Locally Differentially Private and Fair Key-Value Aggregation

Yukun Dong
yukun@udel.edu
University of Delaware
Newark, DE, USA

Zhengwu Lu
zhengwulu@myyahoo.com
Wuhan Engineering Science & Technology Institute
Wuhan, Hubei, China

Rui Zhang
ruizhang@udel.edu
University of Delaware
Newark, DE, USA

ABSTRACT

In the era of Big Data, the ability to extract meaningful insights from vast datasets while maintaining individual privacy has become an increasingly complex challenge. Recent years have witnessed the development of various locally differentially private data aggregation schemes which allow an untrusted data collector to derive meaningful statistics from user data while maintaining strong privacy guarantee for individual users. As a fundamental data type in NoSQL databases, key-value data has two important statistics of interest, the frequency of each key and the corresponding mean value. Current locally differentially private key-value aggregation schemes primarily rely on uniform sampling for mean estimation, i.e., a single key-value pair is selected randomly from each user's key-value set. This approach, however, results in high mean estimation accuracy for frequent keys and low accuracy for infrequent ones. To tackle this problem, this paper presents the design and evaluation of Adaptive, a novel locally differentially private and fair key-value aggregation scheme that can deliver uniformly high mean estimation accuracy across different keys. In the first phase, we utilize a portion of the privacy budget to estimate the frequency of each key. Subsequently, based on the key frequencies estimated in the first phase, we employ non-uniform random sampling for mean estimation, which enables higher probability sampling of values associated with low-frequency keys. Comprehensive theoretical analysis and simulation studies confirm the superiority of Adaptive over previous solutions.

CCS CONCEPTS

• **Security and privacy**; • **Security services**; • **Privacy-preserving protocols**;

KEYWORDS

Key-value; Local Differential Privacy; frequency estimation; mean estimation; fairness

ACM Reference Format:
Yukun Dong, Zhengwu Lu, and Rui Zhang. 2023. Locally Differentially Private and Fair Key-Value Aggregation . In *The 12th International Symposium on Information and Communication Technology (SOICT 2023), December 07–08, 2023, Ho Chi Minh, Vietnam.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3628797.3628807

SOICT 2023, December 07–08, 2023, Ho Chi Minh, Vietnam

ACM ISBN 979-8-4007-0891-6/23/12…$15.00
https://doi.org/10.1145/3628797.3628807

1 INTRODUCTION

In the era of Big Data when massive volumes of user-generated information are continuously collected and analyzed, privacy concerns become particularly acute. Differential Privacy (DP) [6, 9] has become the *de facto* technique for private data release in large-scale environments. This allows data collectors to gather extensive datasets and publish statistics while preserving user privacy. However, one significant limitation arises: under DP, collectors still have access to the raw, and often sensitive, data of users, leading to amplified privacy risks especially when the data collector is not entirely trusted. To address this limitation, Local Differential Privacy (LDP) [3, 4] has emerged as a solution for privacy-preserving data analytics without requiring a trusted data collector. In an LDP framework, data owners individually perturb their raw data to shield their privacy, enabling the data collector to perform statistical analysis on this anonymized dataset. This approach is especially relevant for Big Data applications and has been widely adopted by major technology companies. Notable instances include Google's RAPPOR used in Chrome browser for collecting usage information [10], Apple's LDP adoption in Safari for detecting popular emoji and identifying high energy or memory usage [20, 21], and Microsoft's adoption for collecting telemetry data [2].

While LDP has been extensively utilized for data privacy protection, the majority of existing research focuses on fundamental queries over simple data types. Examples include mean estimation over numeric data [15, 23], frequency or heavy hitter estimation over set-value data [10, 17, 25], marginal release over multidimensional data [14, 30], and synthetic graph release over graph data [18]. However, hybrid queries involving multiple data types remain underexplored. For instance, key-value data, a vital data type in burgeoning NoSQL databases, consists of a categorical key with one or more numeric values. The two important queries involving key-value data include estimating the frequency distribution of key and the mean value estimation for the numeric values of each distinct key. If we treat these two queries independently, the data utility in terms of the query result accuracy would be significantly reduced due to the intrinsic correlation between key and value.

Several studies have considered the inherent correlation between key and value in order to enable locally differentially private key-value data aggregation while enhancing data utility. Specifically, PriKVM [28] samples a random key from the full key domain for frequency estimation and uses a multi-round iteration process to enhance the accuracy of mean value estimation for each key. PriKVM works well when the key domain is limited in size, but it suffers from low estimation accuracy when dealing with larger key domains due

to inadequate samples. To address this large-domain issue, PCKV [12] was proposed, which adopts padding and sampling techniques [17] where users randomly sample one key-value pair from their owned pairs instead of from the full domain. Further research [19] proposed more effective and stable locally differentially private mechanisms for handling conditional analysis related to key-value data. However, a common limitation across these solutions is the uniform random sampling of one key-value pair from each user's set for value mean estimation. This method leads to high accuracy for frequently occurring keys and low accuracy for infrequent keys due to the variation in the number of samples. How to realize locally differentially private key-value aggregation with uniformly high mean estimation accuracies across different keys of varying frequencies remains an open challenge.

