基于本地差分隐私的键-值数据精确收集方法

张啸剑[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 450046)

[2)](中国人民大学信息学院 北京 100872)

摘 要 基于本地差分隐私的键-值数据的收集与分析得到了研究者的广泛关注.键与值的值域大小、二者之间的关联性、报告给收集者的通信方式以及本地扰动机制直接制约着频率与均值估计的精度.针对现有键-值数据本地扰动方法存在的不足,该文提出了一种精确且有效的本地扰动方法 LDPKV(Locally Differentially Private Key-Value data collection),该方法结合输入值域与输出值域之间的整体映射关系,在较少通信代价与不分割隐私预算的情况下,对键-值对进行统一处理.其主要思想是:首先对每个用户所拥有的键-值对进行统一离散化处理;结合每个用户的离散化结果,利用伯努利采样技术随机地抽样一条键-值对进行本地随机扰动;然后将扰动后的键-值对报告给收集者.收集者利用每个用户的报告值估计每个键的频率以及所对应值的均值.理论分析了 LDPKV 方法产生的方差与最大偏差,以及与现有键-值数据收集方法在真实与合成数据集上进行综合比较.实验结果表明 LDPKV 方法均优于同类方法.

关键词 键-值数据;隐私保护;本地差分隐私;频率估计;均值估计

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2020.01479

Key-Value Data Accurate Collection under Local Differential Privacy

ZHANG Xiao-Jian[1)] FU Nan[1)] MENG Xiao-Feng[2)]

[1)](*School of Computer & Information Engineering, Henan University of Economics and Law, Zhengzhou* 450046)

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

Abstract Privacy-preserving data collection is an important problem that has attracted much research attention. The state-of-the-art solution for this problem is local differential privacy, which provides a strong protection by adding random noise from users. Existing methods with local differential privacy, however, rarely focus on key-value data collection. In this paper, given a set U of users, we study locally differentially private algorithms for collecting key-value data over U to estimate the frequency of each key and mean of the value. Existing methods for key-value data collection mostly adopt direct encoding mechanisms and local perturbation mechanisms, which locally encode and perturb the value of each user. The two mechanisms, however, require that we have to transform the whole domain of key-value data and split the privacy budget. The transformation of the whole domain makes the communication cost too high. Furthermore, the partition of the budget leads to excessive noise and lower utility. To remedy the deficiency of existing methods, this paper proposes an accurate and efficient method with local differential privacy, called LDPKV (Locally Differentially Private Key-Value data collection) to collect and analyze key-value data. LDPKV firstly discretes the key-value data of each user in terms of the

收稿日期:2019-08-29;在线发布日期:2020-02-15.本课题得到国家自然科学基金项目(61502146,91746115,91646203,61772131)、河南省自然科学基金资助项目(162300410006)、河南省高等学校青年骨干教师培养计划项目(2017GGJS084)、河南省教育厅高等学校重点科研项目(16A520002)、河南财经政法大学青年拔尖人才资助计划项目资助.**张啸剑**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为差分隐私、数据库.E-mail:xjzhang82@ruc.edu.cn.**付 楠**,硕士研究生,主要研究方向为差分隐私、数据库.**孟小峰**,博士,教授,博士生导师,中国计算机学会(CCF)会士,主要研究领域为数据融合与知识融合、大数据隐私管理、大数据分析等.

mapping between input and output domain. Based on the discretization, each user in LDPKV randomly samples a key-value pair by using Bernoulli sampling technology, and runs the local randomized response on the sample and reports the index of sample along with the perturbed value to the collector. After that, the collector in LDPKV accumulates all of the reports from users to estimate the frequency of each key and mean of values. Finally, we conduct theoretical analysis on variance and error bound of LDPKV. Based on two metrics: *MSE* (Mean Square Error) and *RE* (Relative Error), three groups of experiments were conducted on real and synthetic datasets (TalkingData, JData, GAUSS, PLAW, and LNR) to evaluate the accuracy of the frequency and the mean generated by from LDPKV, Rappor, KRR, PrivKV, PrivKVM, KVOH, and Harmony. For example, based on JData dataset, we fix $\varepsilon=1.6$ to compare the accuracy of the above seven methods. It can be seen from Figure 2(a) that the *MSE* of LDPKV is 0.002, and 1/6 times that of Rappor and KRR, and 1/2 times that of PrivKV, PrivKVM, and KVOH, respectively. From Figure 2(b), when ε varies from 0.1 to 3.2, LDPKV achieves a small value of *MSE* on mean evaluation against Rappor, Harmony, PrivKV, PrivKVM, and KVOH. Besides, based on three synthetic datasets, LDPKV still achieves better accuracy than Rappor, KRR, PrivKV, PrivKVM, KVOH, and Harmony. For instance, from Figure 3 to Figure 5, when varying ε from 0.1 to 0.8 and varying top$k$ from top10 to top50, we can see that the *RE* of six methods decreases gradually. Especially, on fixing $\varepsilon=0.2$ and top$k$ = top10, the *RE* of LDPKV is 1/4 times, 1/6 times that of Rappor and KRR on LNR and PLAW, respectively. From the results of the above experimental analysis, we can see that LDPKV is better than its competitors.

Keywords key-value data; privacy protection; local differential privacy; frequency estimation; mean estimation

1 引 言

大数据技术的迅猛发展,键-值存储已成为关联数据的主要管理范式,该范式通常被应用于 NoSQL 数据系统中.键-值数据类型普遍存在于移动设备、网购与医疗平台等服务之中,每个用户所拥有的数据均是键-值对的一个集合.因此,基于此类数据的收集与分析能够有效改善应用平台的服务质量,以及向用户提供较为全面的个性化体验.然而,当数据收集者不可信时,键-值对中所包含的个人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的购买与评分记录.通过该表中键的频率以及某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 A 的频率为 2/3,A 所对应值的均值为 0.6.不可信收集者分析出 A 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,用户通常无法掌控自己的隐私数据.本地差分隐私保护技术[1-2]的出现有效地解决了此类矛盾.该技术允许每个用户采用随机应答机制或者拉普拉斯机制对自己所拥有的键-值数据进行本地扰动,再把扰动之后的结果报告给收集者,进而可以有效防止不可信收集者泄露自己的隐私.

表 1 键-值数据样例

用户	商品与评分
u_1	$\langle A,0.7\rangle,\langle B,0.8\rangle,\langle C,0.4\rangle,\langle D,0.55\rangle$
u_2	$\langle A,0.5\rangle,\langle B,0.42\rangle,\langle D,0.35\rangle$
u_3	$\langle C,0.6\rangle,\langle D,0.05\rangle$

目前基于本地差分隐私的键-值数据收集与分析研究工作相应较少.文献[3]结合 Rappor[4]与 Duchi[2]方法提出了 PrivKV (Private Key-Value data collection)与 PrivKVM (Private Key-Value data collection with Multiple times)方法,然而这两种方法却由于频繁分割隐私代价致使最终的频率与均值估计精度较低.文献[5]结合 PrivKVM 方法的不足,基于 GRR[4] (Generalized Random Response)框架提出了 KVOH(Key-Value data collection with One Hot)方法,然而该方法却忽视了键-值整个值域的影响,特别是键的值域比较大时,该方法最终的频率与均值的估计精度较低.结合文献[3,5]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方法时,如何维持键与值之间的关联性.如果键与值经过扰动之后失去关联性,则收集者估计出的键频率和该键所对应值的均值的精度会很低;(2)收集者如何以较小的噪音误差与通信代价收集所有用户的键-值数据.如果用户把扰动之后的键-值根据整个值域的大小报告给收集者,则收集者与用户之间的通信代价以及噪音误差会很高.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-值收集与分析方法能够克服上述方法存在的问题.为此,本文基于本地差分隐私技术提出了一种键-值数据收集方法能够兼顾上述的问题需求.本文主要贡献如下:

(1)为了解决挑战(1),本文提出了 LDPKV 方法.不同于文献[3]对键与值分开处理,LDPKV 方法对键-值对进行统一本地扰动,进而维持了键-值之间的关联性;

(2)为了有效解决挑战(2),LDPKV 方法均首先采用离散化操作对每个键所对应的值进行处理.结合离散化结果,随机采样一个键-值对进行本地扰动,然后将扰动结果报告给收集者;

(3)理论分析了 LDPKV 方法满足 ε-本地差分隐私,以及频率与均值估计的误差边界.通过合成数据与真实数据实验分析,LDPKV 方法具有较好的可用性.

2　相关工作

当前,本地差分隐私主要关注频率[4]、直方图[6]、均值[7-8]、频繁项集[9]估计以及图发布[10]等方面的研究.谷歌 Chrome 浏览器所嵌入的 Rappor[4]方法结合 Wamer 提出的 WRR 方法[13]实现了用户浏览数据的本地保护.然而,基本的 Rappor 方法容易受到值域大小的影响,大的值域会对该方法造成大的估计误差与通信代价.改进的 Rappor 方法[4]利用 Bloom filter 技术缩减值域大小来提高估计精度.同时,SHist[11]方法利用哈希与矩阵投影技术将通信代价减少到 log 级别.此外,OUE 方法[12]与 OLH 方法[12]分别利用一元编码与本地哈希技术来提高估计精度,并且估计误差脱离了值域大小的影响.不同于上述的频率估计方法,Duchi 方法[7]结合 WRR 方法[13]与离散化操作实现了$[-1,1]$区间上的均值估计.然而,由于 Duchi 方法输出结果是固定的两个离散值,进而使最终的估计结果偏离了$[-1,1]$区间.针对 Duchi 方法的不足,PM 方法[8]能够将$[-1,1]$区间中的某个真实值扰动到一个受约束的联系区间$[-C,C]$,并在该连续区间中计算出该扰动值的左右边界.尽管目前有多种频率与均值估计方法,而涉及键-值数据收集与分析的工作却很少.文献[3]结合 GRR 与 Duchi 提出了 PrivKV 与 PrivKVM 方法来收集键-值数据,然而这两种方法却由于频繁分割隐私代价导致较高的估计误差.文献[5]同样结合 GRR 方法提出了 KVOH 方法,然而该方法由于继承了 GRR 方法的不足,无法应对稀疏的键-值数据.因此,针对上述方法的不足,本文提出了一种基于离散化与随机采样技术的键-值收集方法 LDPKV,该方法不但对键-值数据进行整体考虑,还具有较高的频率与均值估计精度.

