑动窗分割方法、根据查询敏感性不同引起的噪音添加方法、基于抽样技术的计数值估计方法,以及判断 SHP 算法是否满足ε-差分隐私.

3.1 SHP算法概述

不同于静态数据集上的直方图发布,SHP 算法利用大小固定的滑动窗模型处理数据流.滑动窗每次向前滑动一个时间戳时,SHP 均会发布一个满足差分隐私的直方图.该算法的描述如算法 1 所示.

算法 1. SHP 算法.

输入:长度为 T 的数据流 D,滑动窗 w_i,隐私参数ε,w_i 中的分组个数 k,采样频率 s_i.

输出:隐私处理后的直方图 $\tilde{\boldsymbol{H}}_{w_i}$.

初始化第 1 个直方图 w_1

1. $\boldsymbol{G}$=***Partition*** ($\boldsymbol{H}_{w_1}$,ε,k);

窗口滑动阶段

2. $\boldsymbol{Z}$=***Adaptive-Sampling***(w_i,s_i);
3. **for** $x_j\in\boldsymbol{Z}$ **do**
4. $\tilde{x}_j=x_j+Lap(T/s_i\cdot\varepsilon)$;
5. **end for**
6. **for each** new arriving count x_i **do**
7. **if** x_i is sampled **then**
8. $\tilde{x}_i=x_i+Lap(T/s_i\cdot\varepsilon)$;
9. **else**
10. $\hat{x}_i$=***Estimation***($\boldsymbol{Z}$);
11. **end if**
12. Insert $\hat{x}_i$ or $\tilde{x}_i$ into the current window w_i;
13. $\bar{\boldsymbol{H}}_{w_i}=\{\tilde{x}_{i-w+1},\tilde{x}_{i-w+2},...,\tilde{x}_i\}$;
14. $\boldsymbol{G}$=***Partition*** ($\bar{\boldsymbol{H}}_{w_i}$,$k$);

15. **end for**

16. Compute $\tilde{G}_i=\dfrac{\sum_{x_j\in G_i}\tilde{x}_j}{|G_i|}$ for every group G_i;

17. **Return** $\tilde{\boldsymbol{H}}_{w_i}=\{\tilde{x}_{i-w+1},\tilde{x}_{i-w+2},\ldots,\tilde{x}_i\}$;

根据算法 1 可知:

- 给定一个数据流 D 和相应的参数,SHP 算法首先利用 Partition 方法对第 1 个滑动窗中的数据流实现初始化,并发布该窗口中的直方图(第 1 行).
- 然后,滑动窗基于时间戳向前滑动,SHP 算法分两种情况处理当前窗口中的数据流:一是先利用 Adaptive-Sampling 方法,以 s_i 为抽样频率在当前窗口中进行随机抽样(第 2 行~第 5 行);二是对新来的计数值进行处理(第 6 行~第 15 行).如果新来的数据值恰好被抽样,则直接对该值添加拉普拉斯噪音(第 7 行、第 8 行);否则,基于已经抽样的数据值,利用 Estimation 方法对新值进行估计(第 10 行);在当前窗口中,结合抽样值与估计值再次利用 Partition 方法对当前的数据进行分割(第 12 行~第 14 行).
- 最后,发布当前窗口中的直方图(第 16 行、第 17 行).

3.2 滑动窗分割法

在实际的应用需求中,能够比较精确地返回查询结果,是隐私数据发布的主要目的.基于原始直方图的分割方法可以实现这种目标.然而在分割直方图的过程中,有两种误差是不可避免的:一是由分组均值引起的近似误差(SSE);二是由噪音引起的拉普拉斯误差,该误差由拉普拉斯分布的方差来度量.因此,在当前窗口中,如何均衡这两种误差是一个较大的挑战.

任意给定一个组 $G_i=\{x_i,\ldots,x_j\}$($G_i\in \boldsymbol{G}$),该组所携带的总体误差由公式(6)所示.

$$Err(G_i)=SSE(G_i)+Lap(G_i) \tag{6}$$

其中, $SSE(G_i)=\sum_{l=i}^{j}(x_l)^2-\dfrac{\left(\sum_{l=i}^{j}(x_l)\right)^2}{j-i+1}$, $Lap(G_i)=2\left(\dfrac{\Delta}{\varepsilon}\right)^2$.

文献[16]给出一种基于动态规划技术的直方图划分方法.该方法的主要思想是,如何以最小的搜索代价找出相应的划分边界.该思想的形式化表示如公式(7)所示.

$$HErr(j,k)=\min_{k-1\le i\le j-1}(HErr(j,k),HErr(i,k-1)+Err(G_{i+1})) \tag{7}$$

其中,$HErr(j,k)$表示在区间[1,…,j]划分直方图产生的误差,G_{i+1} 表示一个覆盖区间[i+1,…,j]的分组,k 为给定分组个数.