In this paper, we tackle this challenge by introducing Adaptive, a novel LDP and fair key-value aggregation scheme. We note that the mean estimation accuracy for a key hinges on the number of associated key-value pairs sampled: the more pairs sampled, the higher the accuracy, and vice versa. With uniformly random sampling, values linked with infrequent keys have a lower likelihood of selection, necessitating the adoption of non-uniform sampling to elevate the sampling probability for such keys and thus balance the sample count across different keys. Our proposed Adaptive scheme operates in two phases based on these observations. In the first phase, we estimate the frequency of each key with an LDP guarantee, using a portion of the privacy budget. In the second phase, we estimate the mean value for each key through non-uniform sampling, where the sampling probability for each key pair is established based on the key frequency estimated in the initial phase. This method enables Adaptive to guarantee LDP while also achieving uniformly high mean estimation accuracy across different keys. The main contributions of this paper can be summarized as follows.

- To the best of our knowledge, we are the first to study locally differentially private and fair aggregation for key-value data.
- We introduce Adaptive, a novel locally differentially private and fair key-value aggregation scheme that can guarantee local differential privacy and achieve uniformly high mean estimation accuracy across different keys.
- Detailed theoretical analysis and simulation studies confirm the advantages of our proposed scheme over prior solutions.

The rest of this paper is structured as follows. Section 2 discusses the related work. Section 3 introduces some preliminaries and formulates the problem. Section 4 presents the design of Adaptive. Section 5 analyzes the privacy guarantee of Adaptive. We report the simulation results in Section 6 and finally conclude this paper in Section 7.

2 RELATED WORK

As one of the fundamental statistics, frequency estimation has been extensively studied under the LDP framework. Erlingsson *et al.* [10] introduced RAPPOR as a mean to collect browsing data in Chrome. This approach involves encoding original sensitive data into a binary vector using a Bloom filter and applying the randomized response mechanism to perturb each bit. Subsequent work, RAPPOR-unknown [11], extended this method by enabling the identification of frequency items without requiring an explicit dictionary. Bassily *et al.* [1] introduced Bitstogram and TreeHist to address the high communication cost associated with RAPPOR. Moreover, Wang *et al.* proposed PEM [26] and PrivTrie [22] to identify 'heavy hitters' by scanning a binary prefix tree. Qin *et al.* were the first to consider frequency estimation over set-valued data by proposing a two-phase method called LDP-Miner [17]. Building upon this, Wang *et al.* introduced SVIM [25], which has been proved to be more efficient in estimating frequent items or itemsets. Furthermore, they introduced a framework for choosing the optimal parameter to enhance accuracy and considered consistency issues in the estimation process [24].

Mean estimation represents another well-researched statistic. The most straightforward approach to address this problem is to apply the Laplace mechanism [8]. However, this method introduces unbounded noise to the true value, which often results in unsatisfactory estimation. Duchi *et al.* [5] was the first to investigate this issue and proposed MeanEst [5], which involves transforming numeric data into binary data, followed by the application of the randomized response mechanism. Nonetheless, this solution suffers from high space complexity and communication costs. Nguyen *et al.* further proposed Harmony [16] with reduced computation complexity and higher estimation accuracy. Moreover, Wang *et al.* introduced the Piecewise Mechanism [23], which restricts the original numeric value to a small domain and enables the perturbed value to remain close to the original with high probability. More recently, Li et al. *et al.* proposed the Square Wave mechanism [15], which leverages the numerical ordered nature of the domain to achieve a better balance between privacy and utility.

In contrast to the above two research directions, the aggregation of key-value data has only begun to garner attention recently. Ye *et al.* proposed PrivKVM [28], a mechanism that adaptively combines a randomized response for frequency estimation and the Harmony mechanism for mean estimation, while preserving the correlation between keys and values. This approach also utilizes a multi-round iteration to improve the accuracy of mean estimation. Subsequently, Sun *et al.* [19] proposed a conditional estimator for frequency and mean based on PrivKVM. Gu *et al.* introduced PCKV [12] to address performance issues in large key domains by employing padding and sampling techniques to select one key-value pair from each user's dataset, instead of from the entire key domain. Ye *et al.* [29] also addressed the same issue with a two-phase framework which identifies frequent keys in the first phase and improves the estimate in the second phase through adaptive sampling. However, none of these studies addressed the non-uniform mean estimation accuracy across different keys that results from varying key frequencies.

3 PRELIMINARIES

In this section, we first present the definition of Local Differential Privacy and our problem formulation. We then introduce two mechanisms, Randomized Response and Piecewise Mechanism, which will serve as the building blocks of our proposed scheme.

3.1 Local Differential Privacy

Local Differential Privacy [3, 4] is a privacy framework in data analysis wherein individual data contributors apply a randomized

mechanism to their data locally before submitting them to a untrusted data collector. This mechanism ensures that the statistical properties of the original data remain preserved, while obscuring individual data and preventing unauthorized disclosure of sensitive information. Formally, ϵ-Local Differential Privacy is a rigorous notion for characterizing the privacy guarantee offered by a randomized mechanism, which is defined below.