3　基础知识与问题描述

3.1　本地差分隐私

不同于中心化差分隐私技术,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如下所示.

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 在任意两条不同元组 t 与 $t'(t,t'\in Dom(A))$ 上得到相同输出结果 $o(o\in Range(A))$ 的概率满足下列不等式,则 A 满足 ε-本地差分隐私.

$$\Pr[A(t)=o]\leq \exp(\varepsilon)\times \Pr[A(t')=o] \quad (1)$$

其中,ε 为隐私预算,该值的大小直接决定着隐私保护程度.

随机应答机制[13]是实现本地差分隐私的常用技术.该机制的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 $1-p$ 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 $p=e^{\varepsilon}/(1+e^{\varepsilon})$ 或者其它值,收集者获得所有应答后,即可对真实应答进行分析估计.类似于中心化差分隐私,本地差分隐私同样满足序列组合特性.

性质 1[14].　序列组合性质.给定数据集 D 和 n 个随机算法 $A_1,A_2,\cdots,A_n$,且 $A_i(1\leq i\leq n)$ 满足 ε_i-本地差分隐私,则在 D 上的序列组合满足 ε-本地差分隐私,且 $\varepsilon=\sum_{i=1}^{n}\varepsilon_i$.

3.2　问题描述

给定 n 个用户 $U=\{u_1,u_2,\cdots,u_n\}$,每个用户拥

有不同的键-值对. $K=\{1,2,\cdots,d\}$ 为键的值域，$V=[-1,1]$ 为键所对应值的值域. $S_i=\{(k_j,v_j)\mid 1\leq j\leq l_i, k_j\in K, v_j\in V\}$ 为用户 u_i 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-值数据之后，通常做两种基本的统计分析，即键所对应的频率，以及对应值所对应的均值，具体计算如式(2)与式(3)所示.

$$f_k=\frac{|\{u_i\mid\exists(k,v)\in S_i\}|}{n} \tag{2}$$

其中，$|\{u_i\mid(k,v)\in S_i\}|$ 表示集合 U 中拥有键为 k 的用户个数. 例如表 1 中拥有键为 D 商品的用户个数为 3，则 $f_D=1$.

$$\mu_k=\frac{\sum_i\sum_{j:k_j=k}v_j}{|\{u_i\mid\exists(k,v)\in S_i\}|} \tag{3}$$

其中，$\sum_i\sum_{j:k_j=k}v_j$ 表示键为 k 所有对应值的和. 例如表 1 中键为 D 商品所对应值的和为 0.95，则 $\mu_D=0.32$.

采用均方误差 MSE(Mean Square Estimation)与相对误差 RE(Relative Error)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方法同时，尽可能获得误差较小的估计结果.

4 键-值数据收集方法 LDPKV

4.1 LDPKV 方法

类似于文献[3]，结合键的值域 K(值域大小为 d)与每个用户所拥有的键-值对集合 S_i，首先对 S_i 中所有 (k_j,v_j) 对进行编码操作，生成一个长度为 d 的新型向量 $\boldsymbol{S}_i'$.

$$(k_j,v_j)=\begin{cases}(1,v_j), & k_j\in K, v_j\in[-1,1]\\(0,0), & \text{其他}\end{cases} \tag{4}$$

根据式(4)可知，集合 $\boldsymbol{S}_i'$ 中的元素分别由 $(0,0)$ 与 $(1,v_j)$ 组成，其中 $v_j\in[-1,1]$. 例如，假设键的值域 $K=\{I_1,I_2,I_3,I_4\}$，$S_i=\{(I_2,0.8),(I_3,0.3)\}$. 经由式(4)编码后，$\boldsymbol{S}_i'=\{(0,0),(1,0.8),(1,0.3),(0,0)\}$.

一种最直接的方法是分别采用现有满足本地差分隐私的频率与均值估计方法对 $\boldsymbol{S}_i'$ 中的元素进行频率与均值估计. 然而，这种最直接的方法通常会打破 k_j 与 v_j 之间的关联性. 尽管 PrivKV 与 PrivKVM 方法的可用性优于最直接方法，然而这两种由于分割隐私代价导致最终的可用性较低. 结合集合 $\boldsymbol{S}_i'$ 与 PrivKV 方法可知，扰动方法的输入值域变成了 $\{(0,0)\cup\{(1,v)\mid-1\leq v\leq 1\}\}$，而扰动之后的输出值域为 $\{(0,0),(1,1),(1,-1)\}$. 基于此分析，本文提出了一种频率与其所对应值的均值估计方法 LDPKV，该方法在不分割隐私代价的情况下对键-值对进行统一地扰动. 该方法细节如算法 1 所示.

算法 1. LDPKV 算法.

输入：所有用户的键-值对集合 $S=\{S_1,S_2,\cdots,S_n\}$，隐私预算 ε，扰动概率 p 与 q，键的值域为 K，以及 K 的大小 d

输出：频率向量 $\boldsymbol{f}^*$，均值向量 $\boldsymbol{\mu}^*$

1. $\boldsymbol{f}^*\leftarrow\varnothing,\boldsymbol{\mu}^*\leftarrow\varnothing$;
2. FOR each u_i user DO
3. u_i reports $(j,(k_j^*,v_j^*))\leftarrow\mathrm{LRR}(S_i,\varepsilon)$;
4. END FOR
5. FOR each key $k(k\in K)$ DO
6. Collector aggregates f_k^*;
7. Collector calibrates f_k^* as $f_k^*\leftarrow\dfrac{d\times f_k^*-(1-p)}{p-q}$;
8. Collector counts $(0,0),(1,1)$ and $(1,-1)$ as $c'(0),c'(1)$, and $c'(-1)$;
9. $M\leftarrow c'(0)+c'(1)+c'(-1)$;
10. Collector calibrates $c'(1)$ and $c'(-1)$ as:

$$c^*(1)\leftarrow\frac{c'(1)-M\left(r+\frac{1-p}{2}\right)}{(1-q-r)-\left(r+\frac{1-p}{2}\right)},$$

$$c^*(-1)\leftarrow\frac{c'(-1)-M\left(r+\frac{1-p}{2}\right)}{(1-q-r)-\left(r+\frac{1-p}{2}\right)};$$

11. Collector computes the mean as u_k^*:

$$\mu_k^*\leftarrow\frac{c^*(1)-c^*(-1)}{c^*(1)+c^*(-1)};$$

12. $\boldsymbol{f}^*\leftarrow\boldsymbol{f}^*\cup f_k^*,\boldsymbol{\mu}^*\leftarrow\boldsymbol{\mu}^*\cup u_k^*$;
13. END FOR
14. Return $\boldsymbol{f}^*,\boldsymbol{\mu}^*$.

LDPKV 算法是基于本地差分隐私整体考虑用户键-值对的频率与均值估计问题. 首先每个用户利用 LRR(Local Randomized Response)方法本地扰动自己的键-值对并报告给收集者(步骤 2～步骤 4). 收集者聚集所有用户的报告值后，计算并修正每个键所对应的频率以及值所对应的均值(步骤 5～步骤 13). 根据算法的第 3 行可知，每个用户在不分割隐私预算 ε 的情况下调用 LRR 算法本地统一扰动自己的键-值对. 由于每个用户利用 LRR 算法仅扰动一组键-值对后再报告给收集者，进而使得收集者与用户之间的通信代价为 $\Theta(1)$. 接下来阐述 LRR 算法的实现，其具体实现细节如算法 2 所示.

算法 2.　LRR 算法.

输入：用户 u_i 的键-值集合 S_i，隐私预算 ε

输出：扰动之后的值 $(j,(k_j^*,v_j^*))$

1. $\boldsymbol{S}_i' \leftarrow \text{Encoding}(S_i)$；

2. Sample the j-th pair uniformly at random from $\boldsymbol{S}_i'$；

3. IF the j-th pair (k_j,v_j) is $(0,0)$ THEN

4.　Perturbs $(0,0)$ as：

$$(k_j^*,v_j^*)=\begin{cases}(0,0), & w.p.\ p\\ (1,1), & w.p.\ (1-p)/2\\ (1,-1), & w.p.\ (1-p)/2\end{cases};$$

5. ELSE IF the j-th pair (k_j,v_j) is $(1,v)$ THEN

6.　Perturbs $(1,v)$ as：

$$(k_j^*,v_j^*)=\begin{cases}(0,0), & w.p.\ q\\ (1,1), & w.p.\ r+\dfrac{(1+v)}{2}(1-q-2r)\\ (1,-1), & w.p.\ r+\dfrac{(1-v)}{2}(1-q-2r)\end{cases};$$

7. END IF

8. Return $j,(k_j^*,v_j^*)$.

结合 LRR 算法，每个用户首先利用式(4)对自己的键-值对 S_i 进行编码转换(步骤 1)，进而获得 $\boldsymbol{S}_i'$. 为了减少收集者与用户之间的通信代价以及噪音误差，在 $\boldsymbol{S}_i'$ 中均匀随机地抽取一个键-值对 (k_j,v_j) 进行本地扰动(步骤 2～步骤 7). 无论 (k_j,v_j) 是 $(0,0)$ 还是 $(1,v)$，最终结果均以不同的概率扰动成 $(0,0)$、$(1,1)$ 以及 $(1,-1)$(步骤 4 与步骤 6).

