

基于直方图的隐私键-值数据收集算法

张啸剑¹ 徐雅鑫¹ 付楠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)

Towards Private Key-Value Data Collection with Histogram

Zhang Xiaojian¹, Xu Yaxin¹, Fu Nan¹, and Meng Xiaofeng²

¹(School of Computer & Information Engineering, Henan University of Economics and Law, Zhengzhou 450002)

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

Abstract Recently, user data collection and analysis with local differential privacy has extended into key-value data. The trade-off between the size and sparsity of domain and perturbation method directly constrains the accuracy of the collection and analysis of such data. To remedy the deficiency caused by the domain size and perturbing method, this paper employs histogram technology to propose an efficient solution, called HISKV, to collect key-value data. HISKV firstly uses a user-grouping strategy and partial privacy budget to find the optimal length of truncation and enables each user to truncate his/her key-value data set. And then, based on the truncated set, each user samples one key-value pair and uses the discretization and perturbation method to process this pair. To perturb key-value data efficiently, a novel mechanism in HISKV, named LRR_KV is proposed, which allocates different perturbing probability for different keys. In LRR_KV, each user adopts this mechanism to add noise to his/her sampled pair, and sends the report to a collector. Based on the reports from all of the users, the collector estimates the frequency of each key and the mean of the values. To evaluate the utility of HISKV, we firstly conduct theoretical analysis on unbiased, variance, and error bound of LRR_KV, and then perform experiments on real and synthetic datasets to compare different methods. The experimental results show that HISKV outperforms its competitors.

Key words local differential privacy; randomized response mechanism; key-value data; frequency estimation; mean estimation

摘要 基于本地差分隐私的用户数据收集与分析算法已延伸到了键-值数据类型。然而,该类数据值域大小与稀疏性以及本地扰动机制直接制约着收集与分析精度。针对现有机制难以有效应对该类数据收集的不足,提出了一种基于直方图技术的有效收集与分析算法 HISKV(histogram-based key-value data collection),该算法首先结合用户分组策略寻找最优截断长度,利用最优截断-抽样技术处理值域过大与稀疏性问题,然后结合截断结果随机抽取单个键-值对进行离散化处理。针对离散化结果,设计一种高效的本地扰动机制 LRR_KV(local random response for key-value data),该机制结合具体的键分配不同

收稿日期:2020-05-13;修回日期:2020-11-25
基金项目:国家自然科学基金项目(61502146,91646203,91746115,62072156);河南省自然科学基金项目(162300410006);河南省科技攻关项目(202102310563);河南财经政法大学青年拔尖人才资助计划项目
This work was supported by the National Natural Science Foundation of China (61502146, 91646203, 91746115, 62072156), the Natural Science Foundation of Henan (162300410006), the Key Technologies Research and Development Program of Henan Province (202102310563), and the Young Talents Fund of Henan University of Economics and Law.

的本地扰动概率.每个用户利用 LRR_KV 机制扰动离散化的键-值对之后发送给收集者,收集者结合用户的报告值对每个键的频率及其值所对应的均值进行估计.理论分析了 HISKV 算法的无偏性、所产生的方差以及最大偏差,并与现有的键-值收集算法在真实与合成的数据集上进行比较,实验结果表明 HISKV 算法优于同类算法.

关键词 本地差分隐私;随机应答机制;键-值数据;频率估计;均值估计

中图法分类号 TP392

移动互联网环境下的新兴技术快速发展与应用,使得键-值(key-value)数据的获取与收集变得尤为容易,例如网购、健康医疗、金融等数据的收集与分析.键-值数据收集与分析能够提升产品服务与设备服务的质量,以及向用户提供个性化的体验.然而,键-值数据通常蕴含着丰富的个人敏感信息,在提供给收集者或者第三方应用的同时,个人的敏感信息有可能被泄露.表 1 为某不可信购物网站收集用户购买与评分的记录.通过表 1 中键的频率以及键所对应值的均值分析,可以学习出用户的购物行为模式与偏好.例如,商品 k_2 的频率为 2, k_2 所对应值的均值为 0.61.不可信收集者分析出 k_2 的频率之后,即可对表 1 中的用户进行统计攻击.因此,在此情景下,用户无法控制自己的隐私数据.本地差分隐私保护技术的出现有效地解决了此类矛盾.该技术允许每个用户扰动自身数据之后再响应收集者的需求.

Table 1 Example of Key-Value Data

表 1 键-值数据样例

用户	〈商品,评分〉
u_1	$\langle k_1, 0.7 \rangle, \langle k_2, 0.8 \rangle, \langle k_3, 0.4 \rangle, \langle k_4, 0.55 \rangle, \langle k_5, 0.6 \rangle$
u_2	$\langle k_2, 0.42 \rangle, \langle k_4, 0.6 \rangle$
u_3	$\langle k_5, 0.6 \rangle$

目前基于本地差分隐私的键-值收集算法包括 PrivKV^[1], KVUE^[2], KVOH^[2], PCKV-GRR^[3], LDPKV^[4].PrivKV, KVUE, KVOH, LDPKV 算法基于键的整个值域进行编码,利用简单随机抽样技术减少收集者与用户之间的通信代价.然而,这 4 种算法均没有考虑到键的值域稀疏性与偏斜性问题.例如,假设表 1 中商品的值域为 7.以 u_3 为例,对其键编码之后可以获得长度为 7 的向量 $\langle 1, 0.6 \rangle, \langle 0, 0 \rangle, \langle 0, 0 \rangle, \langle 0, 0 \rangle, \langle 0, 0 \rangle, \langle 0, 0 \rangle, \langle 0, 0 \rangle$.基于该向量进行随机抽样时, $\langle 0, 0 \rangle$ 以 6/7 的概率被抽取到,而 $\langle 0, 0 \rangle$ 对最终统计结果无效.当键的值域非常大且稀疏时,PrivKV, KVUE, KVOH, LDPKV 算法估计误差比较高.

为了解决 PrivKV, KVUE, KVOH, LDPKV 算法无法应对键的值域稀疏性问题, PCKV-GRR 利用填充-抽样^[5]技术减少稀疏性带来的影响.其主要思想是利用给定长度 ℓ 截断每个用户所拥有的键-值对集合,然后利用哑元填充那些小于 ℓ 的键-值对集合.例如,表 1 中用户 u_2 利用 $\ell = 3$ 截断所拥有的键-值对集合后,该集合变成 $\{\langle k_2, 0.42 \rangle, \langle k_4, 0.6 \rangle, \tau\}$,其中 τ 为哑元.然而,该算法存在 3 点不足:

- 1) 与 PrivKV 类似,把隐私预算 ϵ 分割成 ϵ_1 与 ϵ_2 ,其中 ϵ_1 用来本地扰动键,而 ϵ_2 用来扰动每个键所对应的值;
- 2) 在设定本地扰动概率时,没有考虑真实键对最终估计的影响,无论键扰动自身还是其他值均设定相同的扰动概率;
- 3) 没有给定截断长度 ℓ 的计算方法.

若用户直接发送所拥有的键-值对长度,则会导致隐私泄露.此外,如何选择合理的 ℓ 至关重要,直接影响最终键频率与均值的估计精度.例如,以真实稀疏数据 TalkingData 为例,该数据集中非零键-值对所占的比例仅为 6.9%.根据该数据中 60 032 个用户的键-值对长度分布(如图 1 所示)可知,98% 的用户的键-值对长度小于 50.因此, ℓ 的理想选择区间为 $[0, 50]$.

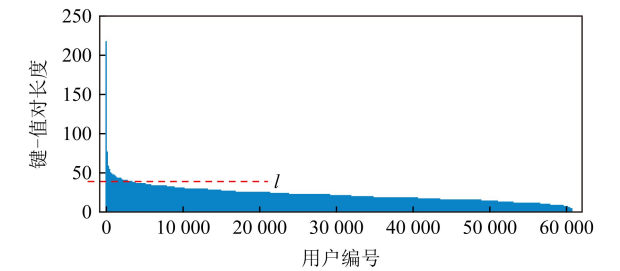


Fig. 1 Length distribution of TalkingData key-value pair
图 1 TalkingData 键-值对长度分布

结合文献[1-4]可知,收集与分析键-值数据过程中存在的挑战包括:1)用户在设计本地扰动算法时,如何设置真实键的扰动概率以及维持键与值之间的关联性;2)由于键-值数据的值域通常很大且

稀疏,如何设计有效的最优截断长度;3)收集者如何以较小的通信代价收集所有用户的键-值数据.总而言之,目前还没有一个行之有效且满足本地差分隐私的键-值收集与分析算法能够克服上述算法存在的问题.为此,本文基于本地差分隐私技术提出了一种键-值数据收集算法能够兼顾3方面挑战需求.

本文主要贡献有3个方面:

1)为了解决挑战1,提出一种基于直方图技术的键-值频率与均值估计算法 HISKV.不同于 PrivKV, KVOH, LDPKV, PCKV-GRR 算法, HISKV 算法在不分割隐私预算的情况下,利用直方图技术对键-值对进行整体扰动,进而维持了键-值之间的关联性.同时,对于真实的键分配较大的本地扰动概率.

2)为了有效解决挑战2与挑战3,提出了一种最优截断算法来寻找截断长度 ℓ .不同于 PCKV-GRR 算法中的经验设置 ℓ , HISKV 利用用户分组策略与部分隐私预算估计 ℓ .结合 ℓ 利用填充-抽样技术处理键值域的稀疏性问题.

3)理论分析了 HISKV 算法满足 ϵ -本地差分隐私,以及频率与均值估计的无偏性、误差边界以及最大偏差.通过合成数据与真实数据实验分析, HISKV 算法具有较高的可用性.

1 相关工作

本地化差分隐私技术在假设收集者完全不可信的情况下,允许每个用户利用随机响应机制^[6]本地扰动自己的数据,并把噪音结果发送给收集者.目前本地化差分隐私研究主要集中于频率估计^[7-11]与均值估计^[12-15]. GRR^[7]算法通过直接编码将整个类别属性的值域进行转换,然后对原始值进行本地扰动,而该算法的通信代价与误差均较高.类似于 GRR 算法, RAPPOR^[8]算法结合一元编码与布隆过滤技术把类别属性的整个值域散列到较小的值域中,结合散列值域统计每个类别属性的频率.该算法能够永久性或者及时性地保护每个用户的数据,相应的通信代价较低.基于 RAPPOR 算法, O-RR^[9]算法借助散列编码与分组操作估计整个类别属性的概率分布情况.无论是 GRR, RAPPOR 及其变种算法,它们的误差均随着值域的增加而增加. OLH^[7]算法克服了对值域大小的依赖,该算法利用优化本地散列技术将整个值域 Hash 到较小地址空间.为了进一步减少通信代价,在较高计算代价的基础上, SHist^[10]算法采用随机矩阵投影技术对类别属性值域进行

编码,随机扰动1个比特位发送给收集者.此外, LDPMiner^[11]算法基于 SHist 算法研究了集-值数据下的 Heavy Hitter 问题.不同于上述基于类别属性的频率估计算法,文献[12-15]集中研究 $[-1, 1]$ 中数值均值的估计. Duchi^[12-13]算法基于离散化操作,对数值型属性的均值进行估计,并取得较小的通信代价与较低的估计误差.基于 Duchi 算法, Harmony^[14]算法在多维数值元组中随机抽取单个数值并扰动成离散值,然而该算法的估计结果远离了区间 $[-1, 1]$ 中的真实值. PM^[15]算法能够将 $[-1, 1]$ 中的真实值扰动到一个连续的区间,并在该连续区间中计算出该扰动值的左右边界.在隐私预算较大时, PM 算法的估计精度明显高于 Duchi 与 Harmony 算法.