Definition 1. *(ϵ-Local Differential Privacy). A randomized mechanism $\mathcal{M}$ satisfies ϵ-Local Differential Privacy if and only if*

$$\frac{\Pr[\mathcal{M}(x) = y]}{\Pr[\mathcal{M}(x') = y]} \leq e^{\epsilon}, \tag{1}$$

for any two possible input x and x' and any output y, where $\epsilon \geq 0$.

The parameter ϵ is commonly referred to as the *privacy budget*. The smaller ϵ is, the more indistinguishable of the two probability distributions induced by any two input values, the more difficult for the adversary to distinguish two input values from the perturbed value, the stronger privacy guarantee for individual users, and vice versa.

3.2 Problem Formulation

We consider a system with a data collector and n users. Each user i possesses a set of key-value pairs denoted by S_i. Without loss of generality, assume that the key domain is $\mathcal{K} = \{1, 2, \ldots, d\}$ and that every key has a value domain $V = [-1, 1]$. Each key-value pair has a form of $\langle k, v\rangle$, where $k \in \mathcal{K}$ and $v \in V$. The value distribution of different keys may or may not be the same. We assume all users have the same number of key-value pairs, i.e. $|S_i| = l$ for all $1 \leq i \leq n$, for some known parameter l. We leave the extension of our work to support varying number of key-value pairs across different users as our future work.

The data collector intends to learn two statistics about the users' data, including the frequency and corresponding value mean for each key $k \in \mathcal{K}$ defined below.

- **Key frequency**: the frequency of a key k is the ratio of users who possess k which is defined as

$$f_k = \frac{|\{i | 1 \leq i \leq n, \exists \langle k, v\rangle \in S_i\}|}{n}. \tag{2}$$

  Let n_k be the number of users who possess key k for all $1 \leq k \leq d$. We have $f_k = n_k/n$. Therefore, estimating f_k is equivalent to estimating n_k.
- **Value mean**: the value mean of a key k is the average of all the values associated with key k across all users who possess it which is defined as

$$m_k = \frac{\sum_{i=1, \langle k, v\rangle \in S_i}^{n} v}{n_k} \tag{3}$$

We assume that the data collector is honest but curious. While the collector is trusted to carry out all system operations faithfully, it is interested in inferring users' true data value from the reported information. We assume the data collector knows the mechanism being deployed as well as all system parameters.

We seek to develop a locally differentially private and fair key-value aggregation scheme that can satisfy ϵ-LDP while ensuring uniformly high estimation accuracies across different keys with diverse frequencies.

3.3 Randomized Response

Randomized Response (RR) [27] is a classical surveying technique developed by Warner in the 1960s for collecting binary data about sensitive topic that respondents might be reluctant to answer unless their privacy can be protected. Specifically, given a question with binary answer "Yes" or "No", the respondent will respond truthful with probability p, and respond the opposite answer with probability $q = 1 - p$. In order to satisfy ϵ-LDP, p needs to be set to $\frac{e^{\epsilon}}{e^{\epsilon}+1}$. As its name suggested, the Generalized Randomized Response (GRR) generalizes the RR to allow the collection of categorical data with domain $D = \{1, 2, \ldots, d\}$. Specifically, a user with a value $v \in D$ reports a perturbed value v' generated according to the following probability distribution

$$\Pr(v'|v) = \begin{cases} \frac{e^{\epsilon}}{e^{\epsilon}+d-1}, & \text{if } v' = v, \\ \frac{1}{e^{\epsilon}+d-1}, & \text{if } v' \in D \setminus \{v\}. \end{cases}$$

It has been proved that the GRR satisfies ϵ-LDP.

3.4 Piecewise Mechanism

Piecewise Mechanism (PM) is a locally differentially private scheme [23] proposed for perturbing numerical value. Let $C = \frac{e^{\epsilon/2}+1}{e^{\epsilon/2}-1}$, where ϵ is the privacy budget. Given a true value $v \in [-1, 1]$, the PM returns a perturbed $v^* \in [-C, C]$ according to the following probability density function

$$f[v^*|v] = \begin{cases} p & \text{if } v^* \in [l(v), r(v)], \\ \frac{p}{\exp(\epsilon)} & \text{if } v^* \in [-C, l(v)) \cup (r(v), C], \end{cases}$$

where

$$\begin{cases} p = \dfrac{\exp(\epsilon) - \exp(\epsilon/2)}{2\exp(\epsilon/2) + 2}, \\ l(v) = \dfrac{C+1}{2}v - \dfrac{C-1}{2}, \\ r(v) = \dfrac{C+1}{2}v + \dfrac{C-1}{2}. \end{cases}$$

Besides satisfying ϵ-LDP, an important property of PM is that $\mathbb{E}[v^*] = v$ for all $v \in [-1, 1]$ [23].