要实现 LRR 算法中的步骤 4 与步骤 6，需要计算出概率 p,q,r. 根据 LRR 算法的描述可知输入值的值域为 $\{(0,0)\cup\{(1,v)\mid -1\le v\le 1\}\}$，输出值域为 $\{(0,0)$、$(1,1)$ 与 $(1,-1)\}$. 因此，可以把输入值域分成两种情况：

(1) 当输入值为 $(0,0)$ 与 $(1,v)$ 时，为使 $(0,0)$、$(1,1)$ 与 $(1,-1)$ 三种输出结果满足 ε-本地差分隐私，则下列三种不等式需成立：

输出值为 $(0,0)$ 时，则式(5)成立：

$$p\le e^{\varepsilon}\times q \tag{5}$$

输出值为 $(1,1)$ 时，则式(6)成立：

$$r+\frac{(1+v)}{2}(1-q-2r)\le e^{\varepsilon}\times\frac{1-p}{2} \tag{6}$$

输出值为 $(1,-1)$ 时，则式(7)成立：

$$r+\frac{(1-v)}{2}(1-q-2r)\le e^{\varepsilon}\times\frac{1-p}{2} \tag{7}$$

(2) 当输入值为 $(1,v_1)$ 与 $(1,v_2)$ 时，其中 v_1 与 v_2 互为近邻，为使 $(0,0)$、$(1,1)$ 与 $(1,-1)$ 三种输出结果满足 ε-本地差分隐私，同样下列三种不等式需同时成立：

输出值为 $(0,0)$ 时，则式(8)成立：

$$q\le e^{\varepsilon}\times q \tag{8}$$

输出值为 $(1,1)$ 时，则式(9)成立：

$$\frac{r+\dfrac{(1+v_1)}{2}(1-q-2r)}{r+\dfrac{(1+v_2)}{2}(1-q-2r)}\le e^{\varepsilon} \tag{9}$$

输出值为 $(1,-1)$ 时，则式(10)成立：

$$\frac{r+\dfrac{(1-v_1)}{2}(1-q-2r)}{r+\dfrac{(1-v_2)}{2}(1-q-2r)}\le e^{\varepsilon} \tag{10}$$

从式(5)至式(10)可知，以上六种不等式应同时满足. 根据本地差分隐私定义可知，若要使式(6)与式(7)同时成立，则需要式(6)中的 $v=1$，式(7)中的 $v=-1$. 进而可得式(11).

$$1-q-r\le e^{\varepsilon}\times\frac{1-p}{2} \tag{11}$$

通过观察式(9)与式(10)可知，若要使这两种等式同时成立，v_1 与 v_2 只需要考虑取极值情况即可. (1) 当 $1-q-2r>0$，$v_1=1$，$v_2=-1$ 时，式(9)取得最大值，则由式(9)可知 $(1-q-r)/r\le e^{\varepsilon}$；(2) 当 $1-q-2r>0$，$v_1=-1$，$v_2=1$ 时，式(10)取得最大值，则由式(10)可知 $(1-q-r)/r\le e^{\varepsilon}$. 因此，结合上述两种情况可知式(9)与式(10)可以同时归结为 $(1-q-r)/r\le e^{\varepsilon}$.

根据式(5)和式(11)以及 $(1-q-r)/r\le e^{\varepsilon}$，可以计算出概率 $p=e^{\varepsilon}/(e^{\varepsilon}+2)$、$q=1/(e^{\varepsilon}+2)$ 以及 $r=1/(e^{\varepsilon}+2)$. 获得 p,q,r 之后，LRR 算法以不同的概率将 (k_j,v_j) 扰动成 $(0,0)$、$(1,1)$ 以及 $(1,-1)$.

4.2　LDPKV 方法的隐私性与可用性分析

本小节主要从 ε-本地差分隐私阐述 LDPKV 方法的隐私性，以及从估计无偏性与方差阐述其可用性. LDPKV 方法中只有 LRR 算法花销 ε，因此，只要 LRR 算法满足 ε-本地差分隐私，则根据性质 1 可知 LDPKV 方法同样满足 ε-本地差分隐私.

定理 1.　LRR 算法满足 ε-本地差分隐私.

证明.　设 (k_1,v_1) 与 (k_2,v_2) 表示两个不同的输入键-值对，(k^*,v^*) 表示 (k_1,v_1) 与 (k_2,v_2) 经过 LRR 算法扰动之后的结果，则 $(k^*,v^*)\in\{(0,0),(1,1),(1,-1)\}$. 已知 $p=e^{\varepsilon}/(e^{\varepsilon}+2)$、$q=1/(e^{\varepsilon}+2)$、$r=1/(e^{\varepsilon}+2)$. 可分三种情况证明 LRR 算法满足 ε-本地差分隐私.

(1) $(k^*,v^*)=(0,0)$

$$\frac{\Pr[(k^*,v^*)=(0,0)|(k_1,v_1)]}{\Pr[(k^*,v^*)=(0,0)|(k_2,v_2)]}\le\frac{\Pr[(k^*,v^*)=(0,0)|(k_1,v_1)=(0,0)]}{\Pr[(k^*,v^*)=(0,0)|(k_2,v_2)=(1,v)]}=\frac{p}{q}=\frac{e^\varepsilon/(e^\varepsilon+2)}{1/(e^\varepsilon+2)}=e^\varepsilon.$$

(2) $(k^*,v^*)=(1,1)$

由定义 1 以及式(5)至式(11)分析可知：当$(k_1,v_1)=(1,1)$时，概率 $\Pr[(k^*,v^*)=(1,1)|(k_1,v_1)]$达到最大；当$(k_2,v_2)=(0,0)$或者$(k_2,v_2)=(1,-1)$时，概率 $\Pr[(k^*,v^*)=(1,1)|(k_2,v_2)]$达到最小，则：

$$\frac{\Pr[(k^*,v^*)=(1,1)|(k_1,v_1)]}{\Pr[(k^*,v^*)=(1,1)|(k_2,v_2)]}\le\frac{\Pr[(k^*,v^*)=(1,1)|(k_1,v_1)=(1,1)]}{\Pr[(k^*,v^*)=(1,1)|(k_2,v_2)=(0,0)]}=\frac{1-q-r}{(1-p)/2}=\frac{e^\varepsilon/(e^\varepsilon+2)}{1/(e^\varepsilon+2)}=e^\varepsilon$$

或者

$$\frac{\Pr[(k^*,v^*)=(1,1)|(k_1,v_1)]}{\Pr[(k^*,v^*)=(1,1)|(k_2,v_2)]}\le\frac{\Pr[(k^*,v^*)=(1,1)|(k_1,v_1)=(1,1)]}{\Pr[(k^*,v^*)=(1,1)|(k_2,v_2)=(1,-1)]}=\frac{1-q-r}{r}=\frac{e^\varepsilon/(e^\varepsilon+2)}{1/(e^\varepsilon+2)}=e^\varepsilon.$$

(3) $(k^*,v^*)=(1,-1)$

由定义 1 以及式(5)至式(11)分析可知：当$(k_1,v_1)=(1,-1)$时，概率 $\Pr[(k^*,v^*)=(1,-1)|(k_1,v_1)]$达到最大；当$(k_2,v_2)=(0,0)$或者$(k_2,v_2)=(1,1)$时，概率 $\Pr[(k^*,v^*)=(1,-1)|(k_2,v_2)]$达到最小，则：

$$\frac{\Pr[(k^*,v^*)=(1,-1)|(k_1,v_1)]}{\Pr[(k^*,v^*)=(1,-1)|(k_2,v_2)]}\le\frac{\Pr[(k^*,v^*)=(1,-1)|(k_1,v_1)=(1,-1)]}{\Pr[(k^*,v^*)=(1,-1)|(k_2,v_2)=(0,0)]}=\frac{1-q-r}{(1-p)/2}=\frac{e^\varepsilon/(e^\varepsilon+2)}{1/(e^\varepsilon+2)}=e^\varepsilon$$

或者

$$\frac{\Pr[(k^*,v^*)=(1,-1)|(k_1,v_1)]}{\Pr[(k^*,v^*)=(1,-1)|(k_2,v_2)]}\le\frac{\Pr[(k^*,v^*)=(1,-1)|(k_1,v_1)=(1,-1)]}{\Pr[(k^*,v^*)=(1,-1)|(k_2,v_2)=(1,1)]}=\frac{1-q-r}{r}=\frac{e^\varepsilon/(e^\varepsilon+2)}{1/(e^\varepsilon+2)}=e^\varepsilon.$$

根据上述三种情况分析可知 LRR 算法满足 ε-本地差分隐私.

尽管通过 LDPKV 算法可以估算出每个键的频率 f_k^* 以及所对应值的均值 μ_k^*，但是由于 LRR 算法的本地扰动，f_k^* 与 μ_k^* 会偏离于原始值，而我们期望 f_k^* 与 μ_k^* 均要满足无偏性. 因此，需要计算出 f_k^* 与 μ_k^* 的修正因子(即算法 1 的步骤 7 与步骤 9).

定理 2. 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 $p=e^\varepsilon/(e^\varepsilon+2)$、$q=1/(e^\varepsilon+2)$时，无偏估计 $E[f_k^*]=f_k$，$E[\mu_k^*]=\mu_k$ 成立.