尽管文献[7-15]对类别属性与数值属性上的频率与均值进行了较为系统的研究,然而这些算法很难直接应用于键-值数据,其主要原因在于这些算法没有考虑键-值对之间的关联性.文献[1-4]对键-值数据上的频率与均值估计进行了研究.基于 RAPPOR 与 Harmony 算法, PrivKV^[1]算法对键-值数据的频率与均值进行了估计, PrivKV 与 Harmony 的主要区别在于 PrivKV 把扰动后的数值修正操作放到了收集者端.然而, PrivKV 算法由于分割隐私预算导致估计精度较低. KVOH^[2]与 LDPKV^[4]算法结合键-值数据值域与扰动输出值域之间的整体映射关系,利用 GRR 算法对离散化后的键-值对进行扰动.然而,这2种算法的误差易受到值域大小的影响.不同于 PrivKV, KVOH, LDPKV 3种算法, PCKV-GRR^[3]算法利用填充-抽样操作对原始键-值序列进行截断与抽样,然后再利用 GRR 算法进行本地扰动.尽管 PCKV-GRR 在值域较大且稀疏的情况下优于 KVOH 与 PrivKV,但是该算法的误差同样受到值域大小的影响,并且没有顾及到如何设置有效键的扰动概率以及截断长度的计算方法.基于上述分析,本文提出一种基于本地差分隐私的键-值收集与分析算法 HISKV,该算法利用直方图技术统一估计键的频率及其均值,利用最优截断长度与填充-抽样技术处理值域稀疏问题.

2 定义与问题

2.1 本地差分隐私

不同于中心化差分隐私保护技术,本地差分隐私技术通常要求用户在本地保护自己的数据,把扰动

之后的数据报告给不可信的收集者,从而实现隐私不被泄露.本地差分隐私的形式化定义为:

定义 1. ϵ -本地差分隐私.给定一个随机算法 A 及其定义域 $Dom(A)$ 和值域 $Range(A)$,若算法 A 在任意 2 条不同元组 t 与 $t'(t, t' \in Dom(A))$ 上得到相同输出结果 $o(o \in Range(A))$ 满足不等式:

$$Pr[A(t)=o] \leq e^\epsilon \times Pr[A(t')=o], \quad (1)$$

则 A 满足 ϵ -本地差分隐私.其中 ϵ 为隐私预算,其值越小则算法 A 的隐私保护程度越高.

2.2 随机应答机制

随机应答机制^[6]是实现本地差分隐私的常用技术.在用户发送数据 t_i 之前,对其进行随机扰动.该机制的原始思想是用户在响应敏感的布尔问题时,以概率 p 真实应答,以 q 的概率给出相反的应答.为了使随机应答机制满足 ϵ -本地差分隐私,通常设置 $p = e^\epsilon / (1 + e^\epsilon)$, $q = 1 / (1 + e^\epsilon)$.目前,基于随机应答机制出现了以 GRR 为代表的本地扰动算法.给定 $k_i, k_j \in \{1, 2, \dots, d\}$, GRR 算法:

$$Pr[GRR(k_i)=k_j] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + d - 1}, & k_i = k_j; \\ q = \frac{1-p}{d-1}, & k_i \neq k_j. \end{cases} \quad (2)$$

2.3 问题描述

类似文献[4],给定 n 个用户 $U = \{u_1, u_2, \dots, u_n\}$ 和一个收集者,每个用户拥有键-值对集合. $K = \{k_1, k_2, \dots, k_d\}$ 为键的值域, d 为该值域大小, $V = [-1, 1]$ 为键所对应值的值域. $S_i = \{\langle k_j, v_j \rangle \mid 1 \leq j \leq l_i, k_j \in K, v_j \in V\}$ 为用户 u_i 所拥有的键-值对集合,其中 l_i 表示 S_i 的长度.收集者收集所有用户的键-值数据后,通常估计键所对应的频率以及值所对应的均值,具体计算为

$$f_k = \frac{1}{n} |\{u_i \mid \exists (k, v) \in S_i\}|, \quad (3)$$

$$\mu_k = \frac{\sum_{i=1}^n \sum_{j: k_j=k} v_j}{|\{u_i \mid \exists (k, v) \in S_i\}|}. \quad (4)$$

采用均方误差与相对误差度量 f_k 与 μ_k 的估计精度.本文要解决的问题是在设计满足本地差分隐私的频率与均值估计算法的同时,要尽可能获得误差较小的估计结果.

3 基于直方图技术的键-值数据收集算法

3.1 HISKV 算法

不同于文献[1-4]中的收集算法, HISKV 算法

不需要对每个用户的键按照值域进行编码,而是仅对键所对应值进行离散化处理.给定某个键-值对 $\langle k, v \rangle$,只对值 v 进行离散化操作,结果为

$$v = \begin{cases} +1, & \text{以概率 } \frac{1+v}{2}; \\ -1, & \text{以概率 } \frac{1-v}{2}. \end{cases} \quad (5)$$

例如,设 $\langle \text{汉堡}, 0.4 \rangle$ 为某一个键-值对,则 0.4 以 0.7 的概率离散化 +1;以 0.3 的概率离散化为 -1.

根据式(5)可知,每个用户所拥有键的对应值均被离散化成 2 个类别,分别为 +1 或者 -1,而所有的键还保持原来的值.因此,可以在 $\{d, +1\}$ 与 $\{d, -1\}$ 这 2 个二维值域内构建相应的直方图来估计键的频率以及所对应值的均值,其中每个键即是直方图的桶.结合桶的频率可以估计每个键的频率及所对应值的均值.如表 1 所示,用户 u_1, u_2, u_3 的键-值对经过离散化之后的值为 $\{\langle k_1, +1 \rangle, \langle k_2, +1 \rangle, \langle k_3, -1 \rangle, \langle k_4, -1 \rangle, \langle k_5, +1 \rangle\}, \{\langle k_2, -1 \rangle, \langle k_4, +1 \rangle, \langle k_5, +1 \rangle\}$.根据 $d = \{k_1, k_2, k_3, k_4, k_5, k_6, k_7\}$ 以及 d 所对应的值 $\{+1, -1\}$ 可获得相应直方图(如图 2 所示).根据图 2 可以估计出 d 中每个键的频率及均值.例如,键 k_2 的频率为 2/3, k_2 所对应值的均值为 1.

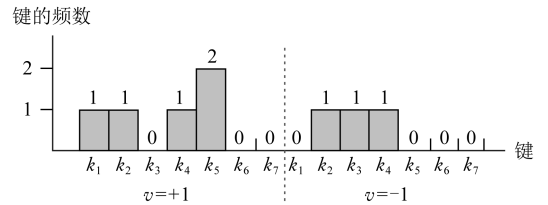


Fig. 2 Two-dimensional range histogram

图 2 二维值域直方图

结合直方图技术的 HISKV 算法如算法 1 所示:

算法 1. HISKV 算法.

输入: n 个用户、每个用户所拥有的键-值对集合 $S = \{S_1, S_2, \dots, S_n\}$ 、隐私预算 ϵ 、键的值域 d ;

输出: 频率向量 f^* 、均值向量 μ^* .

① $f^* \leftarrow \emptyset, \mu^* \leftarrow \emptyset$;

② 收集者将 n 个用户分成 2 组 $\beta n, (1-\beta)n$;

③ $\ell \leftarrow \text{OETL}(\beta n, \epsilon)$; /* 估计最优长度 */

④ for 在 $[(1-\beta)n]$ 内的每个用户 u_i 执行步骤

⑤ ~ ⑥; /* 每个用户扰动 */

⑤ 用户 u_i 报告 $(j, \langle k'_j, v'_j \rangle) \leftarrow$

$\text{LRR_KV}(S_i, \epsilon, \ell)$;

⑥ 用户 u_i 发送 $(j, \langle k'_j, v'_j \rangle)$ 到收集者端;
 ⑦ end for
 ⑧ for 每个用户汇报的键值对 $\langle k'_j, v'_j \rangle$ 执行步骤
 ⑨~⑰; /* 收集者统计每个键的计数 */
 ⑨ if $k_j \in [1, 2(d+\ell)]$ then
 ⑩ if $v'_j = 1$ then
 ⑪ $c_{k_j}^{v'_j=1} \leftarrow c_{k_j}^{v'_j=1} + 1$;
 ⑫ else if $v'_j = -1$ then
 ⑬ $c_{k_j}^{v'_j=-1} \leftarrow c_{k_j}^{v'_j=-1} + 1$;
 ⑭ end if
 ⑮ 收集者计算 $f_{k_j}^* = \frac{\frac{c_{k_j}^{v'_j=1} + c_{k_j}^{v'_j=-1}}{(1-\beta)n} - 2q_2}{p_1 + q_1 - 2q_2} \ell$;
 ⑯ 收集者计算 $\mu_{k_j}^* = \frac{c_{k_j}^{v'_j=1} - c_{k_j}^{v'_j=-1}}{(p_1 - q_1)(1-\beta)n f_{k_j}^*} \ell$;
 ⑰ $f^* \leftarrow f^* \cup f_{k_j}^*, \mu^* \leftarrow \mu^* \cup \mu_{k_j}^*$;
 ⑱ end for
 ⑲ return f^*, μ^* .

HISKV 算法是基于本地差分隐私解决键的频率以及所对应值的估计问题.首先收集者从 n 个用户中抽取 βn 个用户 (β 为抽取概率) (步骤②).基于 βn 个用户利用 OETL 算法估计最优截断长度 ℓ (步骤③).每个用户结合截断长度 ℓ 对自身所拥有的键值对集合进行截断或者填充.然后再均匀随机抽样一个键-值对本地扰动,并把扰动结果报告给收集者 (步骤④~⑦).收集者聚集所有用户的报告值后,估计每个键的频率以及值所对应的均值 (步骤⑧~⑱).由 LRR_KV 算法中的抽样可知,每个用户仅扰动一组键-值对报告给收集者,进而使得收集者与用户之间的通信代价为 $O(1)$.接下来阐述 LRR_KV 与 OETL 算法,其中 LRR_KV 算法具体细节如算法 2 所示:

算法 2. LRR_KV 算法.