虽然文献[1−3]基于动态规划技术或者聚类技术对直方图进行划分,然而这些方法仅考虑近似误差作为分割条件,而忽视了拉普拉斯误差.此外,文献[14]的方法因存在很高的时间复杂度,而无法适应于处理滑动窗中的数据流.例如,给定一个大小为 w 的滑动窗,在该窗口中划分出 k 个分组,则该操作的时间复杂度为 $O((w+1)^2k)$.通过公式(7)可以观察到:随着 i 的增加,函数 $Err(G_{i+1})$递减,而函数 $HErr(i,k-1)$递增.利用这两种函数的特点,本文提出了一种时间复杂度为 $O(k\cdot\log^3 w/4\delta^2)$的直方图划分算法 Partition,其中,$\delta(\delta>0)$表示误差调和参数.算法 2 描述了滑动窗分割算法 Partition 的细节.

算法 2. Partition 算法.

输入: $\tilde{\boldsymbol{H}}_{w_i}$, 隐私参数$\varepsilon$,采样频率 s_i,$\delta>0$,w_i 中的分组个数 k.

输出:分组结果 $\boldsymbol{G}$.

1. $HErr(i,1)=\left(A_2[1,1]-A_2[1,i]-\dfrac{1}{i-1}\cdot(A_1[1,1]-A_1[1,i])^2\right)+\dfrac{2\Delta^2}{\varepsilon^2}$;
2. **for** l=1 to k−1 **do**
3. Initialize lth queue to empty;

4. **if** $i>i-w+1$ **then**
5. $t=HErr(1,l)$;
6. $c=$***BinarySearch***$(1,i)$; //在区间[1,i]中进行二分查找
7. **if** $HErr(c,l)\leq(1+\delta)t$ and $HErr(c+1,l)>(1+\delta)t$ **then**
8. Insert c at the end of lth queue;
9. **end if**
10. c++;
11. **end if**
12. **end for**
13. **for** j is the end point of $(k-1)$'th queue **do**
14. $HErr(i,k)=\min(HErr(i,k),HErr(j,k-1)+Err(G_{j+1}))$;
15. **end for**
16. **return** $\boldsymbol{G}$;

根据公式(6)与公式(7)可知,计算 $HErr(j,k)$的代价主要来自于 $SSE(G_i)$.为了加速 $SSE(G_i)$的计算,维护了两个数组: $A_1[1,i]=\sum_{l=1}^{i}x_l$, $A_2[1,i]=\sum_{l=1}^{i}(x_l)^2$. 因此,$SSE(G_i)$可以表示成公式(8).

$$SSE(G_i)=(A_2[1,j]-A_2[1,i-1])-(A_1[1,j]-A_1[1,i-1])^2 \tag{8}$$

利用 $HErr(i,1)$表示合并区间[1,i]所产生的最小误差(第 1 行).在每一个时间戳,Partition 方法维护着一个计数值区间队列(第 2 行~第 12 行),而该队列中的元素由不同的终结点构成(第 7 行、第 8 行),而这些终结点可以通过折半查找实现(第 6 行).结合队列结构与动态规划,Partition 方法可以对每个窗口中的直方图划分出合适的分组(第 13 行~第 15 行).

3.3 拉普拉斯噪音添加机制

在发布直方图时,为了保护桶计数值不被泄露,需要随机添加拉普拉斯噪音进行保护.然而噪音会造成拉普拉斯误差,噪音的大小与隐私预算和查询的全局敏感性紧密相关.为了尽可能地减少拉普拉斯误差,提出了 3 种噪音添加机制:滑动窗机制 SWM(sliding window mechanism)、时间点机制 TPM(time point mechanism)和自适应抽样机制 ASM(adative sampling mechanism).

3.3.1 滑动窗机制 SWM

给定数据流 $D=\{x_1,x_2,\ldots,x_T\}$和一个大小为 w 的滑动窗.在每一个滑动窗中,SWM 机制把该窗口中的直方图作为一个范围查询,因此,时间区间[1,T]中存在着 $T-w+1$ 个直方图查询.SWM 机制的主要思想是:把每个滑动窗作为一个整体,然后利用拉普拉斯机制为这个整体一次性地添加噪音.因此,根据全局敏感性的定义 3 可知,$\Delta=w\cdot(T-w+1)$.在 SWM 机制下,如果不利用 Partition 方法划分窗口中的直方图,则每个滑动窗中产生的拉普拉斯误差为 $2(w\cdot(T-w+1))^2\cdot w/\varepsilon^2$.为了减少这种误差,我们调用 Partition 方法对每个滑动窗中的直方图进行分割.然后,再利用 SWM 机制添加噪音.为了度量滑动窗中的直方图误差,首先给出每个分组携带误差的度量方法,见定理 2.