证明. 首先证明 $E[f_k^*]=f_k$. 设 $c^*(k)$ 与 $c(k)$ 分别表示键 k 的估计计数以及真实计数. 根据式(2)可知，$f_k^*=c^*(k)/n$，$f_k=c(k)/n$. 因此，只要 $E[c^*(k)]=c(k)$ 成立，则 $E[f_k^*]=f_k$ 成立. 设 $I(k)$ 表示收集者收集到的键 k 为 1 的个数. 设 d 为键的值域大小. 由于每个用户采用伯努利抽样随机抽取一个键-值对进行扰动，则每个键-值对被抽取的概率为 $1/d$. 根据文献[13]可知：

给定 n 个用户，响应键 k 为 0 的概率与响应 k 为 1 的概率分别为 $\Pr[k=0]=\frac{c(k)}{n}q+\left(1-\frac{c(k)}{n}\right)p$、$\Pr[k=1]=\frac{c(k)}{n}(1-q)+\left(1-\frac{c(k)}{n}\right)(1-p)$. 由于无法获知真实的 $c(k)$，需要求得对应的估计量 $c^*(k)$. 根据两种响应概率可知 n 个用户的响应结果构成了符合二项分布的 n 个 0/1 序列. 结合该二项分布，构造相应的似然函数为

$$L(c(k))=\left[\frac{c(k)}{n}(1-q)+\left(1-\frac{c(k)}{n}\right)(1-p)\right]^{I_k}\times\left[\frac{c(k)}{n}q+\left(1-\frac{c(k)}{n}\right)p\right]^{\frac{n}{d}-I_k},$$

则对 $L(c(k))$两侧取对数，再对 $c(k)$求导，即可获得 $c(k)$的估计量 $c^*(k)=\frac{d\times I(k)-n\times(1-p)}{p-q}$.

为了证明 $E[c^*(k)]=c(k)$，下列过程成立即可.

$$E[c^*(k)]=E\left[\frac{d\times I(k)-n\times(1-p)}{p-q}\right]=\frac{d\times E[I(k)]-n\times(1-p)}{p-q}=\frac{d\times\left(\frac{c(k)}{d}\times(1-q)+\left(\frac{n-c(k)}{d}\right)\times(1-p)\right)-n\times(1-p)}{p-q}=\frac{c(k)\times(1-q)+(n-c(k))\times(1-p)-n\times(1-p)}{p-q}=\frac{c(k)\times(p-q)}{p-q}=c(k),$$

则 $E[f_k^*]=f_k$.

接下来证明 $E[\mu_k^*]=\mu_k$ 成立. 设 v_j 与 v_j^* 分别表示键 k 所对应某个值的真实值与扰动值. 根据式(3)可知，只要 $E[v_j^*]=v_j$ 成立，则 $E[\mu_k^*]=\mu_k$ 成立. 根据算法 1 可知，v_j 的输出结果 v_j^* 的值域为 $\{-1,0,1\}$. 如果知道这三种值的输出概率，即可计算出输出 v_j^* 的期望. 设 $\Pr[v_j^*=-1]$，$\Pr[v_j^*=0]$ 与 $\Pr[v_j^*=1]$ 分别表示 -1，0，1 的输出概率. 结合 $p=e^\varepsilon/(e^\varepsilon+2)$、$q=1/(e^\varepsilon+2)$、$r=1/(e^\varepsilon+2)$，则下列等式成立.

$$\Pr[v_j^*=0]=\frac{p+q}{4}+\frac{q}{2}=\frac{p+3q}{4}=\frac{e^\varepsilon+3}{4(e^\varepsilon+2)},$$

$$\begin{aligned}\Pr[v_j^*=1]&=\frac{1}{4}\left(\frac{1-p}{2}+\frac{1-q}{2}\right)+\\&\quad\frac{1}{2}\left(r+\frac{(1+v_j)}{2}(1-q-2r)\right)\\&=\frac{3+e^\varepsilon}{8(2+e^\varepsilon)}+\frac{v_j(e^\varepsilon-1)}{4(2+e^\varepsilon)},\end{aligned}$$

$$\begin{aligned}\Pr[v_j^*=-1]&=\frac{1}{4}\left(\frac{1-p}{2}+\frac{1-q}{2}\right)+\\&\quad\frac{1}{2}\left(r+\frac{(1-v_j)}{2}(1-q-2r)\right)\\&=\frac{3+e^\varepsilon}{8(2+e^\varepsilon)}-\frac{v_j(e^\varepsilon-1)}{4(2+e^\varepsilon)}.\end{aligned}$$

则 $E[v_j^*]$ 可表示成如下公式：

$$\begin{aligned}E[v_j^*]&=0\times\frac{e^\varepsilon+3}{4(e^\varepsilon+2)}+1\times\left(\frac{3+e^\varepsilon}{8(2+e^\varepsilon)}+\frac{v_j(e^\varepsilon-1)}{4(2+e^\varepsilon)}\right)+\\&\quad(-1)\times\left(\frac{3+e^\varepsilon}{8(2+e^\varepsilon)}-\frac{v_j(e^\varepsilon-1)}{4(2+e^\varepsilon)}\right)\\&=\frac{v_j(e^\varepsilon-1)}{2(2+e^\varepsilon)}.\end{aligned}$$

若要使得等式 $E[v_j^*]=v_j$ 成立，则 $E[v_j^*]$ 应当乘以 $\frac{2(2+e^\varepsilon)}{(e^\varepsilon-1)}$，即 $E[v_j^*]=\frac{v_j(e^\varepsilon-1)}{2(2+e^\varepsilon)}\times\frac{2(2+e^\varepsilon)}{(e^\varepsilon-1)}=v_j$，其中 $\frac{2(2+e^\varepsilon)}{(e^\varepsilon-1)}$ 为修正因子. 根据 $E[v_j^*]=v_j$ 可推理出 $E[\mu_k^*]=\mu_k$ 成立.

根据定理 2 可知，修正因子可以使 $E[f_k^*]=f_k$ 与 $E[\mu_k^*]=\mu_k$ 达到无偏. 然而由于利用 LRR 方法对用户的键-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. 而这两种误差的大小是衡量 LRR 方法可用性的主要标准之一. 本文采用方差 $Var(f_k^*)$ 与 $Var(\mu_k^*)$ 度量 LRR 方法产生的误差.

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. 当 $p=e^\varepsilon/(e^\varepsilon+2)$、$q=1/(e^\varepsilon+2)$ 时，f_k^* 的方差为 $Var[f_k^*]=\frac{d}{n}\times\frac{2e^\varepsilon}{(e^\varepsilon-1)^2}-\frac{d}{n}\times\frac{f_k}{e^\varepsilon-1}$. 当 $f_k=0$ 时，f_k^* 的最坏方差为 $Var[f_k^*]=\frac{d}{n}\times\frac{2e^\varepsilon}{(e^\varepsilon-1)^2}$.

证明.　由定理 2 可知 $f_k^*=\dfrac{\frac{d}{n}\times I(k)-(1-p)}{p-q}$，则

$$\begin{aligned}Var[f_k^*]&=Var\left[\frac{\frac{d}{n}I(k)-(1-p)}{p-q}\right]=\frac{d^2Var[I(k)]}{n^2(p-q)^2}\\&=\frac{d^2\left(\frac{n}{d}f_k(1-q)q+\frac{(n-nf_k)}{d}p(1-p)\right)}{n^2(p-q)^2}\\&=\frac{d}{n}\left(\frac{2e^\varepsilon-(e^\varepsilon-1)f_k}{(e^\varepsilon-1)^2}\right)=\frac{d}{n}\left(\frac{2e^\varepsilon}{(e^\varepsilon-1)^2}-\frac{f_k}{e^\varepsilon-1}\right).\end{aligned}$$

当 $f_k=0$ 时，$Var[f_k^*]=\frac{d}{n}\times\frac{2e^\varepsilon}{(e^\varepsilon-1)^2}$.

定理 4.　假设 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 $p=e^\varepsilon/(e^\varepsilon+2)$、$q=1/(e^\varepsilon+2)$、$r=1/(e^\varepsilon+2)$ 时，则

$$Var[\mu_k^*]=\Phi\times\left[\frac{((c(1)-c(-1))(e^\varepsilon-2))^2}{((c(1)+c(-1))(e^\varepsilon-2)-4M)^4}\right]+\Phi\times\left[\frac{1}{((c(1)+c(-1))(e^\varepsilon-2)-4M)^2}\right],$$

其中，$\Phi=4Me^\varepsilon$，$c(1)$ 与 $c(-1)$ 表示(1,1)和(1,-1)的真实计数.