输入: $(1-\beta)n$ 用户中任意 u_i 的键-值集合 S_i 、隐私预算 ϵ 、截断长度 ℓ ;

输出: 扰动之后的值 $(j, \langle k'_j, v'_j \rangle)$.

- ① $\langle k_j, v_j \rangle \leftarrow \text{Padding-Sampling}(S_i, \ell)$;
/* 填充-抽样 */
- ② $\langle k_j, \pm 1 \rangle \leftarrow \text{Discretizing}(\langle k_j, v_j \rangle)$;
/* 离散化 */
- ③ if the j -th pair $\langle k_j, v_j \rangle$ is $\langle k_j, 1 \rangle$ then
- ④ 扰动 $\langle k_j, v_j \rangle$ 为 $\langle k'_j, v'_j \rangle$;

$\langle k'_j, v'_j \rangle =$

$$\begin{cases} \langle k_j, 1 \rangle, \text{以概率 } \frac{e^\epsilon}{e^\epsilon + d' - 1}; \\ \langle k_j, -1 \rangle, \text{以概率 } \frac{(d' - 1)e^\epsilon}{(e^\epsilon + d' - 1)(e^\epsilon + d' - 2)}; \\ \langle k'_j, \pm 1 \rangle, \text{以概率 } \frac{d' - 1}{(e^\epsilon + d' - 1)(e^\epsilon + d' - 2)}. \end{cases}$$

⑤ else if the j -th pair $\langle k_j, v_j \rangle$ is $\langle k_j, -1 \rangle$ then

⑥ 扰动 $\langle k_j, v_j \rangle$ 为 $\langle k'_j, v'_j \rangle$;

$\langle k'_j, v'_j \rangle =$

$$\begin{cases} \langle k_j, 1 \rangle, \text{以概率 } \frac{(d' - 1)e^\epsilon}{(e^\epsilon + d' - 1)(e^\epsilon + d' - 2)}; \\ \langle k_j, -1 \rangle, \text{以概率 } \frac{e^\epsilon}{e^\epsilon + d' - 1}; \\ \langle k'_j, \pm 1 \rangle, \text{以概率 } \frac{d' - 1}{(e^\epsilon + d' - 1)(e^\epsilon + d' - 2)}. \end{cases}$$

⑦ end if
 ⑧ return j 和 $\langle k'_j, v'_j \rangle$.

结合 LRR_KV 算法,用户 u_i 首先利用算法 3 步骤①对集合 S_i 进行填充-抽样操作,进而获得 $\langle k_j, v_j \rangle$.结合 $\langle k_j, v_j \rangle$,用户 u_i 利用式(5)对值 v_j 进行离散化操作 (步骤②).然后结合离散化结果 $\langle k_j, \pm 1 \rangle$ 进行本地扰动 (步骤②~⑦).结合用户 u_i 的键-值对集合 S_i ,LRR_KV 算法的具体实现思路如图 3 所示:

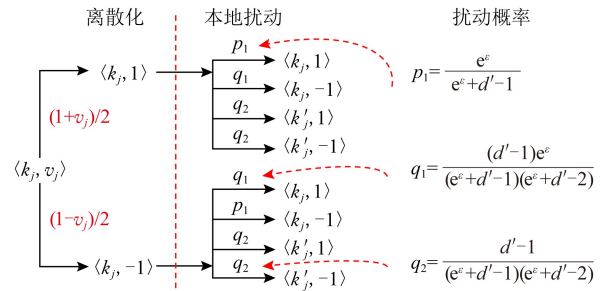


Fig. 3 LRR_KV local disturbance

图 3 LRR_KV 本地扰动示意图

根据图 3 可知,用户 u_i 所抽取的键-值对 $\langle k_j, v_j \rangle$ 经过离散化后变成 $\langle k_j, 1 \rangle$ 或者 $\langle k_j, -1 \rangle$.以概率 p_1 , q_1, q_2 分别将 $\langle k_j, 1 \rangle$ 与 $\langle k_j, -1 \rangle$ 扰动成 4 种不同的情况.其中一种情况是 $\langle k_j, 1 \rangle$ 与 $\langle k_j, -1 \rangle$ 以概率 p_1 扰动成自身.由式(2)可知, $p_1 = e^\epsilon / (e^\epsilon + d' - 1)$, 其中, $d' = 2(d + \ell)$.如何设置剩余 3 种的扰动概率是一个大的挑战.文献[3]中 PCKV-GRR 算法在相应参数取得最优时,剩余 3 种概率均取值为 $1/(e^\epsilon + d' - 1)$.然而,该算法却忽略一种事实,即不论 $\langle k_j, 1 \rangle$ 以

概率 q_1 扰动成 $\langle k_j, -1 \rangle$ 或者 $\langle k_j, 1 \rangle$ 以概率 q_1 扰动成 $\langle k_j, 1 \rangle$, 键 k_j 并没有被扰动成值域 d' 中的其他值. 此种情况对键 k_j 的频率及其所对应值的均值估计均有效. 因此, 概率 q_1 的值不应该取平均值, 应该倾向性地取较大的值. 由定义 1 可知下式成立:

若 $q_1 + (d' - 2)q_2 = 1 - p_1 = \frac{d' - 1}{e^\epsilon + d' - 1}, \frac{q_1}{q_2} \leq e^\epsilon$, 则:

$$\begin{cases} q_1 = \frac{(d' - 1)e^\epsilon}{(e^\epsilon + d' - 1)(e^\epsilon + d' - 2)}, \\ q_2 = \frac{d' - 1}{(e^\epsilon + d' - 1)(e^\epsilon + d' - 2)}. \end{cases}$$

影响 LRR_KV 算法本地扰动效果包括 2 个因素: Padding-Sampling 算法的设计以及如何获得最优截断长度 ℓ . 接下来结合文献[5]设计填充-抽样算法.

算法 3. Padding-Sampling 算法.

输入: S_i 、截断长度 ℓ ;

输出: 所抽取的键-值对 $\langle k_j, v_j \rangle$.

- ① if $l_i > \ell$ then /* l_i 表示 S_i 的长度 */
- ② u_i 不放回地从 S_i 随机采样 ℓ 个键值对, 得到 S'_i ; /* 抽取 ℓ 个键-值对 */
- ③ else if $l_i \leq \ell$ then
- ④ u_i 用 $\ell - l_i$ 个从 $\{\langle \overline{\tau}_1, 0 \rangle, \langle \overline{\tau}_2, 0 \rangle, \dots, \langle \overline{\tau}_l, 0 \rangle\}$ 集合采样的哑元填充 S_i , 得到 S'_i ;
- ⑤ u_i 从 S'_i 中均匀随机采样, 取第 j 个键值对 $\langle k_j, v_j \rangle$; /* 抽取第 j 个键-值对 */
- ⑥ end if
- ⑦ return $\langle k_j, v_j \rangle$.

用户 u_i 利用 S_i 的长度 l_i 与截断长度 ℓ 比较. 如果 $l_i > \ell$, 则从 S_i 中无放回地抽取 ℓ 个键-值对形成 S'_i (步骤①②); 否则, 从哑元键-值对集合 $\{\langle \overline{\tau}_1, 0 \rangle, \langle \overline{\tau}_2, 0 \rangle, \dots, \langle \overline{\tau}_l, 0 \rangle\}$ 中随机抽取 $\ell - l_i$ 个键-值对填充 S_i , 并获得 S'_i (步骤③④). 结合 S'_i , 用户以概率 $1/\ell$ 抽取一个键-值对 $\langle k_j, v_j \rangle$.

根据算法 2 与算法 3 可知, 最优截断长度 ℓ 的选择直接制约着最终频率与均值的估计精度. ℓ 过小, 键的频率以及均值都会被低估. 反之 ℓ 过大, 会导致频率与均值的估计误差过大. 若能够获得用户所拥有的键-值对长度分布 (如图 1 所示), 则可以获得最优的 ℓ . 由于键-值对长度是敏感信息, 用户不能直接发送给收集者. 因此, 结合文献[5]中的第 90 个百分位数求解法, 利用 GRR 算法与 βn 个用户来估计 ℓ , 具有细节如 OETL 算法所示.

算法 4. OETL 算法.

输入: n, β, ϵ 、键的值域 d ;

输出: ℓ .

- ① for 在 $[\beta n]$ 的每个用户 u_i 执行步骤②~③;
/* 每个用户扰动 */
- ② 用户 u_i 报告 $l'_i \leftarrow \text{GRR}(l_i, \epsilon)$;
- ③ 用户 u_i 发送 l'_i 到收集端;
- ④ end for
- ⑤ for 在 $[d]$ 中的每个 l'_i 执行步骤⑥~⑦:
- ⑥ 收集者计算长度 l'_i 的计数 $c(l'_i)$;
- ⑦ 收集者无偏修正 $c(l'_i)$ 后获得 $c'(l'_i)$:
$$c'(l'_i) \leftarrow \frac{(e^\epsilon + d - 1)c(l'_i) - \beta n}{e^\epsilon - 1};$$
- ⑧ end for
- ⑨ 收集者估计最小的 ℓ , 使其满足:
$$\ell \leftarrow \sum_{l'_i=1}^{\ell} c'(l'_i) > 0.9 \times \sum_{l'_i=1}^d c'(l'_i);$$
- ⑩ 收集者发送 ℓ 给 $[(1 - \beta)n]$ 集合中的每个用户;
- ⑪ return ℓ .

βn 中的每个用户利用 GRR 算法本地扰动自己键-值对长度 l_i , 步骤①~④发送给收集者; 步骤⑤~⑧收集者计算并修正每个长度的计数; 步骤⑨利用第 90 个百分位数求解 ℓ 后; 步骤⑩收集者将 ℓ 发送给 $(1 - \beta)n$ 中的每个用户.

3.2 HISKV 隐私性与可用性分析

本节主要从本地差分隐私阐述 HISKV 算法的隐私性以及无偏性估计、方差及其最大偏差来阐述其可用性. 从算法 1 可知, HISKV 算法采用用户分组的形式分别估计频率与均值 (LRR_KV 算法) 以及最优截断长度 (OETL 算法).

定理 1. OETL 算法满足 ϵ -本地差分隐私.