定理 2. 在 SWM 机制下,设 G_i 为当前窗口 w_i 中的任意一个分组,覆盖的时间区域为[i,j],那么 G_i 的总体误差为

$$Err(G_i)=\sum_{x_l\in G_i}(x_l-\bar{G}_i)^2+\frac{2(w(T-w+1))^2}{\varepsilon^2},$$

其中, $\bar{G}_i$ 为 G_i 的均值, $\sum_{x_l\in G_i}(x_l-\bar{G}_i)^2$ 为近似误差, $\frac{2(w(T-w+1))^2}{\varepsilon^2}$ 为拉普拉斯误差.

证明:根据直方图发布误差可知, $\bar{G}_i=\sum_{x_l\in G_i}x_l\big/(j-i+1)$, 设 N_i 为随机添加的拉普拉斯噪音,即

$$N_i=Lap(w(T-w+1)/\varepsilon),$$

其中,$w(T-w+1)$表示查询敏感性. $\tilde{G}_i$ 为 $\bar{G}_i$ 添加噪音之后的表示方式,则分组 G_i 的 $Err(G_i)$的计算过程如下:

$$\begin{aligned}Err(G_i)&=E\left(\sum_{x_l\in G_i}(x_l-\tilde{G}_i)^2\right)\\&=E\left(\sum_{x_l\in G_i}\left(x_l-\bar{G}_i-\frac{\sum_{l=i}^{j}N_l}{j-i+1}\right)^2\right)\\&=\sum_{x_l\in G_i}(x_l-\bar{G}_i)^2-2E\left(\sum_{x_l\in G_i}(x_l-\bar{G}_i)\frac{\sum_{l=i}^{j}N_l}{j-i+1}\right)+E\left(\sum_{x_l\in G_i}\left(\frac{\sum_{l=i}^{j}N_l}{j-i+1}\right)^2\right)\\&=\sum_{x_l\in G_i}(x_l-\bar{G}_i)^2+E\left(\sum_{x_l\in G_i}\left(\frac{\sum_{l=i}^{j}N_l}{j-i+1}\right)^2\right)\\&=\sum_{x_l\in G_i}(x_l-\bar{G}_i)^2+\frac{2(w(T-w+1))^2}{\varepsilon^2}.\end{aligned}$$ □

在 SWM 机制下,设 w_i 有 k 个组, $Err(\tilde{\boldsymbol{H}}_i)_1$ 为 k 个组总的误差,根据定理 2,则 $Err(\tilde{\boldsymbol{H}}_i)_1$ 为

$$Err(\tilde{\boldsymbol{H}}_i)_1=\sum_{i=1}^{k}\sum\nolimits_{x_l\in G_i}(x_l-\bar{G}_i)^2+\frac{2(w(T-w+1))^2}{\varepsilon^2}\cdot k \tag{9}$$

由公式(9)可知:SWM 机制可能产生很大的拉普拉斯误差,特别是 $w=(T+1)/2$.例如,$\boldsymbol{H}=\{1,1,4,2,6,2,2\}$,$w=4$,$T=7$,$k=2$,$\varepsilon=1.0$.因此在当前窗口中,拉普拉斯误差为 1 024.为了减少这种误差,我们提出了第 2 种机制 TPM.

3.3.2 时间点机制 TPM

与 SWM 机制不同,TPM 机制针对每个时间戳上的计数值添加拉普拉斯噪音.由于本文采用的是基于用户的ε-差分隐私,删除或者添加一条用户,有可能影响 T 内所有的计数值.因此,全局敏感性$\Delta=T$,每个计数值的噪音表示形式为 $\tilde{x}_i=x_i+Lap(T/\varepsilon)$.给定一个窗口 w_i,采用 Partition 方法对 w_i 中的直方图进行重新划分,而任意一个组所携带的误差由定理 3 度量.

定理 3. 在 TPM 机制下,设 G_i 为当前窗口 w_i 中的任意一个组,覆盖的时间区域为$[i,j]$,则 G_i 的总体误差为

$$Err(G_i)=\sum\nolimits_{x_l\in G_i}(x_l-\bar{G}_i)^2+\frac{2T^2}{\varepsilon^2}.$$

证明:由定理 2 的方法可直接证明定理 3. □

同样设置 $Err(\tilde{\boldsymbol{H}}_i)_2$ 为当前窗口 w_i 中 k 个组所携带的总体误差,则 $Err(\tilde{\boldsymbol{H}}_i)_2$ 可以表示为公式(10).

$$Err(\tilde{\boldsymbol{H}}_i)_2=\sum_{i=1}^{k}\sum\nolimits_{x_l\in G_i}(x_l-\bar{G}_i)^2+\frac{2T^2}{\varepsilon^2}\cdot k \tag{10}$$

为了说明 TPM 机制下的拉普拉斯误差,我们设置 $\boldsymbol{H}=\{1,1,4,2,6,2,2\}$,$w=4$,$T=7$,$k=2$,$\varepsilon=1.0$.在当前窗口中,拉普拉斯误差为 196.