4 DESIGN OF ADAPTIVE

In this section, we first give an overview of the proposed Adaptive scheme and then detail its design.

4.1 Overview

We find that the key to achieve uniformly high mean estimation accuracy across different keys with varying frequencies is to ensure adequate samples for key-value pairs associated with low frequency keys. Therefore, it is necessary to increase the sampling probabilities of the key-value pairs associated with infrequent keys while reducing the sampling probabilities of those associated with frequent keys. However, the key frequencies are one of the two statistics that we need to estimate and are not known in advance. As a result, we decouple key frequency estimation and key value mean estimation into two phases. The first phase focuses on the estimation of key frequencies. The second phase explores non-uniform random sampling to adjust the number of key-value pairs sampled for each key to achieve a more uniform estimation accuracy of the

mean value across different keys. In what follows, we detail the two phases of the Adaptive scheme.

4.2 Phase 1: Key Frequency Estimation via Generalized Randomized Response

In the first phase, each user i samples one key-value pair uniformly at random, and then perturbs the sampled key using the GRR mechanism [27] with a privacy budget of ϵ_1. Suppose that the sampled key pair is $\langle k_i, v_i \rangle$. User i perturbs key k_i into k_i' according to the following probability distribution

$$\Pr[k_i'|k_i] = \begin{cases} p_1, & \text{if } k_i' = k_i, \\ q_1, & \text{if } k_i' \in \mathcal{K} \setminus \{k_i\}, \end{cases} \tag{4}$$

where $p_1 = \frac{e^{\epsilon_1}}{e^{\epsilon_1}+d-1}$ and $q_1 = \frac{1}{e^{\epsilon_1}+d-1}$. Each user i then submits his perturbed key k_i' to the data collector.

On receiving the perturbed keys $k_1', \ldots, k_n'$ from n users, the collector estimates the frequency of each key. Let f_k' be the frequency of received perturbed key for each $k \in \mathcal{K}$. The data collector estimates the frequency of original key k as

$$\hat{f}_k = \frac{f_k' - q_1}{p_1 - q_1}, \tag{5}$$

which is an unbiased estimator of f_k [27].

4.3 Phase 2: Mean Estimation via Non-uniform Sampling and Joint Perturbation

In the second phase, the data collector first announces the estimated frequency of each key $\langle \hat{f}_1, \ldots, \hat{f}_d \rangle$ obtained in the first phase to all users.

Each user then samples one key-value pair from his set via non-uniform random sampling. Let $\theta \geq 0$ be a system parameter. Each user i samples one pair $\langle k_i, v_i \rangle$ from his set S_i according to the following probability distribution

$$\Pr\left[\langle k_i, v_i \rangle = \langle k, v \rangle\right] = \frac{\hat{f}_k^{-\theta}}{\sum_{\langle k,v \rangle \in S_i} \hat{f}_k^{-\theta}} \tag{6}$$

for all $\langle k, v \rangle \in S_i$. We can see that this sampling procedure is a generalization of the uniform random sampling. In particular, the sampling procedure is equivalent to uniform random sampling when $\theta = 0$. When $\theta > 0$, keys with lower estimated frequencies are sampled with higher probabilities than those with higher frequencies. When $\theta \to \infty$, the key with the lowest frequency in each S_i is sampled with probability one. Therefore, the parameter θ controls the degree of the non-uniformity of the random sampling procedure. We will evaluate the impact of θ in Section 6.

Next, each user i jointly perturbs the sampled key k_i and corresponding value v_i to preserve their correlation and allow the collector to estimate the value mean for each key. Specifically, each user i first perturbs the sampled key k_i into $\tilde{k}_i$ using GRR with a privacy budget of ϵ_2 according to the following probability distribution.

$$\Pr[\tilde{k}_i|k_i] = \begin{cases} p_2, & \text{if } \tilde{k}_i = k_i, \\ q_2, & \text{if } \tilde{k}_i \in \mathcal{K} \setminus \{k_i\}, \end{cases} \tag{7}$$

where $p_2 = \frac{e^{\epsilon_2}}{e^{\epsilon_2}+d-1}$ and $q_2 = \frac{1}{e^{\epsilon_2}+d-1}$. User i then perturbs the sampled value v_i into $\tilde{v}_i$ via the Piecewise Mechanism using a privacy budget of ϵ_3 depending on the perturbed key $\tilde{k}_i$. There are two cases. First, if the perturbed key remains the same as the original key, i.e., $\tilde{k}_i = k_i$, then the value v_i after perturbation can provide useful information for the data collector to estimate the value mean for key k_i. In this case, user i perturbs v_i into $\tilde{v}_i$ via the Piecewise Mechanism using a privacy budget of ϵ_3 as $\tilde{v}_i = PM(v_i, \epsilon_3)$. Second, if the perturbed key is different from the original key, i.e., $\tilde{k}_i \neq k_i$, then value v_i itself cannot provide any useful information to the data collector to estimate the value mean for key $\tilde{k}_i$. To prevent the perturbed value $\tilde{v}_i$ from negatively affecting the estimation of the value mean for key $\tilde{k}_i$, user i replaces v_i by 0 and perturbs 0 into $\tilde{v}_i$ via the Piecewise Mechanism as $\tilde{v}_i = PM(0, \epsilon_3)$. We summarize the perturbation of $\langle k_i, v_i \rangle$ into $\langle \tilde{k}_i, \tilde{v}_i \rangle$ in Alg. 1. Each user i then submits the perturbed key-value pair $\langle \tilde{k}_i, \tilde{v}_i \rangle$ to the data collector.