证明.　为了证明方便，设置 $p_1=1-q-r$，$q_1=\frac{1-p}{2}+r$. 根据算法 1 的第 11 行可知 μ_k^* 可表示为 $\mu_k^*=\left(\frac{c^*(1)-c^*(-1)}{c^*(1)+c^*(-1)}\right)$，则 μ_k^* 的方差可表示为

$$\begin{aligned}Var[\mu_k^*]&=Var\left[\left(\frac{c^*(1)-c^*(-1)}{c^*(1)+c^*(-1)}\right)\right]\\&=Var\left[\frac{c'(1)-c'(-1)}{c'(1)+c'(-1)-2Mq_1}\right].\end{aligned}$$

设 $R=c'(1)-c'(-1)$，$S=c'(1)+c'(-1)-2Mq_1$. 则

$$\begin{aligned}Var\left[\frac{c'(1)-c'(-1)}{c'(1)+c'(-1)-2Mq_1}\right]&=Var\left[\frac{R}{S}\right]\\&=\frac{(E[R])^2}{(E[S])^2}\left[\frac{Var[R]}{(E[R])^2}-2\frac{Cov[R,S]}{E[R]E[S]}+\frac{Var[S]}{(E[S])^2}\right]\\&=\frac{Var[R]}{(E[S])^2}-2\frac{E[R]Cov[R,S]}{(E[S])^3}+\frac{(E[R])^2Var[S]}{(E[S])^4}\end{aligned}\tag{12}$$

R 的期望为

$$E[R]=E[c'(1)-c'(-1)]$$
$$=E[c'(1)]-E[c'(-1)]$$
$$=(c(1)-c(-1))(p_1-q_1).$$

同理可知,S 的期望为

$$E[S]=E[c'(1)+c'(-1)-2Mq_1]$$
$$=E[c'(1)]+E[c'(-1)]-2Mq_1$$
$$=(c(1)+c(-1))(p_1-q_1)-2Mq_1.$$

R 的方差为

$$Var[R]=Var[c'(1)-c'(-1)]$$
$$=Var[c'(1)]+Var[c'(-1)]$$
$$=(c(1)+c(-1))(p_1-q_1)(1-p_1-q_1)+2Mq_1(1-q_1).$$

同理可知,S 的方差为

$$Var[S]=Var[c'(1)]+Var[c'(-1)]$$
$$=(c(1)+c(-1))(p_1-q_1)(1-p_1-q_1)+2Mq_1(1-q_1).$$

此外,

$$Cov[R,S]=Cov[c'(1)-c'(-1),c'(1)+c'(-1)]$$
$$=Var[c'(1)]-Var[c'(-1)]$$
$$=(c(1)-c(-1))(p_1-q_1)(1-p_1-q_1).$$

将 $E[R]$、$E[S]$、$Var[R]$、$Var[S]$、$Cov[R,S]$ 代入式(12)可知:

$$Var[\mu_k^*]=\Phi\times\left[\frac{((c(1)-c(-1))(e^\varepsilon-2))^2}{((c(1)+c(-1))(e^\varepsilon-2)-4M)^4}\right]+\Phi\times\left[\frac{1}{((c(1)+c(-1))(e^\varepsilon-2)-4M)^2}\right],$$

其中 $\Phi=4Me^\varepsilon$.

尽管通过定理 3 与定理 4 可以计算出某个键 k 的频率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 之间的最大偏差是具有挑战性的问题.接下来由定理 5 与定理 6 给出详细说明.

定理 5. 假设 f_k^* 与 f_k 分别为键 k 的估计频率与真实频率.当 $p=e^\varepsilon/(e^\varepsilon+2)$、$q=1/(e^\varepsilon+2)$ 时,则 $|f_k-f_k^*|=O(d\sqrt{\log(d/\beta)}/\varepsilon\sqrt{n})$ 至少以概率 $1-\beta$ 成立,其中 n 为用户个数,d 为键 k 的值域大小.

证明. 根据定理 2 可知下列等式成立.

$$|f_k-f_k^*|=\frac{|c(k)-c^*(k)|}{n}$$
$$=\frac{1}{n}\times\left|\sum_{j=1}^{n}\frac{|d|\times k_j-(1-p)}{p-q}-\sum_{j=1}^{n}\frac{|d|\times k_j^*-(1-p)}{p-q}\right|$$
$$=\frac{|d|}{n}\times\sum_{j=1}^{n}\left|\frac{k_j-k_j^*}{p-q}\right|,$$

其中 $c^*(k)$ 与 $c(k)$ 表示键 k 的估计计数和真实计数;k_j 与 k_j^* 表示第 j 个用户拥有键 k 的真实值与估计值.

因此,可知 $k_j-k_j^*$ 为一个随机变量,根据定理 2 可知其均值为 0.随机变量 $k_j-k_j^*$ 的方差可以表示为

$$Var[k_j-k_j^*]=E[(k_j-k_j^*)^2]-(E(k_j-k_j^*))^2$$
$$=\frac{3}{2(e^\varepsilon+2)}-\frac{1}{4(e^\varepsilon+2)^2}$$
$$=\frac{6e^\varepsilon+11}{4(e^\varepsilon+2)^2}=O(1/\varepsilon^2).$$

由于 $k_j\in\{0,1\}$,$k_j^*\in\{0,1\}$,则 $k_j-k_j^*\in\{-1,0,1\}$.进而可知 $|k_j-k_j^*|\leq 1$,$\left|\frac{k_j-k_j^*}{p-q}\right|\leq\frac{e^\varepsilon+2}{e^\varepsilon-1}$.

根据 Bernstein 不等式可知:

$$\Pr[|f_k-f_k^*|\geq\lambda/n]=\Pr[|c(k)-c^*(k)|\geq\lambda]$$
$$=\Pr\left[\frac{d}{n}\sum_{j=1}^{n}\left|\frac{k_j-k_j^*}{p-q}\right|\geq\lambda\right]=\Pr\left[\sum_{j=1}^{n}\left|\frac{k_j-k_j^*}{p-q}\right|\geq\frac{n\lambda}{d}\right]$$
$$\leq 2\times\exp\left(-\frac{\frac{n^2\lambda^2}{2d^2}}{\sum_{j=1}^{n}Var\left[\frac{k_j-k_j^*}{p-q}\right]+\frac{n\lambda}{3d}\times\frac{e^\varepsilon+2}{e^\varepsilon-1}}\right)$$
$$=2\times\exp\left(-\frac{\frac{n}{2}\lambda^2}{d^2O(1/\varepsilon^2)+d\lambda O(1/\varepsilon)}\right).$$

结合上述不等式可知 $\lambda=O\left(\frac{d\sqrt{\log(d/\beta)}}{\varepsilon\sqrt{n}}\right)$,则可知 $|f_k-f_k^*|<\lambda/n$ 至少以概率 $1-\beta$ 成立.

定理 6. 设 μ_k^* 与 μ_k 分别为键 k 所对应值的估计均值与真实均值.下列等式至少以概率 $1-\beta$ 成立.

$|\mu_k-\mu_k^*|\leq\frac{1}{nf_k/\lambda-1}$,其中 n 为用户个数,f_k 为键 k 的频率.

证明. 由算法 1 的第 10 行可知 k 对应值的估计均值为 $u_k^*=\frac{c^*(1)-c^*(-1)}{c^*(1)-c^*(-1)}$.设 $\theta_1=c^*(1)-c(1)$,$\theta_{-1}=c^*(-1)-c(-1)$.

根据式(2)可知,$f_k=\frac{d}{n}[c(1)+c(-1)]$,以及 $f_k^*=\frac{d}{n}[c^*(1)+c^*(-1)]$.则 $|f_k-f_k^*|=\frac{d}{n}|\theta_1+\theta_{-1}|$ 成立.根据定理 5 可知 $\Pr\left(|f_k-f_k^*|<\frac{\lambda}{n}\right)\geq 1-\beta$ 成立,则 $\Pr\left(|\theta_1+\theta_{-1}|<\frac{\lambda}{d}\right)\geq 1-\beta$ 成立.进而使得下列不等式成立:

$$|u_k-u_k^*|=\left|\frac{c(1)-c(-1)}{c(1)+c(-1)}-\frac{c^*(1)-c^*(-1)}{c^*(1)+c^*(-1)}\right|$$

$$=\left|\frac{c(1)-c(-1)}{c(1)+c(-1)}-\frac{c(1)-c(-1)+\theta_1-\theta_{-1}}{c(1)+c(-1)+\theta_1+\theta_{-1}}\right|$$

$$=\left|\frac{[c(1)-c(-1)](\theta_1+\theta_{-1})-[c(1)+c(-1)](\theta_1-\theta_{-1})}{[c(1)+c(-1)]\times[c(1)+c(-1)+\theta_1+\theta_{-1}]}\right|$$

$$=\left|\frac{(\theta_1+\theta_{-1})\times u_k-(\theta_1-\theta_{-1})}{\frac{nf_k}{d}+\theta_1+\theta_{-1}}\right|$$

$$\leq\frac{|\theta_1|(1-u_k)+|\theta_{-1}|(1+u_k)}{\frac{nf_k}{d}+\theta_1+\theta_{-1}}\leq\frac{\frac{\lambda}{d}}{\frac{nf_k}{d}-\frac{\lambda}{d}}$$

$$=\frac{1}{nf_k/\lambda-1},$$

则定理 6 成立.

定理 2 至定理 6 从频率与均值的无偏性、方差以及最大偏差的角度分析了 LRR 的可用性. 接下来结合当前基于整个值域的频率估计方法 Rappor[4] 与 KRR[15],以及针对键-值数据的均值与频率估计方法 PrivKV[3]、PrivKVM[3] 与 KVOH[5],阐述 LRR 方法的可用性. 根据 LRR 方法的描述可知,该方法利用伯努利抽样技术随机抽取某一键-值对进行本地扰动. 该操作减少了 LRR 方法的渐进误差. 给定键的值域大小 d 以及用户 u_i 转化后的键-值对集合 $\boldsymbol{S}_i'$. 如果 LRR 方法对 $\boldsymbol{S}_i'$ 中的每个键-值对均本地扰动,则 $p=e^{\frac{\varepsilon}{d}}/e^{\frac{\varepsilon}{d}}+2$, $q=1/e^{\frac{\varepsilon}{d}}+2$. 分别利用此时的 p 与 q 代入定理 5 可知估计键 k 频率所产生的渐进误差为 $O(d^2\sqrt{\log(d/\beta)}/\varepsilon\sqrt{n})$. 本文 LRR 方法利用伯努利抽样技术,则估计键 k 频率所产生的渐进误差为 $O(d\sqrt{\log(d/\beta)}/\varepsilon\sqrt{n})$.