通过上述两个例子可知,TPM 机制下的拉普拉斯误差低于 SWM 机制.基于此,给出定理 4.

定理 4. $Err(\tilde{\boldsymbol{H}}_i)_2<Err(\tilde{\boldsymbol{H}}_i)_1$.

证明:设当前的窗口 w_i 中有 k 个组,SWM 与 TPM 下的 $\sum_{i\in[1,\ldots,k]}\sum_{x_l\in G_i}(x_l-\bar{G}_i)^2$ 相同, $\tilde{\boldsymbol{H}}_i$ 为 w_i 中直方图,w_i 的大小为 w,而在实际的应用中,$1<w<<T$.利用反证法证明,设 $Err(\tilde{\boldsymbol{H}}_i)_2\geq Err(\tilde{\boldsymbol{H}}_i)_1$.

$$\begin{aligned}Err(\tilde{\boldsymbol{H}}_i)_2-Err(\tilde{\boldsymbol{H}}_i)_1&=\left(\sum_{i=1}^{k}\sum\nolimits_{x_l\in G_i}(x_l-\bar{G}_i)^2+\frac{2T^2}{\varepsilon^2}\cdot k\right)-\left(\sum_{i=1}^{k}\sum\nolimits_{x_l\in G_i}(x_l-\bar{G}_i)^2+\frac{2(w(T-w+1))^2}{\varepsilon^2}\cdot k\right)\\&=\frac{2T^2}{\varepsilon^2}\cdot k-\frac{2(w(T-w+1))^2}{\varepsilon^2}\cdot k=\frac{2k}{\varepsilon^2}(T^2-(w(T-w+1))^2).\end{aligned}$$

由 $Err(\tilde{\boldsymbol{H}}_i)_2\geq Err(\tilde{\boldsymbol{H}}_i)_1$ 可知,$T^2-(w(T-w+1))^2\geq 0$,即$(w-1)(w-T)\geq 0$.

由于 $1<w<<T$, $Err(\tilde{\boldsymbol{H}}_i)_2 \geq Err(\tilde{\boldsymbol{H}}_i)_1$ 不成立,故 $Err(\tilde{\boldsymbol{H}}_i)_2 < Err(\tilde{\boldsymbol{H}}_i)_1$ 成立. □

尽管 TPM 机制的误差低于 SWM 机制,然而若数据流持续的时间 T 过长,则每次发布时都要消耗隐私预算,进而使得 TPM 产生很大的拉普拉斯误差,即 $2T^2 \cdot k/\varepsilon^2$ 过大.为了节省隐私预算,提出了 ASM 机制.

3.3.3 自适应抽样机制 ASM

为了减少拉普拉斯误差与节省隐私预算,ASM 机制利用抽样技术在当前窗口 w_i 中随机地抽取采样点进行添加噪音,通过这些抽样点对新的计数值进行估计.

设 s_i 为抽样频率, $\boldsymbol{Z}=\{\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_m\}$ 为 m 个抽样点, $\tilde{x}_i=x_i+Lap(T/s_i\cdot\varepsilon)$,当 x_{m+1} 到达时,利用 Estimation 方法对其进行估计.

算法 3. Estimation 算法.

输入:当前滑动窗 w_i,采样频率 s_i,$\boldsymbol{Z}$.

输出:预测值 $\hat{x}_{m+1}$.

1. **if** x_{m+1} 不是采样点 **then**
2. $\hat{x}_{m+1}=\beta^{\boldsymbol{t}}\cdot\boldsymbol{Z}$; //$\boldsymbol{t}$ 是由抽样时间点形成的向量
3. **end if**
4. **if** $|x_{m+1}-\hat{x}_{m+1}|\leq\dfrac{\sqrt{2}T}{\varepsilon}$ **then**
5. release $\hat{x}_{m+1}$ at (m+1)'th time stamp rather than $\tilde{x}_{m+1}$;
6. **else**
7. release $\tilde{x}_{m+1}$;
8. adjust the sampling interval: $s_i=s_{i+1}$;
9. **end if**

在算法 Estimation 中,如果 x_{m+1} 不是一个抽样点,则利用以前的抽样点进行估计(第 2 行).若估计值满足条件 $|x_{m+1}-\hat{x}_{m+1}|\leq\sqrt{2}T/\varepsilon$ (第 4 行),则发布估计值 $\hat{x}_{m+1}$;否则发布噪音值 $\tilde{x}_{m+1}$,进而再调节抽样频率.其中, $\sqrt{2}T/\varepsilon$ 表示拉普拉斯噪音分布的标准差(第 5 行~第 8 行).如何实现估计与自适应地抽样,是算法 Estimation 的两个挑战.