证明. 结合 $[\beta n]$ 集合中用户所给定任意的 l_i 与 l_j , 算法 OETL 的输出为 l' , 则不等式成立:

$$\frac{\Pr[\text{OETL}(l_i) = l']}{\Pr[\text{OETL}(l_j) = l']} \leq \frac{e^\epsilon / (e^\epsilon + d - 1)}{1 / (e^\epsilon + d - 1)} = e^\epsilon,$$

则 OETL 算法满足 ϵ -本地差分隐私.

证毕.

定理 2. LRR_KV 算法满足 ϵ -本地差分隐私.

证明. 设 $\langle k_j, v_j \rangle$ 与 $\langle k_i, v_i \rangle$ 表示 2 个不同的输入键-值对, $\langle k^*, v^* \rangle$ 表示 $\langle k_j, v_j \rangle$ 与 $\langle k_i, v_i \rangle$ 经过 LRR_KV 算法本地扰动之后的输出结果, 根据图 3 的表示可知, $\langle k^*, v^* \rangle \in \{\langle k_j, 1 \rangle, \langle k_j, -1 \rangle, \langle k'_j, 1 \rangle, \langle k'_j, -1 \rangle\}$. 分 4 种情况证明 LRR_KV 算法满足本地差分隐私.

1) 当 $\langle k^*, v^* \rangle = \langle k_j, 1 \rangle$ 时, 根据定义 1 可知:

$$\begin{aligned} \frac{Pr[\langle k^*, v^* \rangle = \langle k_j, 1 \rangle | \langle k_j, v_j \rangle]}{Pr[\langle k^*, v^* \rangle = \langle k_j, 1 \rangle | \langle k_i, v_i \rangle]} &\leq \\ \frac{Pr[\langle k^*, v^* \rangle = \langle k_j, 1 \rangle | \langle k_j, v_j \rangle = \langle k_j, 1 \rangle]}{Pr[\langle k^*, v^* \rangle = \langle k_j, 1 \rangle | \langle k_i, v_i \rangle = \langle k_j, -1 \rangle]} &= \\ \frac{p_1}{q_1} = \frac{e^\epsilon / (e^\epsilon + d' - 1)}{(d' - 1)e^\epsilon / (e^\epsilon + d' - 1)(e^\epsilon + d' - 2)} &= \\ \frac{e^\epsilon + d' - 2}{d' - 1} \leq e^\epsilon. \end{aligned}$$

2) 当 $\langle k^*, v^* \rangle = \langle k_j, -1 \rangle$ 时, 类似于第 1 种情况:

$$\begin{aligned} \frac{Pr[\langle k^*, v^* \rangle = \langle k_j, -1 \rangle | \langle k_j, v_j \rangle]}{Pr[\langle k^*, v^* \rangle = \langle k_j, -1 \rangle | \langle k_i, v_i \rangle]} &\leq \\ \frac{Pr[\langle k^*, v^* \rangle = \langle k_j, -1 \rangle | \langle k_j, v_j \rangle = \langle k_j, -1 \rangle]}{Pr[\langle k^*, v^* \rangle = \langle k_j, -1 \rangle | \langle k_i, v_i \rangle = \langle k_j, 1 \rangle]} &= \\ \frac{p_1}{q_1} &\leq e^\epsilon. \end{aligned}$$

3) 当 $\langle k^*, v^* \rangle = \langle k'_j, 1 \rangle$ 时, 由定义 1 与图 3 可知:

$$\begin{aligned} \frac{Pr[\langle k^*, v^* \rangle = \langle k'_j, 1 \rangle | \langle k_j, v_j \rangle]}{Pr[\langle k^*, v^* \rangle = \langle k'_j, 1 \rangle | \langle k_i, v_i \rangle]} &\leq \\ \frac{Pr[\langle k^*, v^* \rangle = \langle k'_j, 1 \rangle | \langle k_j, v_j \rangle = \langle k_j, 1 \rangle]}{Pr[\langle k^*, v^* \rangle = \langle k'_j, 1 \rangle | \langle k_i, v_i \rangle = \langle k_j, -1 \rangle]} &= \\ \frac{q_2}{q_2} &\leq e^\epsilon. \end{aligned}$$

4) 当 $\langle k^*, v^* \rangle = \langle k'_j, -1 \rangle$ 时, 类似于第 3 种情况:

$$\begin{aligned} \frac{Pr[\langle k^*, v^* \rangle = \langle k'_j, -1 \rangle | \langle k_j, v_j \rangle]}{Pr[\langle k^*, v^* \rangle = \langle k'_j, -1 \rangle | \langle k_i, v_i \rangle]} &\leq \\ \frac{Pr[\langle k^*, v^* \rangle = \langle k'_j, -1 \rangle | \langle k_j, v_j \rangle = \langle k_j, -1 \rangle]}{Pr[\langle k^*, v^* \rangle = \langle k'_j, -1 \rangle | \langle k_i, v_i \rangle = \langle k_j, 1 \rangle]} &= \\ \frac{q_2}{q_2} &\leq e^\epsilon. \end{aligned}$$

通过 4 种情况可知 LRR_KV 算法满足 ϵ -本地差分隐私。证毕。

由定理 1 与定理 2 以及差分隐私的序列组合性质可知, HISKV 收集键-值数据的过程满足本地差分隐私。HISKV 的可用性主要从 LRR_KV 算法与 OETL 算法的误差来度量。OETL 算法的误差主要来自于 GRR 算法的本地扰动操作, 根据文献[7]可知, OETL 算法产生的方差为 $\beta n (e^\epsilon + d - 2) / (e^\epsilon - 1)^2$ 。LRR_KV 算法的可用性主要从无偏性、方差、最大偏差来度量。尽管收集者可以估计出键 k 的频率 f_k^* 以及所对应值的均值 μ_k^* , 但由于 LRR_KV

算法的本地扰动使得 f_k^* 与 μ_k^* 会偏离真实值。而我们期望 f_k^* 与 μ_k^* 均要满足无偏性, 即 $E[f_k^*] = f_k$ 与 $E[\mu_k^*] = \mu_k$ 成立。

定理 3. 假设 f_k^* 与 f_k 分别为键 k 的估计频率与真实频率, μ_k^* 与 μ_k 分别为键 k 所对应值的估计均值与真实均值。当已知概率 p_1, q_1, q_2 时, 无偏估计 $E[f_k^*] = f_k, E[\mu_k^*] = \mu_k$ 成立。

证明. 为了证明 $E[f_k^*] = f_k$, 首先要获得统计量 f_k^* 的表达式。 $(1-\beta)n$ 个用户中回答 k 的概率和回答非 k 的概率为

$$\begin{aligned} Pr[k] &= \frac{f_k}{\ell} (p_1 + q_1) + \left(1 - \frac{f_k}{\ell}\right) 2q_2, \\ Pr[k'] &= \frac{f_k}{\ell} (1 - p_1 - q_1) + \left(1 - \frac{f_k}{\ell}\right) (1 - 2q_2). \end{aligned}$$

根据二项分布构造极大似然函数:

$$L(f_k) = (Pr[k])^{(c_k^{v'=1} + c_k^{v'=-1})} \times (Pr[k'])^{(1-\beta)n - (c_k^{v'=1} + c_k^{v'=-1})},$$

其中, $c_k^{v'=1}$ 与 $c_k^{v'=-1}$ 分别表示收集者收集到的 $\langle k, 1 \rangle$ 和 $\langle k, -1 \rangle$ 的计数。

对极大似然函数 $L(f_k)$ 的等号两侧取对数, 再对 $L(f_k)$ 求导, 即可获得 f_k 的估计量 $f_k^* = \frac{c_k^{v'=1} + c_k^{v'=-1}}{(1-\beta)n} - 2q_2 \frac{p_1 + q_1 - 2q_2}{p_1 + q_1 - 2q_2} \ell$ 。

$$\begin{aligned} E[f_k^*] &= E \left[\frac{c_k^{v'=1} + c_k^{v'=-1}}{(1-\beta)n} - 2q_2 \frac{p_1 + q_1 - 2q_2}{p_1 + q_1 - 2q_2} \ell \right] = \\ &= \frac{1}{\ell} \frac{E[c_k^{v'=1} + c_k^{v'=-1}] - 2q_2}{p_1 + q_1 - 2q_2}. \end{aligned} \quad (6)$$

因此, 式(6)中只要获得 $E[c_k^{v'=1} + c_k^{v'=-1}]$ 即可证明 $E[f_k^*] = f_k$ 成立。 $c_k^{v'=1}$ 与 $c_k^{v'=-1}$ 可分别表示为

$$\begin{aligned} c_k^{v'=1} &= (1-\beta)n \left(\frac{1}{\ell} (f_k^{v=1} p_1 + f_k^{v=-1} q_1) + \left(1 - \frac{f_k^{v=1} + f_k^{v=-1}}{\ell}\right) q_2 \right), \\ c_k^{v'=-1} &= (1-\beta)n \left(\frac{1}{\ell} (f_k^{v=-1} p_1 + f_k^{v=1} q_1) + \left(1 - \frac{f_k^{v=1} + f_k^{v=-1}}{\ell}\right) q_2 \right), \end{aligned}$$

其中, $f_k^{v=1}$ 与 $f_k^{v=-1}$ 表示键-值对 $\langle k, 1 \rangle$ 和 $\langle k, -1 \rangle$ 的真实频率。进而可知 $f_k = f_k^{v=1} + f_k^{v=-1}$ 。

由 $c_k^{v'=1}$ 与 $c_k^{v'=-1}$ 的表达式可获得:

$$E[c_k^{v'=1} + c_k^{v'=-1}] = (1-\beta)n \left(\frac{f_k}{\ell} (p_1 + q_1 - 2q_2) + 2q_2 \right). \quad (7)$$

将式(7)带入式(6)可推理出:

$$E[f_k^*] = \ell \frac{\frac{f_k}{\ell} (p_1 + q_1 - 2q_2)}{p_1 + q_1 - 2q_2} = f_k,$$

则 $E[f_k^*] = f_k$ 成立.

接下来证明 $E[\mu_k^*] = \mu_k$ 成立.