Algorithm 1: Joint Key-Value Perturbation

Input : $\langle k_i, v_i \rangle$, ϵ_2 and ϵ_3

Output: Perturbed key-value pair $\langle \tilde{k}_i, \tilde{v}_i \rangle$

Perturb key k_i into $\tilde{k}_i$ according to p.d.f.

$$\Pr[\tilde{k}_i|k_i] = \begin{cases} \frac{e^{\epsilon_2}}{e^{\epsilon_2}+d-1}, & \text{if } \tilde{k}_i = k_i, \\ \frac{1}{e^{\epsilon_2}+d-1}, & \forall \tilde{k}_i \in \mathcal{K} \setminus \{k_i\}. \end{cases}$$

Perturb value v_i into $\tilde{v}_i$ as follows.

$$\tilde{v}_i = \begin{cases} \tilde{v}_i = PM(v_i, \epsilon_3), & \text{if } \tilde{k}_i = k_i, \\ \tilde{v}_i = PM(0, \epsilon_3), & \text{otherwise.} \end{cases}$$

return $\langle \tilde{k}_i, \tilde{v}_i \rangle$;

On receiving the n perturbed key-value pairs $\langle \tilde{k}_1, \tilde{v}_1 \rangle, \ldots, \langle \tilde{k}_n, \tilde{v}_n \rangle$, the data collector estimates the key count and value mean for each key $k \in \mathcal{K}$. Specifically, for each key $k \in \mathcal{K}$, the data collector does the following.

First, the collector counts the number of received perturbed key-value pairs with key k as $\tilde{n}_k = \sum_{i=1}^{n} I(\tilde{k}_i = k)$, where $I(\cdot)$ denotes the indicator function. The collector aims to estimate n_k, which represents the number of key-value pairs whose original key is k and remains k after perturbation. To do this, the collector needs to calibrate $\tilde{n}_k$. This calibrated value accounts for two types of key-value pairs: those that originally had a key of k and remain k after perturbation, and those whose key changed to k as a result of perturbation. Since each k_i was perturbed into $\tilde{k}_i$ via GRR, n_k can be estimated as

$$\hat{n}_k = \frac{\tilde{n}_k - nq_2}{p_2 - q_2}, \tag{8}$$

where $p_2 = \frac{e^{\epsilon_2}}{e^{\epsilon_2}+d-1}$ and $q_2 = \frac{1}{e^{\epsilon_2}+d-1}$.

Next, the collector estimates the value mean for key k, which can be computed as the ratio between the total sum of the values associated with key k and n_k. The reason is that for any key-value pair whose original key was not k but became k after perturbation, their perturbed value is generated as $PM(0, \epsilon_3)$ with a mean of 0. It follows that

$$\mathbb{E}[\sum_{\{i|\tilde{k}_i=k\}} \tilde{v}_i] = \mathbb{E}[\sum_{\{i|k_i=k\}} v_i]. \tag{9}$$

Therefore, the data collector estimates the value mean of key k as

$$\hat{m}_k = \frac{\sum_{\{i|\tilde{k}_i=k\}} \tilde{v}_i}{\hat{n}_k p_2} . \tag{10}$$

In summary, the estimated frequency $\hat{f}_k$ and value mean $\hat{m}_k$ are given in Eq. (5) and Eq. (10), respectively.

5 THEORETICAL ANALYSIS

In this section, we analyze the privacy guarantee of Adaptive. We have the following theorem regarding the privacy guarantee of the proposed scheme.

Theorem 1. *With privacy budget $\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3$, the proposed scheme satisfies ϵ-LDP.*

Proof. We first show that the sampling and perturbation carried out in Phase 1 satisfies ϵ_1-LDP. Denote the conditional probability of Phase 1 producing a perturbed key k' given a key-value set S by $\Pr[k'|S]$. Consider any two possible key-value sets S_1 and S_2, respectively. Suppose the sampled key are k_1 and k_2 correspondingly, we have

$$\begin{aligned}\frac{\Pr[k'|S_1]}{\Pr[k'|S_2]} &= \frac{\sum_{k_1\in\mathcal{K}} \Pr[k_1|S_1]\cdot\Pr[k'|k_1]}{\sum_{k_2\in\mathcal{K}} \Pr[k_2|S_2]\cdot\Pr[k'|k_2]} \\ &\le \frac{\sum_{k_1\in\mathcal{K}} \Pr[k_1|S_1]\cdot p_1}{\sum_{k_2\in\mathcal{K}} \Pr[k_2|S_2]\cdot q_1} \\ &= \frac{p_1}{q_1} \\ &= e^{\epsilon_1} ,\end{aligned} \tag{11}$$

where we use the fact that $q_1 \le \Pr[k'|k] \le p_1$ for all $k, k' \in \mathcal{K}$ under the GRR and that $\sum_{k\in\mathcal{K}} \Pr[k|S] = 1$ for any S. Therefore, the sampling and perturbation carried out in Phase 1 satisfies ϵ_1-LDP.