同理,利用 Rappor[4] 方法对 $\boldsymbol{S}_i'$ 中的每个键-值对扰动,则 $p=e^{\frac{\varepsilon}{d}}/e^{\frac{\varepsilon}{d}}+1$, $q=1/e^{\frac{\varepsilon}{d}}+1$. 估计键 k 频率所产生的渐进误差为 $O(d^2/\varepsilon\sqrt{n})$. 此时 Rappor 方法产生的渐进误差大于本文的基于抽样技术的 LRR 方法,即 $O(d^2/\varepsilon\sqrt{n})>O(d\sqrt{\log(d/\beta)}/\varepsilon\sqrt{n})$. 此外,利用 KRR[15] 方法扰动 $\boldsymbol{S}_i'$ 中所有键-值对后,估计键 k 频率所产生的渐进误差为 $O(d^2\sqrt{d}/\varepsilon\sqrt{n})$. 此时 KRR[15] 方法产生的渐进误差大于 LRR 方法,即 $O(d^2\sqrt{d}/\varepsilon\sqrt{n})>O(d\sqrt{\log(d/\beta)}/\varepsilon\sqrt{n})$.

根据上述分析可知,如果 Rappor 方法与 KRR 方法分别作用于 $\boldsymbol{S}_i'$ 中的每个键-值对,则产生的频率估计误差均高于 LRR 方法. 接下来利用定理 7 阐述附带 LRR 方法的 LDPKV 优于现有键-值对频率与均值估计方法 PrivKV、PrivKVM 与 KVOH.

定理 7.　结合现有的键-值对收集方法 PrivKV、PrivKVM 与 KVOH 产生的方差,以及当前可知,LDPKV 优于这四种方法.

证明.　类似于定理 3 与定理 4,给定键 k,PrivKV、PrivKVM 以及 KVOH 方法估计键 k 频率产生的误差为 $\frac{d}{n}\times\frac{e^{\varepsilon/2}}{(e^{\varepsilon/2}-1)^2}$. LDPKV 方法估计键 k 频率产生的误差为 $\frac{d}{n}\times\frac{2e^{\varepsilon}}{(e^{\varepsilon}-1)^2}$. 给定隐私预算 ε,可知 $\frac{d}{n}\times\frac{e^{\varepsilon/2}}{(e^{\varepsilon/2}-1)^2}\geq\frac{d}{n}\times\frac{2e^{\varepsilon}}{(e^{\varepsilon}-1)^2}$,则在基于键 k 频率产生误差的情况下,LDPKV 方法优于 PrivKV、PrivKVM 以及 KVOH 方法.

根据定理 4 可知,PrivKV、PrivKVM 以及 KVOH 方法的均值方差为

$$Var_{\text{PrivKV}}[\mu_k^*]=Var_{\text{PrivKVM}}[\mu_k^*]=$$
$$4M(e^{\varepsilon}+1)e^{\varepsilon/2}\times\left[\frac{(c(1)-c(-1))^2(e^{\varepsilon/2}-1)^4}{((c(1)-c(-1))(e^{\varepsilon/2}-1)^2-4Me^{\varepsilon/2})^4}\right]+$$
$$4M(e^{\varepsilon}+1)e^{\varepsilon/2}\times\left[\frac{1}{((c(1)-c(-1))(e^{\varepsilon/2}-1)^2-4Me^{\varepsilon/2})^2}\right]\tag{13}$$

$$Var_{\text{KVOH}}[\mu_k^*]=$$
$$2Me^{\varepsilon/2}\times\left[\frac{((c(1)-c(-1))^2(e^{\varepsilon/2}-1))^2}{((c(1)+c(-1))(e^{\varepsilon/2}-1)-2M)^4}\right]+$$
$$2Me^{\varepsilon/2}\times\left[\frac{1}{((c(1)+c(-1))(e^{\varepsilon/2}-1)-2M)^2}\right]\tag{14}$$

根据式(13)与式(14)可知,PrivKV 的均值方差大于 KVOH 的均值方差,则 KVOH 方法优于 PrivKV 与 PrivKVM 方法. 同理,基于定理 4 可知,LDPKV 方法优于 KVOH.

5　实验结果与分析

实验平台是 4 核 Intel i7-4790 CPU(4 GHz),8GB 内存,Win7 系统. 所有算法均采用 Python 实现. 实验采用五个数据集 TalkingData、Jdata、GAUSS、PLAW 和 LNR. 其中 TalkingData 是从与移动应用的 SDK 中收集而得,该数据集包含 60822 个设备和 306 类应用程序. JData 数据集来自于 JD.com,其中包含 2016 年的 5060 万条销售记录,该数据集记录了 2016 年间 105180 个用户对 442 个品牌的 5060 万条购买记录. GAUSS、PLAW 和 LNR 三个数据集

为合成数据集，键和值分别遵循高斯、幂律和线性分布. 五种数据集具体细节如表 2 所示.

表 2 五种数据集信息描述

数据集	分布	用户数	Keys 数量
GAUSS	gaussian	1 000 000	100
PLAW	power-law	1 000 000	100
LNR	linear	1 000 000	100
Talking Data	—	60 822	306
JData	—	105 180	442

结合上述五种数据集，采用均方误差 *MSE* 和相对误差 *RE* 度量 Rappor、KRR(K-Random Response)、PrivKV、PrivKVM、KVOH、LDPKV 以及 Harmony[7] 方法的频率和均值计算精度. 其中 KRR 是单值频率估计方法的典型代表，Harmony 是高维分类数据与数值型数据下估计频率与均值的典型代表. 键与值所对应的均方误差 *MSE* 可以表示为 $\frac{1}{d}\times\sum_{k\in K}(u_k-u_k^*)^2$ 与 $\frac{1}{d}\times\sum_{k\in K}(f_k-f_k^*)^2$，相对误差 *RE* 可表示为 $RE=\frac{|f_k-f_k^*|}{f_k}$，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.

5.1 实验结果比较

隐私参数 ε 的取值为 0.1、0.2、0.4、0.8、1.6、3.2. 实验中分别对五种数据集中的频率最高的 top10、top20、top30、top40、top50 的键-值进行的统计，在每种情况下进行 500 次试验并求取平均值作为最终的计算结果.

5.1.1 基于 TalkingData 数据集的七种算法的 *MSE* 和 *RE* 值比较

图 1(a)～(f)描述了 Rappor、KRR、PrivKV、PrivKVM、KVOH 以及 LDPKV 算法 *MSE* 值和 *RE* 值的比较结果. 由图 1(a)的验结果可以发现，当 ε 从 0.1 变化到 3.2 时，六种方法的 *MSE* 值均减少，而 LDPKV 方法的精度优于其它五种方法. 当 ε=3.2 时，LDPKV 方法的精度是 Rappor、KRR 算法的近 4 倍，是 PrivKV、PrivKVM 和 KVOH 算法的将近 2 倍. 同时从图 1(c)～(f)中发现，当 ε 固定时，键的数目从 top10 增加到 top50，六种方法的 *RE* 值均增大，且 LDPKV 算法的精度同样优于其它四种方法，其原因在于随着键-值个数增加，累计的误差也就越多，导致精度越差. 图 1(b)是 Rappor、Harmony、PrivKV、PrivKVM、KVOH 以及 LDPKV 方法关于均值的 *MSE* 值比较结果. 从实验结果可以发现，ε 从 0.1 变化到 3.2 时，六种方法关于均值的 *MSE* 值均呈现下降趋势，然而 LDPKV 方法的均值精度明显优于其它五种方法，并且其下降趋势明显大于其它五种方法，其原因在于 LDPKV 方法将键-值对看成一个整体进行编码扰动，避免了隐私预算的分割.