根据式(4)可知,若要使 $E[\mu_k^*] = \mu_k$ 成立,只要 $E[v_j'] = v_j$ 成立即可.根据 LRR_KV 算法的扰动概率可知, v_j 的输出结果 v_j' 的值域为 $[-1, 1]$.如果知道这 2 种值的输出概率,即可计算出输出 v_j' 的期望值.设 $Pr[-1]$ 和 $Pr[1]$ 分别表示 -1 与 1 的输出概率,则等式成立:

$$\begin{aligned} Pr[1] &= \frac{1+v_j}{2}p_1 + \frac{1-v_j}{2}q_1 + \frac{1+v_j}{2}q_2 + \frac{1-v_j}{2}q_2 \\ &= \frac{1+v_j}{2}(p_1+q_2) + \frac{1-v_j}{2}(q_1+q_2), \\ Pr[-1] &= \frac{1+v_j}{2}q_1 + \frac{1-v_j}{2}p_1 + \frac{1+v_j}{2}q_2 + \frac{1-v_j}{2}q_2 \\ &= \frac{1-v_j}{2}(p_1+q_2) + \frac{1+v_j}{2}(q_1+q_2). \end{aligned}$$

则 $E[v_j']$ 可以表示为

$$\begin{aligned} E[v_j'] &= 1 \times \left[\frac{1+v_j}{2}(p_1+q_2) + \frac{1-v_j}{2}(q_1+q_2) \right] + \\ &(-1) \times \left[\frac{1-v_j}{2}(p_1+q_2) + \frac{1+v_j}{2}(q_1+q_2) \right] = \\ &v_j(p_1 - q_1), \end{aligned} \quad (8)$$

其中, $p_1 - q_1 = \frac{e^\epsilon(e^\epsilon - 1)}{(e^\epsilon + d' - 1)(e^\epsilon + d' - 2)}$.

根据式(8)可知,若使 $E[v_j'] = v_j$ 成立,式(8)右端乘以 $1/(p_1 - q_1)$ 即可,其中 $1/(p_1 - q_1)$ 为修正因子.根据 $E[v_j'] = v_j$, 可知 $E[\mu_k^*] = \mu_k$ 成立.

证毕.

尽管定理 3 保证了键-值的频率与均值的无偏性,然而由于利用 LRR_KV 算法对用户的键-值对进行本地扰动,在估计 f_k^* 与 μ_k^* 时会产生相应的误差.本文采用方差 $Var[f_k^*]$ 与 $Var[\mu_k^*]$ 度量 LRR_KV 算法产生的误差.

定理 4. 假设 f_k^* 与 f_k 分别为键 k 的估计频率与真实频率.当已知概率 p_1, q_1, q_2 时, f_k^* 的方差为

$$Var[f_k^*] = \left(\frac{\ell}{(1-\beta)n(p_1 + q_1 - 2q_2)} \right)^2 Var[c_k^{v'=1} + c_k^{v'=-1}].$$

证明. 由定理 3 可知 $f_k^* = \frac{c_k^{v'=1} + c_k^{v'=-1}}{(1-\beta)n} - 2q_2$, 则:

$$\begin{aligned} Var[f_k^*] &= Var \left[\frac{c_k^{v'=1} + c_k^{v'=-1}}{(1-\beta)n} - 2q_2 \right] = \\ &\left(\frac{\ell}{(1-\beta)n(p_1 + q_1 - 2q_2)} \right)^2 Var[c_k^{v'=1} + c_k^{v'=-1}], \end{aligned} \quad (9)$$

$$\begin{aligned} Var[c_k^{v'=1} + c_k^{v'=-1}] &= (1-\beta)n \frac{f_k}{\ell} (p_1 + q_1) \times \\ &(1 - p_1 - q_1) + (1-\beta)n \left(1 - \frac{f_k}{\ell} \right) 2q_2 (1 - 2q_2). \end{aligned} \quad (10)$$

将式(10)带入式(9)可获得 $Var[f_k^*]$,

$$\begin{aligned} Var[f_k^*] &= Var \left[\frac{c_k^{v'=1} + c_k^{v'=-1}}{(1-\beta)n} - 2q_2 \right] = \\ &\frac{\ell(f_k(p_1 + q_1)(1 - p_1 - q_1) + (\ell - f_k)2q_2(1 - 2q_2))}{(1-\beta)n(p_1 + q_1 - 2q_2)^2}. \end{aligned}$$

证毕.

定理 5. 设 u_k^* 与 μ_k 分别为 k 所对应值的估计均值与真实均值.当已知概率 p_1, q_1, q_2 时, u_k^* 的方差为

$$\begin{aligned} Var[\mu_k^*] &= \\ &\left(\frac{p_1 + q_1 - 2q_2}{p_1 - q_1} \right)^2 Var \left[\frac{c_k^{v'=1} - c_k^{v'=-1}}{c_k^{v'=1} + c_k^{v'=-1} - 2n'q_2} \right], \end{aligned}$$

其中 $n' = (1-\beta)n$.

证明. 根据算法 1 的步骤⑮⑯可知 u_k^* 的表达式为 $u_k^* = \frac{c_k^{v'=1} - c_k^{v'=-1}}{c_k^{v'=1} + c_k^{v'=-1} - 2n'q_2} \times \frac{p_1 + q_1 - 2q_2}{p_1 - q_1}$. 则 u_k^* 的方差可表示为

$$Var[\mu_k^*] = Var \left(\frac{c_k^{v'=1} - c_k^{v'=-1}}{c_k^{v'=1} + c_k^{v'=-1} - 2n'q_2} \times \frac{p_1 + q_1 - 2q_2}{p_1 - q_1} \right). \quad (11)$$

分别设置 $R = c_k^{v'=1} - c_k^{v'=-1}$, $S = c_k^{v'=1} + c_k^{v'=-1} - 2n'q_2$. 则 $Var[\mu_k^*] = Var \left[\frac{R}{S} \right] \left(\frac{p_1 + q_1 - 2q_2}{p_1 - q_1} \right)^2$. 因此,只要计算出 $Var \left[\frac{R}{S} \right]$, 即可获得 $Var[\mu_k^*]$. 根据

文献[3]可知, $Var \left[\frac{R}{S} \right]$ 可以表示为

$$\begin{aligned} Var \left[\frac{R}{S} \right] &\approx \frac{Var[R]}{(E[S])^2} - \frac{2E[R]}{(E[S])^3} Cov(R, S) + \\ &\frac{(E[R])^2}{(E[S])^4} Var[S], \end{aligned} \quad (12)$$

$$E[R] = E[c_k^{v'=1} - c_k^{v'=-1}] = \frac{p_1 - q_1}{\ell} (c_k^{v=1} - c_k^{v=-1}),$$

$$E[S] = E[c_k^{v'=1} + c_k^{v'=-1} - 2n'q_2] = \frac{p_1 + q_1 - 2q_2}{\ell} (c_k^{v=1} + c_k^{v=-1}).$$

$$\begin{aligned} \text{Var}[R] &= \text{Var}[c_k^{v'=1}] + \text{Var}[c_k^{v'=-1}] = \\ &= (c_k^{v=1} + c_k^{v=-1})[(p_1 - q_2)(1 - p_1 - q_2) + \\ &+ (q_1 - q_2)(1 - q_1 - q_2)] + 2n'q_2(1 - q_2), \\ \text{Var}[S] &= \text{Var}[c_k^{v'=1}] + \text{Var}[c_k^{v'=-1}] = \\ &= (c_k^{v=1} + c_k^{v=-1})[(p_1 - q_2)(1 - p_1 - q_2) + \\ &+ (q_1 - q_2)(1 - q_1 - q_2)] + 2n'q_2(1 - q_2), \end{aligned}$$

$$\text{Cov}(R, S) = \text{Cov}(c_k^{v'=1} - c_k^{v'=-1},$$

$$\begin{aligned} c_k^{v'=1} + c_k^{v'=-1} - 2n'q) &= \text{Var}[c_k^{v'=1}] - \\ \text{Var}[c_k^{v'=-1}] &= (c_k^{v=1} - c_k^{v=-1})[(p_1 - q_2) \\ &+ (1 - p_1 - q_2) - (q_1 - q_2)(1 - q_1 - q_2)]. \text{证毕.} \end{aligned}$$

上述 $E[R], E[S], \text{Var}[R], \text{Var}[S], \text{Cov}(R, S)$ 中的 $c_k^{v=1}$ 与 $c_k^{v=-1}$ 分别表示键-值对 $\langle k, 1 \rangle$ 和 $\langle k, -1 \rangle$ 的真实计数. 将 $E[R], E[S], \text{Var}[R], \text{Var}[S], \text{Cov}(R, S)$ 带入式(12)即可计算出 $\text{Var}\left[\frac{R}{S}\right]$, 进而可计算出 $\text{Var}[\mu_k^*]$. 证毕.

根据定理4与定理5可以计算出每个键 k 及其对应均值的方差. 而如何度量 f_k^* 与 f_k 以及 μ_k^* 与 μ_k 之间的最大偏差是一个挑战性问题, 由定理6与定理7给出详细说明.

定理6. 设 f_k^* 与 f_k 分别为 k 的估计频率与真实频率. 当已知 p_1, q_1, q_2 时,

$$|f_k - f_k^*| = O\left(\frac{\ell \sqrt{\ln(\ell/\eta)}}{\epsilon \sqrt{(1-\beta)n}}\right)$$

至少以概率 $1-\eta$ 成立, 其中 n 为用户个数, ℓ 为最优截断长度, β 为用户分组抽样率.

证明. 设 c'_k 与 c_k 分别表示键 k 的估计计数与真实计数, 则

$$\begin{aligned} |f_k - f_k^*| &= \frac{|c'_k - c_k|}{(1-\beta)n} = \frac{\ell}{(1-\beta)n} \times \\ &\left| \sum_{j=1}^{(1-\beta)n} \frac{I_{k_j} - 2q_2}{p_1 + q_1 - 2q_2} - \sum_{j=1}^{(1-\beta)n} \frac{I_{k'_j} - 2q_2}{p_1 + q_1 - 2q_2} \right| = \\ &\frac{\ell}{(1-\beta)n} \sum_{j=1}^{(1-\beta)n} \left| \frac{I_{k_j} - I_{k'_j}}{p_1 + q_1 - 2q_2} \right|, \end{aligned}$$

$$\text{其中}, I_{k_j} = \begin{cases} 1, & k_j = k; \\ 0, & k_j \neq k; \end{cases}, I_{k'_j} = \begin{cases} 1, & k'_j = k; \\ 0, & k'_j \neq k. \end{cases}$$

可知 $I_{k_j} - I_{k'_j}$ 为一个随机变量, 对应的方差可以表示为

$$\text{Var}[I_{k_j} - I_{k'_j}] = E[(I_{k_j} - I_{k'_j})^2] -$$

$$\begin{aligned} [E(I_{k_j} - I_{k'_j})]^2 &= \frac{1 - p_1 - q_1 + 2q_2}{2} - \\ &\frac{(1 - p_1 - q_1 - 2q_2)^2}{4} = O(1/\epsilon^2). \end{aligned}$$