Next, we show that the joint key-value perturbation in Phase 2 satisfies $\epsilon_2 + \epsilon_3$-LDP. Similar to the analysis in Eq. (11), for any two possible key-value sets S_1 and S_2 and any perturbed key $\tilde{k}$ generated from key perturbation in Phase 2, we have

$$\frac{\Pr[\tilde{k}|S_1]}{\Pr[\tilde{k}|S_2]} \le e^{\epsilon_2} . \tag{12}$$

Moreover, under Piecewise Mechanism [23], for any two values v_1 and $v_2 \in [-1, 1]$ and any perturbed value $\tilde{v} \in [-C, C]$, we have

$$\begin{aligned}\frac{\Pr[PM(v_1,\epsilon_3)=\tilde{v}]}{\Pr[PM(v_2,\epsilon_3)=\tilde{v}]} &\le \frac{p_3}{p_3/e^{\epsilon_3}} \\ &= e^{\epsilon_3} ,\end{aligned} \tag{13}$$

where $p_3 = \frac{\exp(\epsilon_3)-\exp(\epsilon_3/2)}{2\exp(\epsilon_3/2)+2}$ and we use the fact that $\frac{p_3}{e^{\epsilon_3}} \le f(PM(v,\epsilon_3)=\tilde{v}) \le p_3$ for any $v \in [-1, 1]$ and any $\tilde{v} \in [-C, C]$.

Furthermore, considering any two key-value sets S_1 and S_2 any perturbed key-value pair $\langle\tilde{k}, \tilde{v}\rangle$. Suppose the sampled key value are $\langle k_1, v_1\rangle$ and $\langle k_2, v_2\rangle$ correspondingly, we have

$$\begin{aligned}\frac{\Pr[\langle\tilde{k},\tilde{v}\rangle|S_1]}{\Pr[\langle\tilde{k},\tilde{v}\rangle|S_2]} &= \frac{\sum_{k_1\in\mathcal{K}} \Pr[k_1|K_1]\cdot\Pr[\tilde{k}|k_1]\cdot\Pr[\tilde{v}|v_1]}{\sum_{k_2\in\mathcal{K}} \Pr[k_2|K_2]\cdot\Pr[\tilde{k}|k_2]\cdot\Pr[\tilde{v}|v_2]} \\ &\le \frac{\sum_{k_1\in\mathcal{K}} \Pr[k_1|K_1]\cdot p_2\cdot p_3}{\sum_{k_2\in\mathcal{K}} \Pr[k_2|K_2]\cdot q_2\cdot q_3} \\ &= \frac{\sum_{k_1\in\mathcal{K}} \Pr[k_1|K_1]}{\sum_{k_2\in\mathcal{K}} \Pr[k_2|K_2]}\cdot e^{\epsilon_2}\cdot e^{\epsilon_3} \\ &= e^{\epsilon_2+\epsilon_3} .\end{aligned} \tag{14}$$

Therefore, the joint key-value perturbation in Phase 2 satisfies $\epsilon_2 + \epsilon_3$-LDP.

Finally, the composition property of LDP [7] indicates that the proposed scheme satisfies ϵ-LDP, where $\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3$. □

6 SIMULATION RESULTS

In this section, we evaluate the performance of Adaptive via detailed simulation studies.

6.1 Simulation Settings

We consider a system composed of $n = 20,000$ users each with a set of $l = 5$ key-value pairs. We set the key space as $\mathcal{K} = \{1, 2, \ldots, 20\}$. To demonstrate the estimation accuracy across keys with differing frequencies, we generate a synthetic dataset where the frequency of each key is linearly dependent on its key ID. More specifically, we express the key frequency ratio f_k as a linear function of the key id k, namely, $f_k - 1 = s(k-1)$. Here, $s \ge 0$ represents the variance in key frequency: the larger the s, the greater the discrepancy, and the converse holds true. For $s = 0$, the frequencies of all keys are identical.