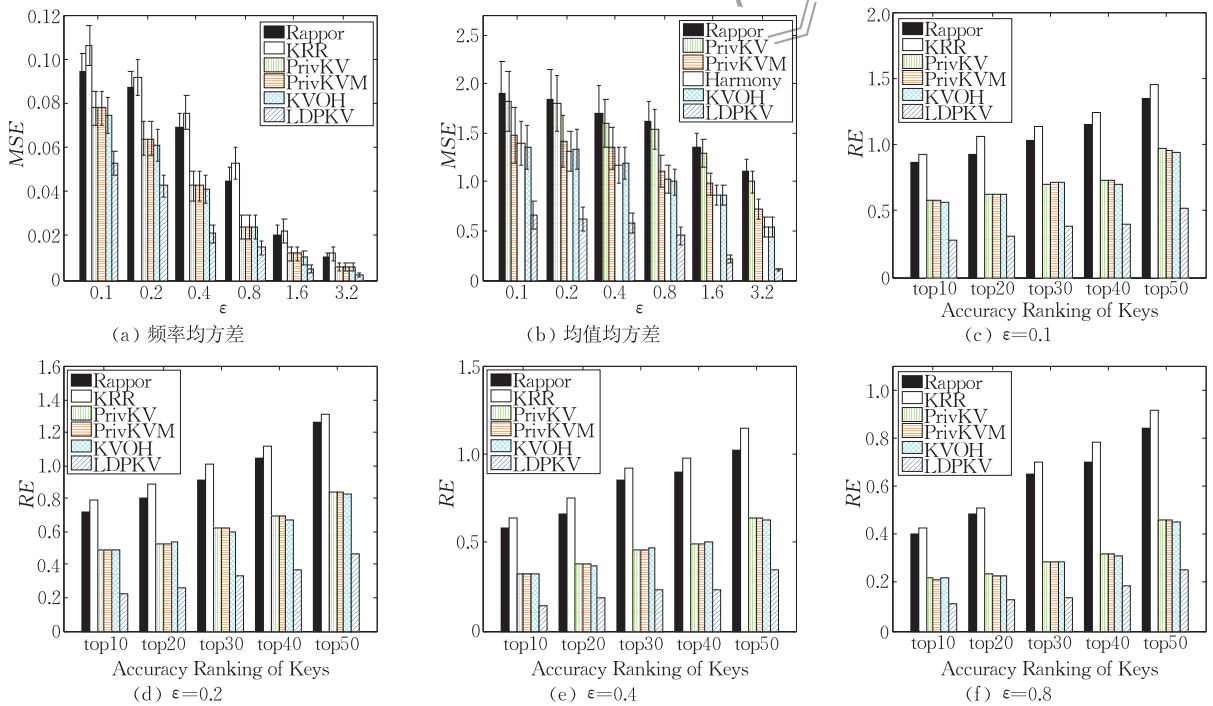


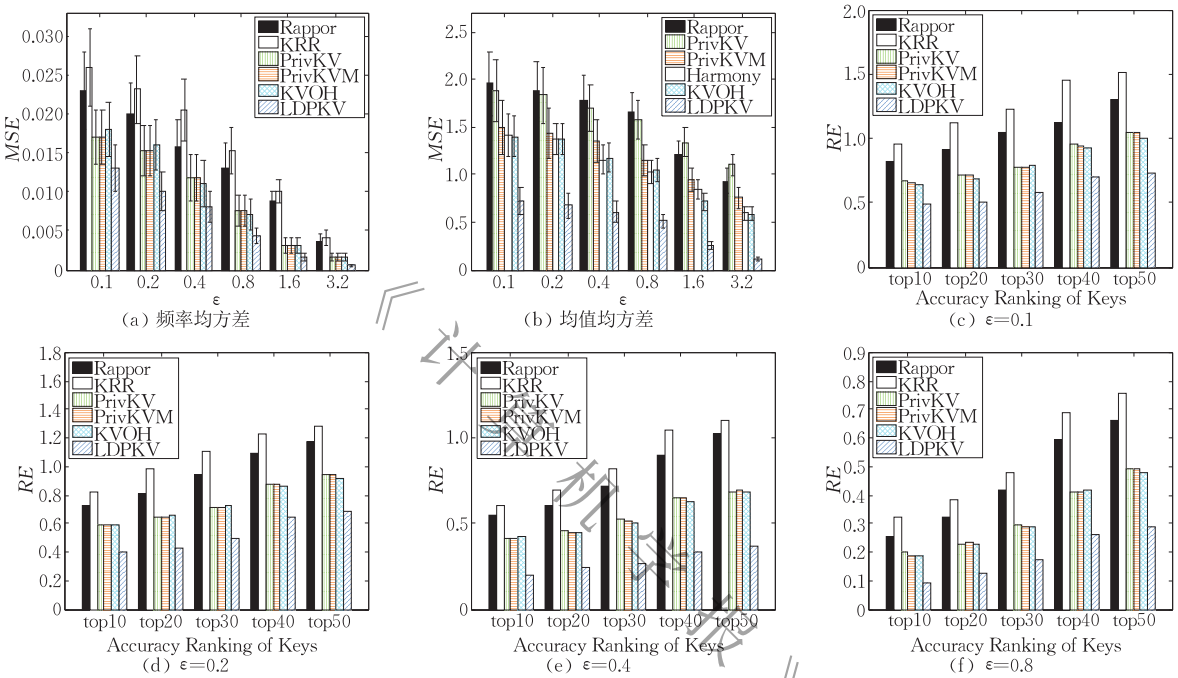
图 1 TalkingData 数据集频率和均值计算结果

5.1.2　基于 JData 数据集的七种算法的 *MSE* 和 *RE* 值比较

图 2(a)～(f)描述了 Rappor、KRR、PrivKV、PrivKVM、KVOH 以及 LDPKV 方法 *MSE* 值和 *RE* 值的比较结果. 由图 2(a)的实验结果可以发现，当 ε 从 0.1 变化到 3.2 时，六种方法的 *MSE* 值均减少，当 ε＝1.6 时，LDPKV 方法的精度是 Rappor、KRR 算法的近 6 倍，是 PrivKV、PrivKVM 和 KVOH 方法的将近 2 倍. 图 2(b)描述了 Rappor、Harmony、PrivKV、PrivKVM、KVOH 以及 LDPKV 方法有关均值的 *MSE* 值比较，从实验结果发现，当 ε 从 0.1 变化到 3.2 时，LDPKV 方法的均值精度优于其它五种方法. 当 ε＝3.2 时，LDPKV 方法的精度是 PrivKV 方法的近 8 倍，是 PrivKVM 算法的近 6 倍. 其原因为 LDPKV 算法不仅避免了隐私预算的分割，同时避免了 PrivKV 和 PrivKVM 方法中对键所对应值的多次扰动.