为了衡量 $\boldsymbol{Z}=\{\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_m\}$ 中每个抽样点对 x_{m+1} 估计的贡献,定义了一种权重因子,如公式(11)所示.

$$\alpha_j=\frac{1}{(m+1)-j}\cdot\left(\frac{1}{|x_j-\tilde{x}_j|+\sigma}\right) \tag{11}$$

其中,$1\leq j\leq m$,参数σ用来避免除以 0.

根据公式(11),可以得到一个由权重因子组成的向量 $\boldsymbol{A}$,即 $\boldsymbol{A}=\{\alpha_1,\alpha_2,\ldots,\alpha_m\}$.基于向量 $\boldsymbol{Z}$ 与 $\boldsymbol{A}$,$\beta^{\boldsymbol{t}}$ 可以表示成公式(12).

$$\beta^{\boldsymbol{t}}=((\boldsymbol{AZ})^T\boldsymbol{AZ})^{-1}\cdot(\boldsymbol{AZ})^T\boldsymbol{A} \tag{12}$$

因此,基于$\beta^{\boldsymbol{t}}$ 与 $\boldsymbol{Z}$,x_{m+1} 的估计值如公式(13)所示.

$$\hat{x}_{m+1}=\beta^{\boldsymbol{t}}\cdot\boldsymbol{Z} \tag{13}$$

调节抽样频率是为了适应于数据流的动态变化与提高估计精度(第 8 行),若 $|x_{m+1}-\hat{x}_{m+1}|>\sqrt{2}T/\varepsilon$, 则需要调整下一个抽样频率.

设 s_i 与 s_{i+1} 分别表示当前的与下一个抽样点的抽样频率, $\hat{x}_{m+1}-x_m$ 与 $x_{m+1}-x_m$ 分布表示预测值与真实值的变化率,设 $R=|\hat{x}_{m+1}-x_m/x_{m+1}-x_m|$. 基于已经得到的估计值,可以获得 $R_{\min}$ 与 $R_{\max}$ 值.因此,基于 R,s_i 与 s_{i+1},自适应抽样策略如公式(14)所示.

$$s_{i+1}=\begin{cases}s_i\cdot(1+R), & R>R_{\max}\\ s_i\cdot\eta, & R_{\min}<R<R_{\max}\\ s_i\cdot R, & R<R_{\min}\end{cases} \tag{14}$$

其中,η为调和参数.

由公式(14)可知:若 $R>R_{\max}$,则表明 $x_{m+1}-x_m$ 低于 $\hat{x}_{m+1}-x_m$, 需要提高下一个抽样点的抽样频率;若 $R<R_{\min}$,则表明 $x_{m+1}-x_m$ 高于 $\hat{x}_{m+1}-x_m$, 需降低抽样频率;若 R 满足 $R_{\min}<R<R_{\max}$,则用参数η调节即可.

类似于 SWM 机制与 TPM 机制,可以得到定理 5.

定理 5. 在 ASM 机制下,设 G_i 为当前窗口 w_i 中的任意一个组,覆盖的时间区域为$[i,j]$,则 G_i 的总体误差为 $Err(G_i)=\sum_{x_l\in G_i}(x_l-\bar{G}_i)^2+\frac{2T^2}{(s_i)^2\cdot\varepsilon^2}$, 其中,$s_i(s_i>1)$表示第 i 个抽样点的抽样频率.

证明:由 s_i 可知,每个抽样点所消耗的隐私预算为$\varepsilon/(T/s_i)$.利用该预算可以直接计算抽样点的拉普拉斯误差,整个推理过程如定理 2 所示. □

在 ASM 机制下,同样设置 $Err(\tilde{\boldsymbol{H}}_i)_3$ 为当前窗口 w_i 中 k 个组所携带的误差,则 $Err(\tilde{\boldsymbol{H}}_i)_3$ 为

$$Err(\tilde{\boldsymbol{H}}_i)_3=\sum_{i=1}^{k}\sum_{x_l\in G_i}(x_l-\bar{G}_i)^2+\sum_{i=1}^{k}\frac{2T^2}{(s_i)^2\varepsilon^2} \tag{15}$$

通常抽样频率 $s_i>1$,则 $Err(\tilde{\boldsymbol{H}}_i)_3<Err(\tilde{\boldsymbol{H}}_i)_2$.

3.4 SHP算法隐私分析

定理 6. 算法 SHP 满足ε-差分隐私.

证明:分别从 3 种机制下证明 SHP 是否满足ε-差分隐私.设数据流的长度为 T,滑动窗大小为 w.在 SWM 机制下,共有 $T-w+1$ 个直方图查询,每个查询所分得的隐私预算为$\varepsilon/T-w+1$,则根据差分隐私的序列组合性质,SHP 满足ε-差分隐私;在 TPM 机制下,每个时间点所分得的隐私预算为ε/T,同样有序列组合性质得 SHP 满足ε-差分隐私;在 ASM 机制下,设 n 为 T 中的所有抽样点,每个抽样点的预算为ε/n,则根据序列组合性质得 SHP 满足ε-差分隐私. □

性质 1(序列组合性质)[17]. 设 D 为一个隐私数据集,设 $A_1,\ldots,A_n$ 为 n 个随机算法,且 $A_i(1\le i\le n)$满足ε_i-差分隐私,则$\{A_1,\ldots,A_n\}$在 D 上序列组合性操作满足ε-差分隐私,其中, $\varepsilon=\sum_{i=1}^{n}\varepsilon_i$.