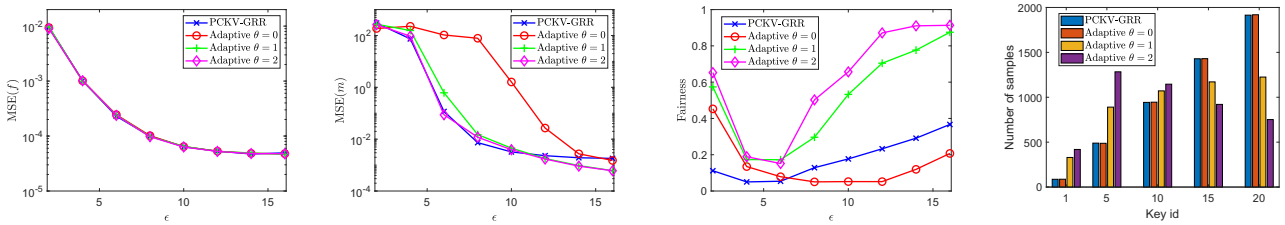
We compare Adaptive with the state-of-art solution PCKV-GRR [12]. As PCKV-GRR apportions the total privacy budget into two parts, one for key perturbation and the other for value perturbation, while Adaptive distributes the total privacy budget into three parts, we equalize the privacy budget allocated for key perturbation in both Adaptive and PCKV-GRR to ensure a fair comparison. Specifically, given a total privacy budget of ϵ we set $\epsilon_1 = \epsilon_2 + \epsilon_3 = \epsilon/2$. Furthermore, unless otherwise specified, we set $\epsilon_2 = \epsilon_3 = \epsilon/4$, $s = 1$, and $\theta = 2$.

We use averaged Mean Square Error (MSE) to measure the estimation accuracy for both key frequency and value mean across different keys. Specifically, we first define the variance of the estimated frequency for key $k \in \mathcal{K}$ as

$$\mathrm{Var}[f_k] = \frac{\sum_{j=1}^{N}(f_k^* - \hat{f}_k^j)^2}{N} \tag{15}$$

where f_k^* is the true frequency of key k, $\hat{f}_k^j$ is the estimated frequency of key k in the jth simulation run, and $N = 100$ is the total number of simulation runs. The averaged MSE of the estimated frequencies across all keys is then defined as

$$\mathrm{MSE}(f) = \frac{\sum_{k\in\mathcal{K}} \mathrm{Var}[f_k]}{|\mathcal{K}|} . \tag{16}$$



(a) MSE(f) vs. ϵ (b) MSE(m) vs. ϵ (c) Fairness vs. ϵ (d) Sampled Count

Figure 1: MSE and Fairness versus ϵ.

Similarly, we define the variance of the estimated value mean for key $k \in \mathcal{K}$ as

$$\text{Var}[m_k] = \frac{\sum_{j=1}^{N}(m_k^* - \hat{m}_k^j)^2}{N}, \tag{17}$$

where m_k^* is the true value mean and $\hat{m}_k^j$ is the estimated value mean of key k in the jth simulation run. The averaged MSE of the estimated value mean across all keys is then defined as

$$\text{MSE}(m) = \frac{\sum_{k\in\mathcal{K}}\text{Var}[m_k]}{|\mathcal{K}|}. \tag{18}$$

Moreover, we borrow Jain's fairness index [13] to evaluate the fairness of mean estimation accuracy across different keys, which is defined as

$$\text{Fairness} = \frac{(\sum_{k\in\mathcal{K}}\text{Var}[m_k])^2}{|\mathcal{K}|\cdot\sum_{k\in\mathcal{K}}(\text{Var}[m_k])^2}. \tag{19}$$

The fairness value ranges from $\frac{1}{|\mathcal{K}|}$ to 1. The larger the fairness value, the more uniform the mean estimation accuracy across different keys, and vice versa.

6.2 Simulation Results

We now report our simulation results.

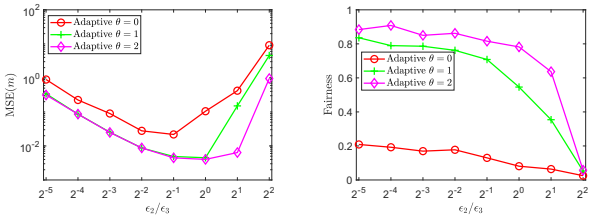
6.2.1 Impact of ϵ. Fig. 1 shows the accuracy and fairness of the Adaptive compared to PCKV-GRR, as the total privacy budget ϵ increases from 2 to 16 under $s = 1$. Fig. 1(a) shows a comparison between the average MSE of frequency estimation under PCKV-GRR and our Adaptive scheme across different θ values. All these methods share an identical MSE(f), as they all employ the GRR with half the total privacy budget dedicated to key perturbation. As anticipated, MSE(f) declines as privacy budget ϵ increases, i.e., the larger the privacy budget, the higher the probability of users reporting the original true key, thereby reducing the error.

The average MSE and Fairness of the mean estimations are shown in Figs. 1(b) and 1(c). As we can see, the MSE(m) for all methods decrease as ϵ increases. This can be attributed to the fact that with a larger ϵ, users are more likely to report the accurate value (i.e., less privacy protection), thereby enabling the collector to infer the mean value with greater precision. Additionally, although PCKV-GRR assigns $\epsilon/2$ to value perturbation as opposed to Adaptive's $\epsilon/4$, Adaptive (for example, $\theta = 1, 2$) still maintains a comparable MSE(m) to that of PCKV-GRR. This is primarily because Adaptive significantly enhances the estimation accuracy for infrequent keys while still delivering relatively high accuracy for frequent keys. Furthermore, Adaptive surpasses PCKV-GRR for larger ϵ values, e.g., $\epsilon > 12$. At this point, the number of samples becomes the primary factor influencing the estimation accuracy, given that users are nearly always reporting their true key-values with such an ample privacy budget. By contrast, Adaptive, at $\theta = 0$, records the highest MSE$_m$. This is because $\theta = 0$ signifies uniform sampling, which doesn't benefit from varying the number of samples but instead depletes the privacy budget for key perturbation in the second phase. As such, it has the lowest accuracy, even when compared to PCKV-GRR.