由 $I_{k_j} \in \{0, 1\}, I_{k'_j} \in \{0, 1\}$ 可知 $I_{k_j} - I_{k'_j} \in \{-1, 0, 1\}$, 进而可知 $|I_{k_j} - I_{k'_j}| \leq 1$. 则

$$\left| \frac{I_{k_j} - I_{k'_j}}{p_1 + q_1 - 2q_2} \right| \leq \frac{(\epsilon^\epsilon + d' - 1)(\epsilon^\epsilon + d' - 2)}{(\epsilon^\epsilon - 1)(\epsilon^\epsilon + 2d' - 2)},$$

$$\Pr[|f_k - f_k^*| \geq \lambda/(1-\beta)n] =$$

$$\Pr[|c_k - c'_k| \geq \lambda] =$$

$$\Pr\left[\frac{\ell}{(1-\beta)n} \sum_{j=1}^{(1-\beta)n} \left| \frac{I_{k_j} - I_{k'_j}}{p_1 + q_1 - 2q_2} \right| \geq \lambda\right] =$$

$$\Pr\left[\sum_{j=1}^{(1-\beta)n} \left| \frac{I_{k_j} - I_{k'_j}}{p_1 + q_1 - 2q_2} \right| \geq \frac{(1-\beta)n\lambda}{\ell}\right].$$

则可知:

$$\Pr[|f_k - f_k^*| \geq \lambda/(1-\beta)n] \leq$$

$$\begin{aligned} 2e^{-\left(\frac{\frac{(1-\beta)n^2\lambda^2}{2\ell^2}}{\sum_{j=1}^{(1-\beta)n} \text{Var}\left[\frac{I_{k_j} - I_{k'_j}}{p_1 + q_1 - 2q_2}\right] + \frac{(1-\beta)n\lambda(\epsilon^\epsilon + d' - 1)(\epsilon^\epsilon + d' - 2)}{3\ell(\epsilon^\epsilon - 1)(\epsilon^\epsilon + 2d' - 2)}}\right)} = \\ 2e^{-\left(\frac{\frac{(1-\beta)n\lambda^2}{2}}{\ell^2 O(1/\epsilon^2) + O(1/\epsilon)}\right)}. \end{aligned} \quad (13)$$

结合式(13)可知 $\lambda = O\left(\frac{\ell \sqrt{\ln(\ell/\eta)}}{\epsilon \sqrt{(1-\beta)n}}\right)$, 则可知

$|f_k - f_k^*| < \lambda/(1-\beta)n$ 至少以概率 $1-\eta$ 成立.

证毕.

定理7. 设 u_k^* 与 μ_k 分别为 k 所对应值的估计均

值与真实均值. 不等式 $|\mu_k - \mu_k^*| \leq \frac{1}{\frac{(1-\beta)nf_k}{\lambda} - 1}$ 至

少以概率 $1-\eta$ 成立. 其中 n 为用户个数, f_k 为 k 的真实频率.

证明. 为了证明方便, 令 $c_k^{v*=1}$ 与 $c_k^{v*=-1}$ 分别表示 $c_k^{v'=1}$ 与 $c_k^{v'=-1}$ 的修正值. 因此, u_k^* 与 f_k^* 可以分别表示为

$$u_k^* = \frac{c_k^{v*=1} - c_k^{v*=-1}}{c_k^{v*=1} + c_k^{v*=-1}},$$

$$f_k^* = \frac{\ell}{(1-\beta)n} (c_k^{v*=1} + c_k^{v*=-1}).$$

令 $\theta_1 = c_k^{v*=1} - c_k^{v=1}, \theta_{-1} = c_k^{v*=-1} - c_k^{v=-1}$, 其

中, $c_k^{v=1}$ 与 $c_k^{v=-1}$ 使得 $f_k = \frac{\ell}{(1-\beta)n} (c_k^{v=1} + c_k^{v=-1})$

成立. 根据 f_k 与 f_k^* 可知 $|f_k - f_k^*| = \frac{\ell}{(1-\beta)n} \times |\theta_1 + \theta_{-1}|$ 成立.

根据定理6可知 $\Pr\left(|f_k - f_k^*| < \frac{\lambda}{(1-\beta)n}\right) \geq$

$1-\eta$ 成立, 则 $Pr\left(\left|\theta_1+\theta_{-1}\right|<\frac{\lambda}{\ell}\right) \geq 1-\eta$ 成立. 进而使得不等式成立:

$$\begin{aligned} \left|u_k-u_k^*\right| &= \left|\frac{c_k^{v=1}-c_k^{v=-1}}{c_k^{v=1}+c_k^{v=-1}}-\frac{c_k^{v^*=1}-c_k^{v=-1}}{c_k^{v^*=1}+c_k^{v=-1}}\right|= \\ &= \left|\frac{c_k^{v=1}-c_k^{v=-1}}{c_k^{v=1}+c_k^{v=-1}}-\frac{c_k^{v=1}-c_k^{v=-1}+\theta_1-\theta_{-1}}{c_k^{v=1}+c_k^{v=-1}+\theta_1+\theta_{-1}}\right|= \\ &= \left|\frac{\left(c_k^{v=1}-c_k^{v=-1}\right)\left(\theta_1+\theta_{-1}\right)-\left(c_k^{v=1}+c_k^{v=-1}\right)\left(\theta_1-\theta_{-1}\right)}{\left(c_k^{v=1}+c_k^{v=-1}\right) \times\left(c_k^{v=1}+c_k^{v=-1}+\theta_1+\theta_{-1}\right)}\right|= \\ &= \left|\frac{\left(\theta_1+\theta_{-1}\right) u_k-\left(\theta_1-\theta_{-1}\right)}{\frac{(1-\beta) n f_k}{\ell}+\theta_1+\theta_{-1}}\right| \leq \\ &= \frac{\left|\theta_1\right|\left(1-u_k\right)+\left|\theta_{-1}\right|\left(1+u_k\right)}{\frac{(1-\beta) n f_k}{\ell}+\theta_1+\theta_{-1}} \leq \\ &= \frac{\frac{\lambda}{\ell}}{\frac{(1-\beta) n f_k}{\ell}-\frac{\lambda}{\ell}}=\frac{1}{\frac{(1-\beta) n f_k}{\lambda}-1} . \quad \text { 证毕. } \end{aligned}$$

4 实验结果与分析

文献[4]结论部分所给出的未来研究点正是本文的研究工作。因此,从实验平台、实验数据集以及隐私参数 ϵ 的设置均与文献[4]保持一致。实验平台是 4 核 Intel i7-4790 CPU(4 GHz), 8 GB 内存, Win7 系统, Python 实现所有算法. 采用 TalkingData, JData, GAUSS, LNR 四个数据集验证文中相关算法. 4 个数据集具体细节如表 2 所示:

Table 2 Characteristics of Datasets

表 2 数据集的属性

数据集	分布	用户数	Keys 数量
GAUSS	Gaussian	1 000 000	100
LNR	Linear	1 000 000	100
TalkingData		60 822	306
JData		105 180	442

其中 GAUSS 和 LNR 这 2 个数据集为合成数据集, 键和值分别遵循高斯和线性分布. TalkingData 是从与移动应用的 SDK 中收集而得, 包含 60 822 个设备和 306 类应用程序. JData 来自于 JD.com, 包含 2016 年的 5 060 万条销售记录, 该数据集记录了 2016 年间 105 180 个用户对 442 个品牌的 5 060 万条购买记录.

结合上述 4 个数据集, 采用均方误差 (mean

square error, MSE) 和相对误差 (relative error, RE) 度量 PrivKV^[1], KVOH^[2], PCKV-GRR^[3], LDPKV^[4], HISKV, RAPPOR^[8], Harmony^[14] 算法的频率和均值计算精度. 其中, RAPPOR 与 Harmony 分别是类别数据与数值型数据下频率与均值估计的典型代表. 键频率与相应均值所对应的均方误差 MSE 可以表示为 $\frac{1}{d} \sum_{k \in K}\left(f_k-f_k^*\right)^2$ 与 $\frac{1}{d} \sum_{k \in K}\left(u_k-u_k^*\right)^2$, 频率的相对误差 RE 可表示为 $\left|f_k-f_k^*\right| / f_k$. 其中, f_k 表示 k 的真实频率, f_k^* 表示 k 的估计频率; u_k 表示 k 所对应值的真实均值, u_k^* 表示 k 所对应值的估计均值.

隐私参数 ϵ 分别取值为 0.1, 0.2, 0.4, 0.8, 1.6, 3.2. 此外, 实验分别对 4 个数据集中频率最高的 top10, top20, top30, top40, top50 的键值进行统计, 在每种情况下进行 500 次实验并求取平均值作为最终的计算结果.

在文献[4]中, 我们分别实现了 LDPKV, PrivKV, KVOH, RAPPOR 以及 Harmony 算法, 并结合 MSE 与 RE 进行了综合对比分析. 因此, 本文主要对比 HISKV 算法与上述 6 种算法的性能.

1) 基于 TalkingData 数据集的多种算法的 MSE 和 RE 值比较.

图 4(a)~(f) 描述了 HISKV, PCKV-GRR, PrivKV, KVOH, LDPKV, RAPPOR, Harmony 算法的 MSE 值和 RE 值的比较结果. 由图 4(a) 的实验结果可以发现, 当 ϵ 从 0.1 变化到 3.2 时, 6 种算法的 MSE 值均减少, 而 HISKV 算法的精度优于其他 5 种算法. 当 $\epsilon=1.6$ 时, HISKV 算法的精度是 PCKV-GRR 算法的近 4 倍, 是 PrivKV 和 KVOH 算法的将近 3 倍, 是 LDPKV 算法的将近 1 倍. 尽管文献[4]中提到 LDPKV 算法优于 PrivKV 和 KVOH 算法, 然而由于该算法没有考虑到键-值的稀疏性(如图 1 所示), 过多的 $\langle 0,0\rangle$ 对使得该算法的 MSE 大于 HISKV 算法.