5.1.3　基于 LNR、GAUSS、PLAW 数据集的七种算法的 *MSE* 和 *RE* 值比较

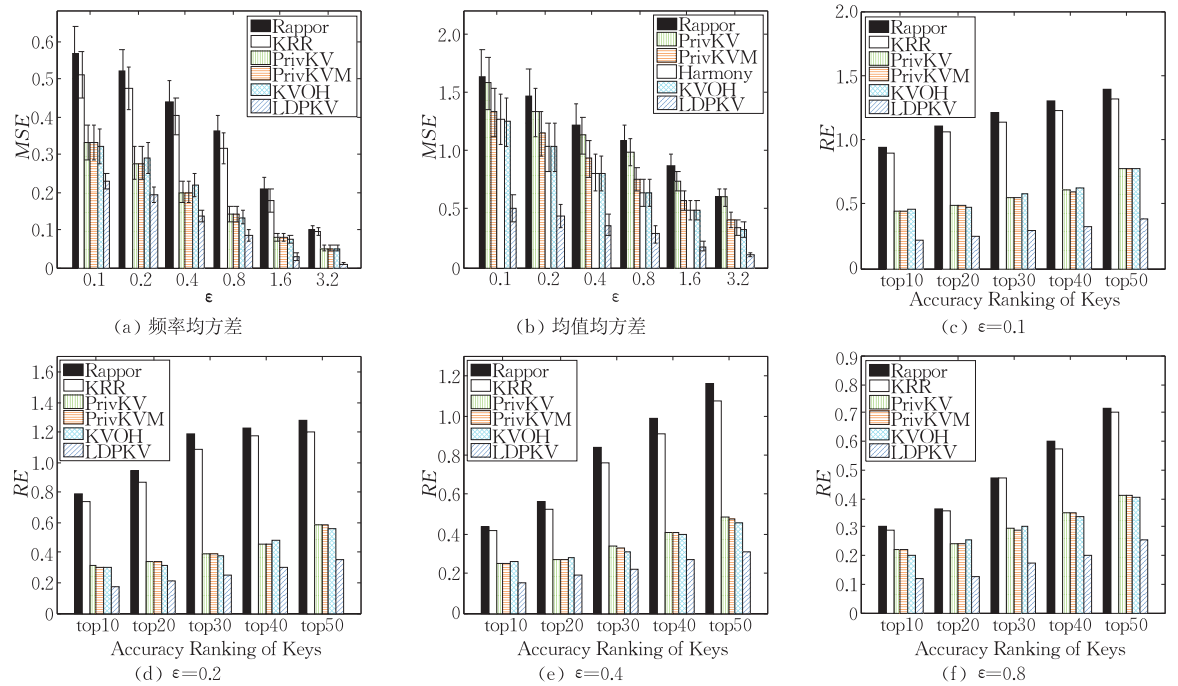
图 3～图 5 分别描述了 Rappor、KRR、PrivKV、



(a) 频率均方差　(b) 均值均方差　(c) ε=0.1

(d) ε=0.2　(e) ε=0.4　(f) ε=0.8

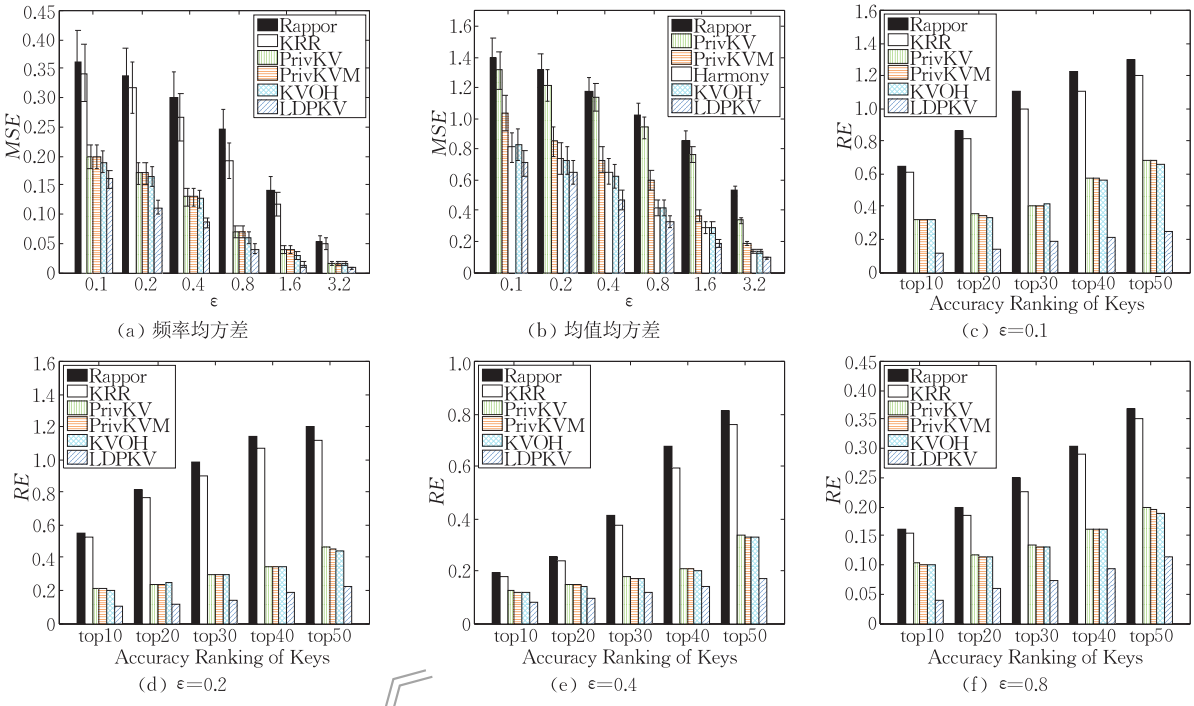
图 2　JData 数据集频率和均值计算结果



(a) 频率均方差　(b) 均值均方差　(c) ε=0.1

(d) ε=0.2　(e) ε=0.4　(f) ε=0.8

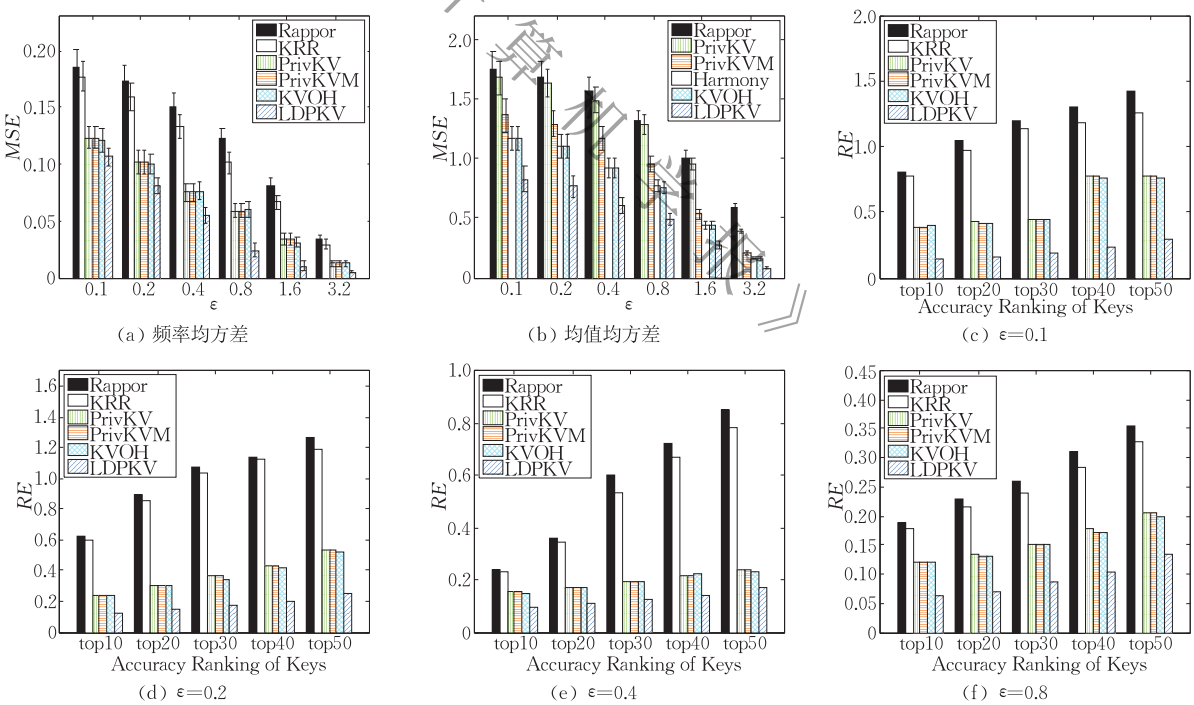
图 3　LNR 数据集频率和均值计算结果



(a) 频率均方差 (b) 均值均方差 (c) ε=0.1

(d) ε=0.2 (e) ε=0.4 (f) ε=0.8

图 4 GAUSS 数据集频率和均值计算结果



(a) 频率均方差 (b) 均值均方差 (c) ε=0.1

(d) ε=0.2 (e) ε=0.4 (f) ε=0.8

图 5 PLAW 数据集频率和均值计算结果

PrivKVM、KVOH 以及 LDPKV 方法在 LNR、GAUSS、PLAW 数据集上的 *MSE* 值和 *RE* 值的比较结果. 由实验结果可以发现，当 ε 从 0.1 变化到 3.2 时，LDPKV 方法的精度优于其它六种方法. 当固定键-值的个数时，从图 3、图 4、图 5(c)～(f)的实验结果可以看出，当 ε 从 0.1 变化到 0.8 时，六种算法在三种数据集上的 *RE* 值均下降. 当 ε=0.2 时，键-值等于 top10 时，LDPKV 方法的精度相对于 Rappor、KRR 方法，较 LNR 数据集上达到了近 4 倍，较 GAUSS 数据集上达到了近 5 倍，在 PLAW 数据集上达到了近 6 倍. 其主要原因为 LDPKV 方法将键-值对看成一个整体进行编码扰动，减少了隐私预算分割和扰动次数.

6 总　结

本文针对键-值数据的收集与分析问题，提出了一种基于本地差分隐私的键-值数据收集方法 LDPKV. 该方法在不分割隐私代价的情况下，每个用户利用离散化、本地扰动以及伯努利抽样技术统一保护了自己的键-值数据，这几种操作既维护了键与值之间的关联性，还保持了用户与收集者之间较小的通信代价. 收集者结合整体扰动的结果，对每个键的频率以及键所对应值的均值进行估计. 从理论角度分析了 LDPKV 方法估计频率与均值产生的误差以及最大偏差. 从实验角度分析了 LDPKV 与同类方法相比具有较好的估计精度. 今后的研究考虑如下两个方面：(1) 如何利用直方图技术估计键的频率与均值；(2) 由于键-值的值域通常比较稀疏，如何设计高精度的本地扰动方法来克服此类问题.

参 考 文 献

[1] Duchi J C, Jordan M I, Wainwright M J. Local privacy and statistical minimax rates//Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science. Berkeley, USA, 2013: 429-438

[2] Duchi J C, Jordan M I, Wainwright M J. Minimax optimal procedures for locally private estimation. Journal of the American Statistical Association, 2018, 113(521): 182-201

[3] Ye Q, Hu H, Meng X, et al. PrivKV: Key-value data collection with local differential privacy//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2019: 127-143

[4] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale, USA, 2014: 1054-1067

[5] Sun L, Zhao J, Ye X, et al. Conditional analysis for key-value data with local differential privacy. arXiv: 1907.05014, 2019

[6] Acharya J, Sun Z. Communication complexity in locally private distribution estimation and heavy hitters//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 51-60

[7] Nguyên T T, Xiao X, Yang Y, et al. Collecting and analyzing data from smart device users with local differential privacy. arXiv: 1606.05053, 2016

[8] Wang N, Xiao X, Yang Y, et al. Collecting and analyzing multidimensional data with local differential privacy//Proceedings of the 35th IEEE International Conference on Data Engineering. Macao, China, 2019: 638-649

[9] Wang T, Li N, Jha S. Locally differentially private frequent itemset mining//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2018: 127-143

[10] Qin Z, Yu T, Yang Y, et al. Generating synthetic decentralized social graphs with local differential privacy//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA, 2017: 425-438

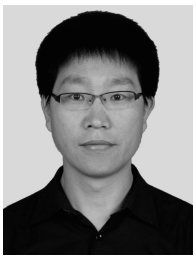
[11] Cormode G, Kulkarni T, Srivastava D. Answering range queries under local differential privacy. Proceedings of the VLDB Endowment, 2019, 12(10): 1126-1138

[12] Wang T, Blocki J, Li N, et al. Locally differentially private protocols for frequency estimation//Proceedings of the 26th USENIX Security Symposium. Vancouver, Canada, 2017: 729-745

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

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

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



ZHANG Xiao-Jian, Ph. D., associate professor. His research interests include differential privacy and database.

FU Nan, M. S. candidate. His research interests include local differential privacy and database.

MENG Xiao-Feng, Ph. D., professor, Ph. D. supervisor. His research interests include data fusion and knowledge fusion, big data privacy management, and big data analysis, etc.

Background

Key-value data collection is the workhorse of analysis of big data, which is a popular storage paradigm designed for NoSQL database. Key-value data, however, is inherently sensitive. Directly collecting and analyzing such data may

break the privacy of individual's. For example, an attacker will rely on key-value data protected by k-anonymity and the background knowledge to discover the shopping behavior of individuals. While the impractical background assumptions of k-anonymity make the proposed solutions weak. Local differential privacy is the state-of-the-art privacy notion that provides the strongest privacy guarantee independent of an adversarys background knowledge. Therefore, a series of studies have been proposed on different types of data, such as spatial data, streaming data, web data, etc. While seldom work with local differential privacy focused on key-value data. In this paper, the authors employ the discretization and local randomized response model to sanitize the key-value pair. With the sanitized pair, an attacker cannot learn any user's sensitive information.

WAMDM lab has studied privacy protection since 2016, a series of interesting problems are solved, and related works are published in VLDB, IEEE TKDE, SIAM SDM, and S&P, etc. We have studied differential privacy since 2012 and solved several key problems, such as histogram release, spatial data release, high-dimensional data release, etc. Related research papers are published in SIAM SDM, Frontiers of Computer Science, and Chinese Journal of Computers, etc.

This work was partially supported by the National Natural Science Foundation of China (61502146, 91746115, 91646203, 61772131), the Natural Science Foundation of Henan Province (162300410006), Training Program of Young Backbone Teachers in Colleges and Universities of Henan Province (2017GGJS084), the Research Program of the Higher Education of Henan Educational Committee (16A520002), and the Young Talents Fund of Henan University of Economics and Law.