4 实验结果与分析

4.1 实验设置

实验环境为 Inter Core 2 Duo CPU 2.94GHz,4GB 内存,Windows 7 操作系统.

使用 Census(http://www.ipums.org),Flu(http://www.cdc.gov/flu)以及 Location(http://www.stats.govt.nz/Census/2006CensusHomePage/MeshblockData.aspx)这 3 个真实的数据集进行验证.Census 抽取了巴西人口普查网站 IPUM 上 42 11 484 条记录,每条记录包含年龄与工资属性,其中,年龄区间为[0,103];Flu 抽取了疾病控制与防御中心网站上在 2004~2011 年期间的 6 280 条记录,年龄的范围是[5,100];Location 是新西兰 2006 年的人口街区普查数据,抽取 7 725 街区作为计数数据.

表 1 给出了实验所用参数的描述以及默认值.

Table 1 Description of experimental parameters

表 1 实验参数描述

对应参数	描述	默认值
k	当前窗口中组个数	$w/10$
η	抽样时的调节参数	2
δ	分割直方图时的精度参数	$1/2k$

本文采用两种评估标准度量 SHP 算法的可用性,分别为查询负载误差(workload error)、绝对误差(absolute error).二者的表示如下:

$$absolute\ error=|noisy-\text{true}|,$$

其中,*noisy* 与 true 分别表示响应某个查询的噪音值与真实值;

$$workload\ error=\sum_{w_i\in W^Q}\frac{Err(\tilde{\boldsymbol{H}}_i)}{T-w+1},$$

其中,$W^Q=\{w_1,w_2,\ldots,w_{T-w+1}\}$表示查询 $Q[w_i,\tilde{\boldsymbol{H}}_i]$ 落入每个滑动窗口所组成的集合.

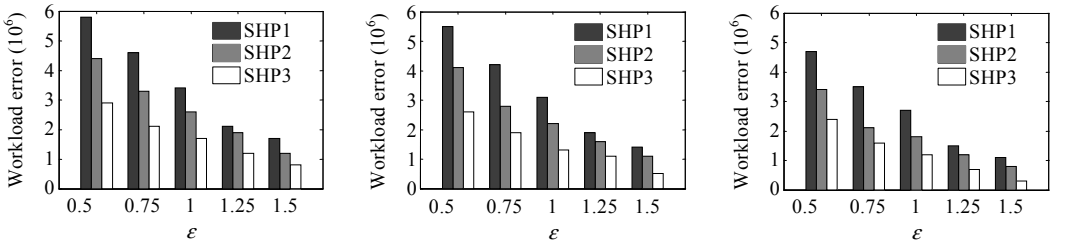
设 SWM,TPM 与 ASM 机制下 SHP 算法分别称为 SHP1,SHP2 以及 SHP3.

4.2 可用性度量

基于上述 3 种数据集,变化隐私预算ε、滑动窗大小 w 以及抽样点数 n 与数据流长度 T 的比例 n/T,来度量 SHP 的可用性.

4.2.1 ε变化对可用性的影响

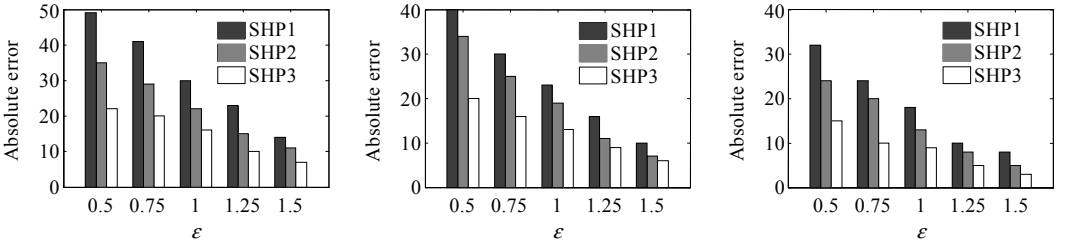
本组实验固定 w=200,n/T=0.4,然后隐私预算参数ε从 0.5 变化到 1.5.图 5 与图 6 显示了算法 SHP1,SHP2 以及 SHP3 的评估结果.