由图 4(c)~(f) 可知, 当 ϵ 固定, 键的数目从 top10 增加到 top50 时, 相应算法的 RE 值均增大, 但 HISKV 算法的精度却优于其他 4 种算法, 其原因在于 HISKV 算法利用最优截断长度削弱了键-值域稀疏带来的影响. 由图 4(b) 中 6 种算法的均值 MSE 比较结果可知, 当 ϵ 从 0.1 变化到 3.2 时, 6 种算法的均值 MSE 均呈现下降趋势, 而 HISKV 算法估计出的均值精度明显优于其他 5 种算法. 当 $\epsilon=1.6$ 时, HISKV 算法的均值精度是 LDPKV 算法

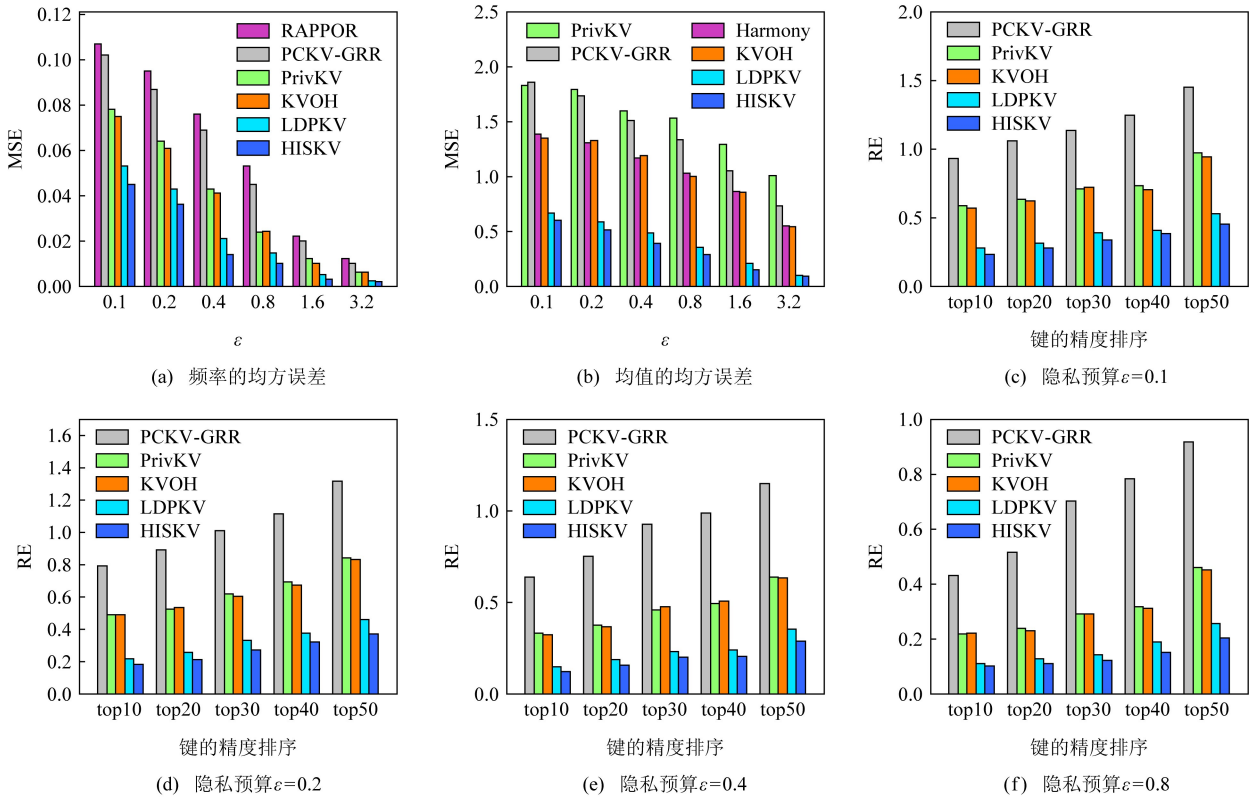


Fig. 4 The key-value frequency and mean estimation of TalkingData
图 4 基于 TalkingData 的键-值频率与均值估计结果

的将近 1 倍.其主要原因在于 HISKV 算法利用直方图将键-值对进行统一编码扰动,避免了隐私预算的分割,合理增大了扰动成真实键的本地扰动概率.此外, HISKV 算法利用截断-抽样技术避免了键-值稀疏性带来的影响,实现了隐私增强.

2) 基于 JData 数据集的多种算法的 MSE 和 RE 值比较.

图 5(a)~(f)描述了 JData 数据集上 7 种算法的 MSE 值和 RE 值的比较结果.根据图 5(a)的实验结果可知,当 ϵ 从 0.1 变化到 3.2 时,6 种算法的 MSE 值均减少.当 $\epsilon=1.6$ 时, HISKV 算法的精度是 PCKV-GRR 算法的近 6 倍,是 PrivKV 和 KVOH 算法的将近 3 倍,是 LDPKV 的将近 2 倍.其主要原因是 PCKV-GRR 算法没有顾及真实键本地扰动概率的倾向性设置问题, PrivKV 和 KVOH 算法隐私预算进行分割, LDPKV 算法没有考虑值域稀疏性带来的影响.图 5(b)描述了 HISKV 算法与其他 5 种算法有关均值 MSE 值的比较.当 ϵ 从 0.1 变化到 3.2 时, HISKV 算法的均值精度优于其他 5 种算法.当 $\epsilon=0.8$ 时, HISKV 算法的均值精度是 PrivKV 算法的近 4 倍,是 PCKV-GRR 算法的近 3 倍,是 Harmony

与 KVOH 的近 2 倍,是 LDPKV 算法的近 1 倍.在固定 ϵ 且变换 top10 到 top50 的情况下, HISKV 算法同样优于其他 4 种算法.其主要原因是 HISKV 算法利用抽样-截断技术、本地扰动概率的倾向性设置以及隐私预算不分割策略提升了键-值的频率与均值估计精度.

3) 基于 LNR, GAUSS 数据集的多种算法的 MSE 和 RE 值比较.

图 6 和图 7 描述了 7 种算法基于 LNR 与 GAUSS 合成数据集上的 MSE 与 RE 的对比结果.在百万级别的数据上,当 ϵ 从 0.1 变化到 3.2 时, HISKV 算法的精度优于其他 6 种算法,如图 6(a)、图 6(b)、图 7(a)、图 7(b)所示.文献[4]给出了 LDPKV 算法在 LNR 与 GAUSS 数据集上优于 PrivKV, KVOH, RAPPOR 以及 Harmony 算法.而本文的 HISKV 算法却明显有 LDPKV 算法,其主要原因在于 LNR 与 GAUSS 数据集中同样存在键-值的稀疏性问题,而 HISKV 算法的最优截断长度技术却可以有效减弱键-值稀疏性带来的影响.此外,当固定键值的个数且 ϵ 从 0.1 变化到 0.8 时,所有算法在 LNR 与 GAUSS 数据集上的 RE 值均下降,如图 6(c)~(f)、图 7(c)~