(a) Census (b) Location (c) Flu

Fig.5 Workload error with different ε

图 5 不同ε值下的负载误差



(a) Census (b) Location (c) Flu

Fig.6 Absolute error with different ε

图 6 不同ε值下的绝对误差

图 5 与图 6 给出了ε变化下,负载误差与绝对误差的变化趋势.两图显示的结果表明:ε从 0.5 变化到 1.5 时,SHP3 方法的负载误差与绝对误差低于 SHP2 与 SHP1 方法.尽管大的ε值会相对减少这两种误差,然而在负载误差上,SHP1 方法将近是 SHP3 的 3 倍;而在绝对误差上将近 10 倍.其原因是 SHP3 中的自适应抽样方法可以有效减少拉普拉斯误差.

4.2.2 n/T 变化对可用性的影响

本组实验固定隐私预算ε=1.0,滑动窗 w=200.抽样比率 n/T 分别取 0.2,0.4,0.6,0.8.图 7 显示了算法 SHP1,SHP2 以及 SHP3 在 n/T 变化时的评估结果.从实验结果可以看出,当 n/T 变化时,自适应抽样技术下的 SHP3 方法的绝对误差变化比较平稳,具有较强的鲁棒性.例如,Census 数据集下,当 n/T=0.2 时,SHP3 的绝对误差为 19;当 $n/T\geq$0.4 时,SHP3 方法的误差变化趋于平滑状态;SHP1 与 SHP3 方法的绝对误差随着的 n/T 增加而逐渐变大.

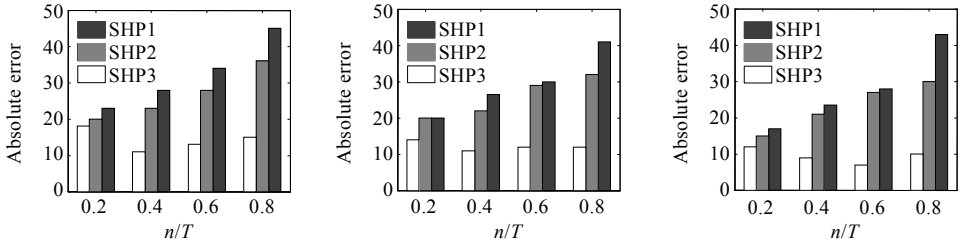


Fig.7 Absolute error with different n/T

图 7 不同 n/T 值下的绝对误差

4.2.3 w 变化对可用性的影响

第 3 组实验固定隐私预算ε=1.0,抽样比率 n/T=0.4,而滑动窗大小 w 分别取值 100,200,300,400,500.实验结果如图 8 与图 9 所示.图 8 与图 9 的结果表明,3 种方法的负载误差与绝对误差都随着参数 w 的增加而增加.其原因是,大的 w 通常会积累较多的拉普拉斯误差.然而,SHP3 方法的这两种误差的增加趋势明显低于 SHP1 与 SHP2 方法,其主要原因是 SHP3 方法采用抽样技术添加噪音,进而避免累积滑动窗中所有计数点携带的误差.

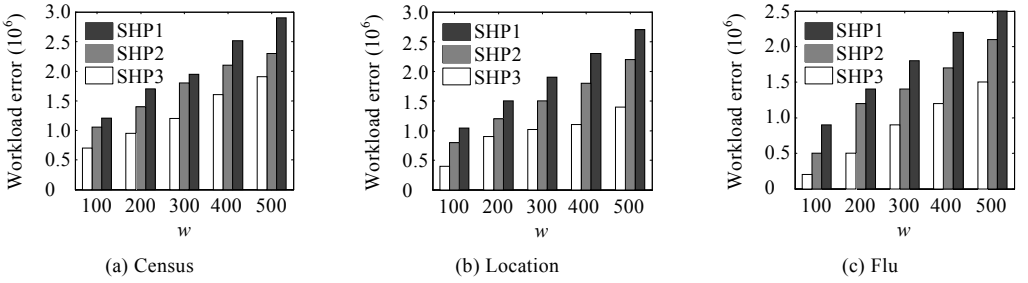


Fig.8 Workload error with different w

图 8 不同 w 值下的负载误差

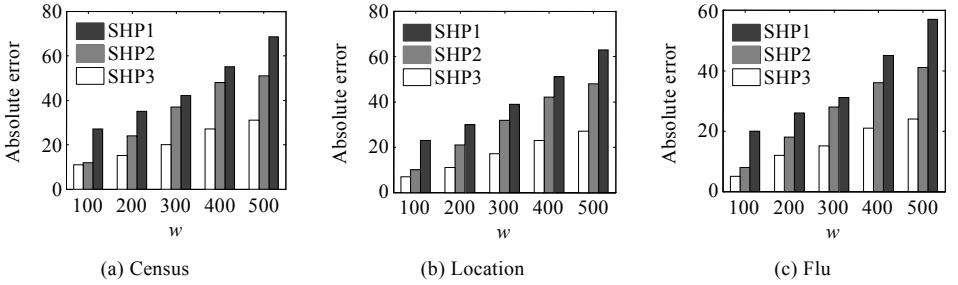


Fig.9 Absolute error with different w

图 9 不同 w 值下的绝对误差

5 结束语

本文主要研究了差分隐私保护下发布数据流形式的直方图,提出了一种新的基于动态规划的发布方法 SHP.该方法利用滑动窗对数据流进行建模,对每个窗口的直方图进行分割.为了提高最终发布结果的可用性,提出了 3 种拉普拉斯噪音添加机制,这 3 种机制都可以平衡近似误差与拉普拉斯误差.在第 3 种机制下,SHP 采用自适应抽样技术节省了隐私预算,进而提高了直方图的发布精度.在 3 种真实数据集上进行 3 种机制下的 SHP