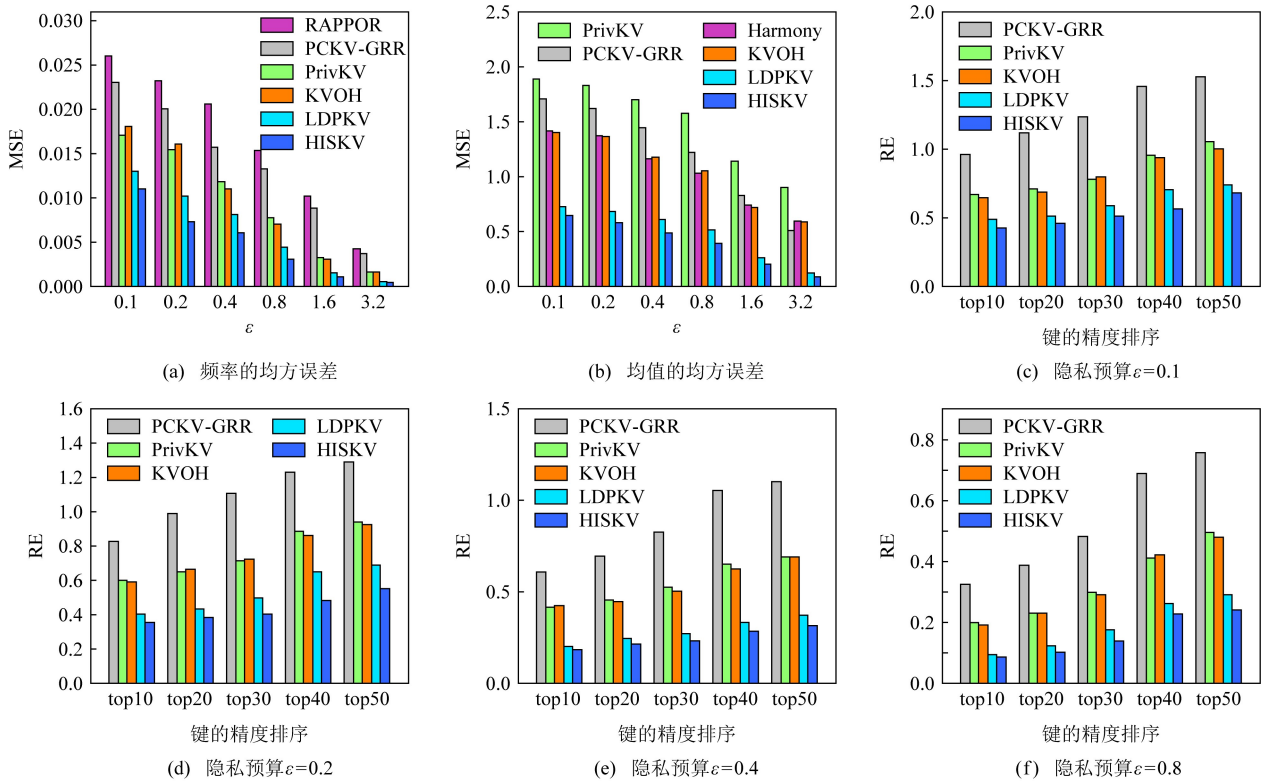


Fig. 5 The key-value frequency and mean estimation of JData

图 5 基于 JData 的键-值频率与均值估计结果

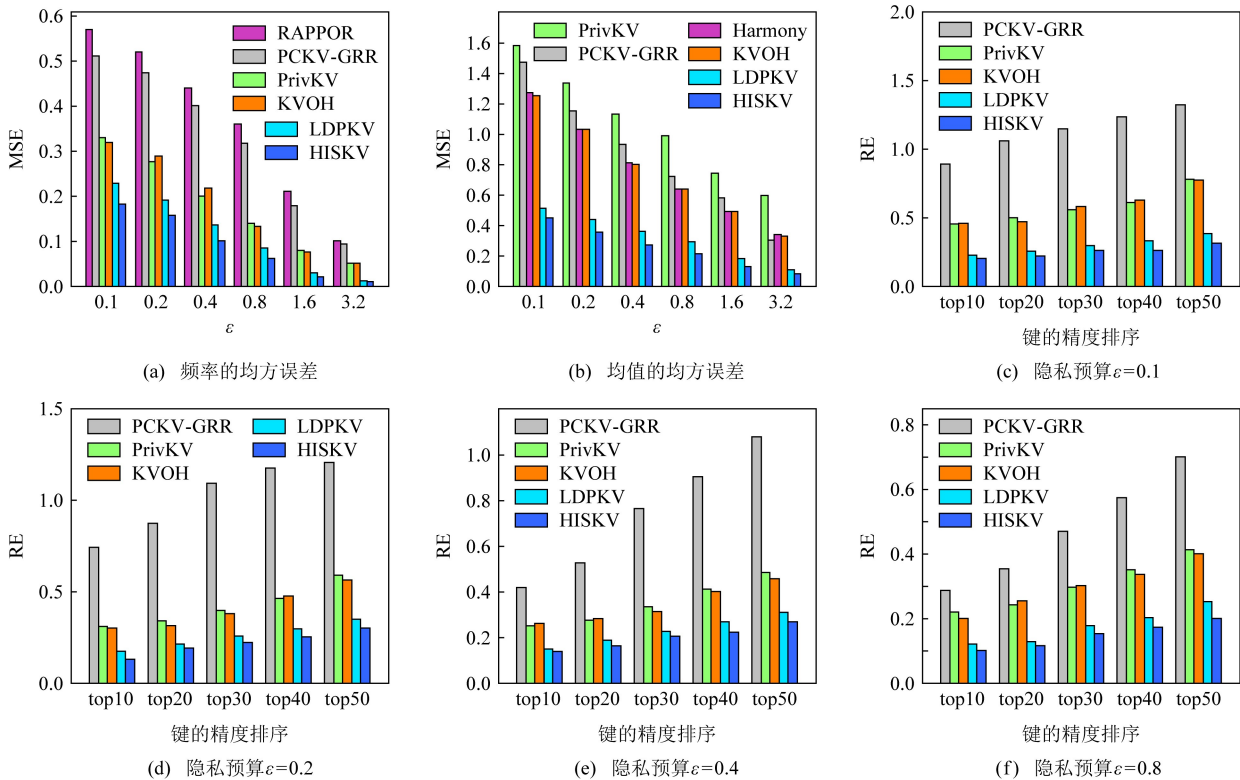


Fig. 6 The key-value frequency and mean estimation of LNR

图 6 基于 LNR 的键-值频率与均值估计结果

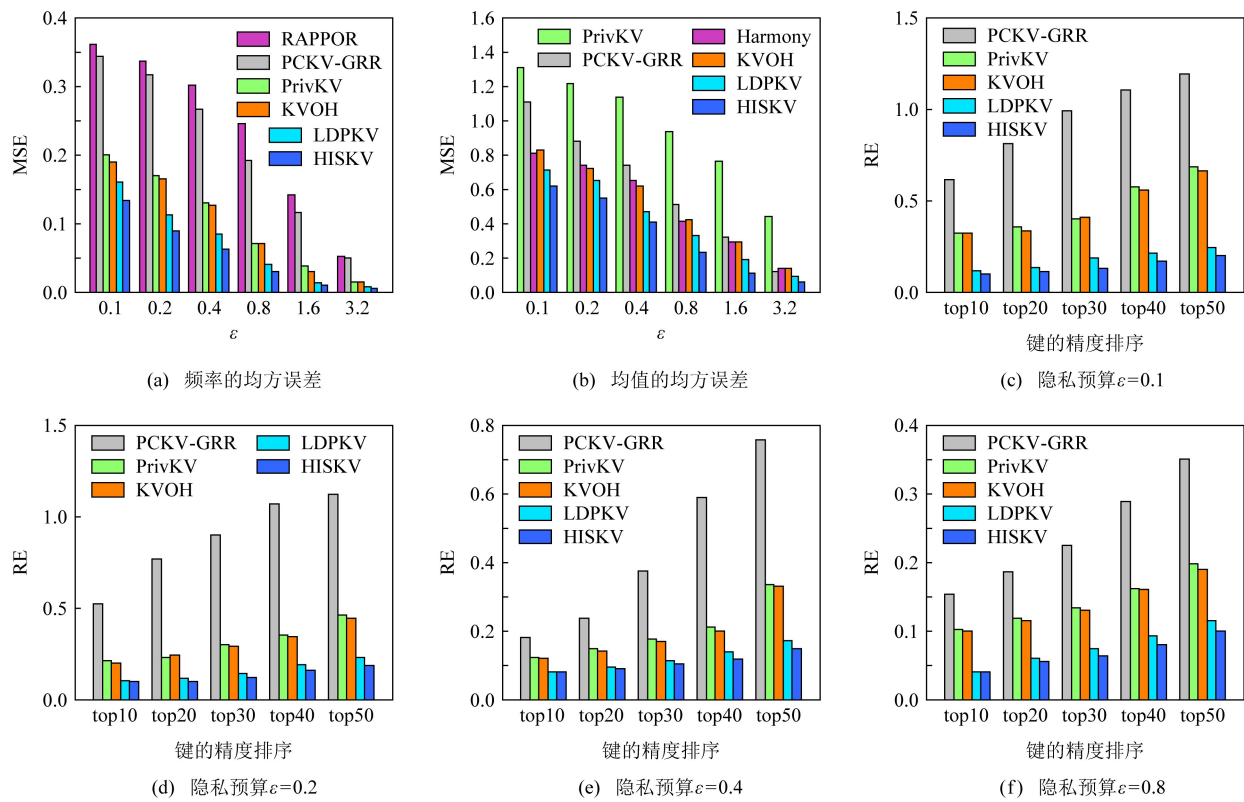


Fig. 7 The key-value frequency and mean estimation of GAUSS

图 7 基于 GAUSS 的键-值频率与均值估计结果

(f)所示.当 $\epsilon=0.2$ 时,键值为 top40 时,HISKV 算法的精度相对于 PCKV-GRR 算法,在 LNR 数据集上达到了近 5 倍,在 GAUSS 数据集上达到了近 6 倍,其主要原因是 HISKV 算法利用直方图将键-值对看成一个整体进行扰动,不但减少了隐私预算分割和扰动次数,同时其增大了扰动成真实数据的概率.

5 结束语

针对本地差分隐私保护下收集键-值数据存在的问题,本文结合现有的该类数据收集算法存在的不足,提出了基于直方图技术的收集算法 HISKV,引入基于用户分组策略的最优截断长度估计算法,结合真实键的实际扰动情况设置合理的扰动概率,并利用填充-抽样技术减少用户与收集者之间的通信代价.从本地差分隐私定义角度分析了 HISKV 满足 ϵ -本地差分隐私.理论分析了 HISKV 的无偏性、所产生的方差以及最大偏差,最后通过 4 种数据集验证了 HISKV 算法的可用性.实验结果表明,HISKV 明显优于现有的同类算法.未来工作考虑动态环境下的隐私键-值数据的收集与分析问题.

参 考 文 献

[1] Ye Qingqing, Hu Haibo, Meng Xiaofeng, et al. PrivKV: Key-value data collection with local differential privacy [C] // Proc of the 28th USENIX Security Symp (S&P 2019). New York; ACM, 2019: 317-331

[2] Sun Lin, Zhao Jun, Ye Xiaojun, et al. Conditional analysis for key-value data with local differential privacy [J]. arXiv preprint, arXiv:1907.05014, 2019

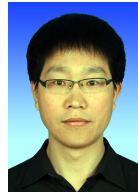
[3] Gu Xiaolan, Li Ming, Cheng Yueqiang, et al. PCKV: Locally differentially private correlated key-value data collection with optimized utility [J]. arXiv preprint, arXiv: 1911.12834, 2019

[4] Zhang Xiaojian, Fu Nan, Meng Xiaofeng. Key-value data collection under local differential privacy [J]. Chinese Journal of Computers, 2020, 43(8): 1479-1492 (in Chinese) (张啸剑,付楠,孟小峰.基于本地化差分隐私的键-值数据精确收集算法[J].计算机学报,2020,43(8):1479-1492)

[5] Wang Tianhao, Li Ninghui, Jha S. Locally differentially private frequent itemset mining [C] //Proc of IEEE Symp on Security and Privacy (SP2018). Piscataway, NJ: IEEE, 2018: 127-143

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

- [7] Wang Tianhao, Bloi J. Locally differentially private protocols for frequency estimation [C] //Proc of the 26th USENIX Security Symp (SP 2017). Berkeley, CA: USENIX Association, 2017: 729-745
- [8] Erlingsson Ú, Pihur V, Korolova A. RAPPOR: Randomized aggregatable privacy-preserving ordinal response [C] //Proc of the 2014 ACM SIGSAC Conf on Computer and Communications Security (CCS 2014). New York: ACM, 2014: 1054-1067
- [9] Kairouz P, Bonawitz K, Ramage D. Discrete distribution estimation under local privacy [C] //Proc of the 33rd Int Conf on Machine Learning (ICML 2016). New York: ACM, 2016: 2436-2444
- [10] Bassily R, Smith A. Local, private, efficient protocols for succinct histograms [C] //Proc of the 47th Annual ACM on Symp on Theory of Computing (STOC 2015). New York: ACM, 2015: 127-135
- [11] Qin Zhan, Yu Ting, Yang Yin, et al. Heavy hitter estimation over set-valued data with local differential privacy [C] //Proc of ACM Conf on Computer and Communications Security (CCS 2016). New York: ACM, 2016: 192-203
- [12] Duchi J C, Jordan M I. Local privacy and statistical minimax rates [C] //Proc of the 54th Annual IEEE Symp on Foundations of Computer Science (FOCS 2013). Piscataway, NJ: IEEE, 2013: 429-438
- [13] Duchi J C, Jordan M I, Wainwright M J. Minimax optimal procedures for locally private estimation [J]. Journal of the American Statistical Association, 2018, 113(521):182-201
- [14] Nguyễn T, Xiao Xiaokui, Yang Yin, et al. Collecting and analyzing data from smart device users with local differential privacy [J]. arXiv preprint, arXiv: 1606.0505, 2016
- [15] Wang Ning, Xiao Xiaokui, Yang Yin, et al. Collecting and analyzing multidimensional data with local differential privacy [C] //Proc of the 35th IEEE Int Conf on Data Engineering (ICDE 2019). Los Alamitos, CA: IEEE Computer Society, 2019: 638-649



Zhang Xiaojian, born in 1980. PhD, associate professor. His main research interests include differential privacy, data mining, and graph data management.

张啸剑, 1980年生, 博士, 副教授, 主要研究方向为差分隐私、数据挖掘、图形数据管理。



Xu Yaxin, born in 1996. Master candidate. Her main research interest is local differential privacy. (xyx19@huel.edu.cn)

徐雅鑫, 1996年生, 硕士研究生, 主要研究方向为本地差分隐私。



Fu Nan, born in 1988. PhD candidate. His main research interests include local differential privacy, data mining. (funan1988@qq.com)

付楠, 1988年生, 博士研究生, 主要研究方向为本地差分隐私、数据挖掘。



Meng Xiaofeng, born in 1964. Professor and PhD supervisor at the Renmin University of China. Executive director of CCF. His main research interests include cloud data management, Web data management, native XML databases, flash-based databases, privacy-preserving, and etc. (xfmeng@ruc.edu.cn)

孟小峰, 1964年生, 教授, 博士生导师, 中国计算机学会常务理事, 主要研究方向为云数据管理、Web数据集成、XML数据库系统、基于flash的数据库、隐私保护等。