评估,实验结果表明,第 3 种机制下的 SHP 方法均优于其他两种机制.

References:

[1] Acs G, Castelluccia C, Chen R. Differentially private histogram publishing through lossy compression. In: Proc. of the 11th IEEE Int'l Conf. on Data Mining. Piscataway: IEEE, 2012. 84−95. [doi: 10.1109/ICDM.2012.80]

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

[3] Xu J, Zhang Z, Xiao X, Yang Y, Yu G. Differentially private histogram publication. In: Proc. of IEEE the 28th Int'l Conf. on Data Engineering. Piscataway: IEEE, 2012. 32−43. [doi: 10.1109/ICDE.2012.48]

[4] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Proc. of the 3rd Theory of Cryptography Conf. Berlin: Springer-Verlag, 2006. 265−284. [doi: 10.1007/11681878_14]

[5] Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency. Proc. of the VLDB Endowment, 2010,3(1):1021−1032. [doi: 10.14778/1920841.1920970]

[6] Xiao X, Xiong L, Yuan C. Differential privacy via wavelet transforms. IEEE Trans. on Knowledge and Data Engineering, 2010, 23(8):1200−1214. [doi: 10.1109/TKDE.2010.247]

[7] Li C, Hay M, Rastogi V, Miklau G, McGregor A. Optimizing linear counting queries under differential privacy. In: Proc. of the 29th ACM Symp. on Principles of Database Systems. New York: ACM Press, 2010. 123−134. [doi: 10.1145/1807085.1807104]

[8] Li C, Miklau G. An adaptive mechanism for accurate query answering under differential privacy. Proc. of the VLDB Endowment, 2012,5(6):514−525. [doi: 10.14778/2168651.2168653]

[9] Yuan G, Zhang Z, Winslett M, Xiao X, Yang Y, Hao Z. Low-Rank mechanism: optimizing batch queries under differential privacy. Proc. of the VLDB Endowment, 2012,5(11):1352−1363. [doi: 10.14778/2350229.2350252]

[10] Dwork C, Naor M, Pitassi T, Rothblum GN. Differential privacy under continual observation. In: Proc. of the 42nd ACM Symp. on Theory of Computing. New York: ACM Press, 2010. 715−724. [doi: 10.1145/1806689.1806787]

[11] Chan THH, Shi E, Song D. Private and continual release of statistics. ACM Trans. on Information and System Security, 2011,14(3): 26−49. [doi: 10.1145/2043621.2043626]

[12] Fan L, Xiong L. An adaptive approach to real-time aggregate monitoring with differential privacy. IEEE Trans. on Knowledge and Data Engineering, 2014,26(9):1041−4347. [doi: 10.1109/TKDE.2013.96]

[13] Cao J, Xiao Q, Ghinita G, Li N, Bertino E, Tan KL. Efficient and accurate strategies for differentially-private sliding window queries. In: Proc. of the 16th Int'l Conf. on Extending Database Technology. New York: ACM Press, 2013. 191−202. [doi: 10.114 5/2452376. 2452400]

[14] Kellaris G, Papadopoulos S, Xiao X, Papadias D. Differentially private event sequences over infinite streams. Proc. of the VLDB Endowment, 2014,7(12):1155−1166. [doi: 10.14778/2732977.2732989]

[15] Friedman A, Sharfman I, Keren D, Schuster A. Privacy-Preserving distributed stream monitoring. In: Proc. of the 21st Annual Network and Distributed System Security Symp. Washington: The Internet Society, 2014. [doi: 10.14722/ndss.2014.23128]

[16] Jagadish HV, Koudas N, Muthukrishnan S, Poosala V, Sevcik KC, Suel T. Optimal histograms with quality guarantees. In: Proc. of the 24th Conf. of Very Large Databases. San Francisco: Morgan Kaufmann Publishers, 1998. 275−286.

[17] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Presss, 2009. 19−30. [doi: 10.1145/1559845.1559850]

张啸剑(1980－),男,河南淮阳人,博士,讲师,CCF 学生会员,主要研究领域为隐私保护,数据挖掘,图数据管理.

孟小峰(1964－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为大数据管理,面向新型硬件大数据管理,大数据隐私管理,大数据分析,社会计算.