Fig. 1(c) shows a comparison of the fairness of mean estimation under various methods, with ϵ ranging from 2 to 16. As depicted, the fairness of all methods initially decreases and subsequently increases as ϵ increases. When ϵ is small, these methods have uniformly low estimation accuracy across all keys, leading to relatively high fairness. However, with a moderate privacy budget, for instance, $\epsilon \in [4, 6]$, keys with higher frequency exhibit superior estimation accuracy, while keys with lower frequency have markedly low accuracy, resulting in diminished fairness. As ϵ continues to rise, these methods achieve similar high estimation accuracy for all keys, reducing the accuracy differential among keys and enhancing fairness. In general, the fairness of Adaptive methods with $\theta = 1, 2$ exceeds that of PCKV-GRR. This can be attributed to the fact that the proposed Adaptive scheme secures more samples for infrequent keys and ample samples for frequent keys, ensuring fair estimation across different keys. Conversely, PCKV-GRR exhibits reduced accuracy for infrequent keys due to limited samples, resulting in lower fairness despite high estimation accuracy for frequent keys. The inferior fairness of Adaptive with $\theta = 0$ is owed to the squandering of the privacy budget on key perturbation in the second phase without reaping the benefits of adaptive sampling.

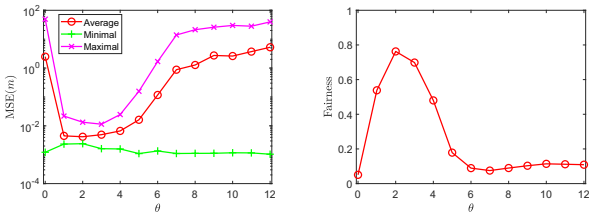
Fig. 1(d) exhibits the number of samples for different keys in the second phase. The sampled distributions of PCKV-GRR and Adaptive $\theta = 0$ mirror the original key frequency distribution, as expected with uniform sampling. More significantly, Adaptive $\theta = 1, 2$ gathers more samples for infrequent keys and fewer samples for highly frequent keys when juxtaposed with uniform sampling. This illustrates the merit of the proposed Adaptive method, obtaining a fair estimation for different keys by balancing the number of samples.

In summary, Fig. 1 underscores that the proposed Adaptive method with $\theta = 1$ or $\theta = 2$ surpasses the PCKV-GRR method



(a) MSE(m) vs. ϵ_2/ϵ_3 (b) Fairness vs. ϵ_2/ϵ_3

Figure 2: MSE and Fairness under budget allocation.



(a) MSE(m) vs. θ (b) Fairness vs. θ

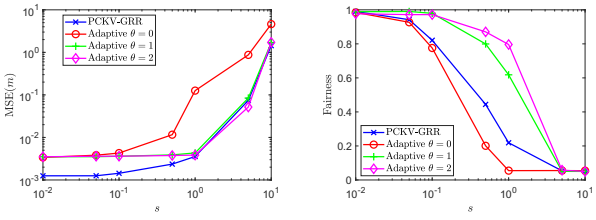
Figure 3: MSE and Fairness versus θ.

in terms of both accuracy and fairness for mean estimation. The Adaptive method mitigates the issue of low accuracy for infrequent keys by adjusting the sampling probabilities for different keys, suggesting that the Adaptive method is better suited for scenarios where the key frequency distribution is skewed.

6.2.2 Impact of Privacy Budget Allocation. The averaged MSE and fairness of mean estimation under different budget allocations, i.e., ϵ_2/ϵ_3, are depicted in Fig. 2. As we can see from Fig. 2(a), the MSE first decreases and subsequently increases as the ϵ_2/ϵ_3 elevates from 2^{-5} to 2^2. The accuracy of mean estimation in the second phase is dependent on both key and value. A smaller ϵ_2/ϵ_3 implies a reduced ϵ_2, which correlates with a low probability of users reporting their actual sampled keys in the second phase, thereby yielding low frequency estimate accuracy. Conversely, a larger ϵ_2/ϵ_3 suggests a smaller ϵ_3, in which case users tend to report perturbed values with considerable noise. Both scenarios result in a high MSE for mean estimation. Fig. 2(b) shows the fairness of mean estimation with varying ϵ_2/ϵ_3. We can see that the fairness initially decreases at a slow rate before plummeting sharply post $\epsilon_2/\epsilon_3 = 1$ and nearly bottoming out at 0. This implies that there exists a ϵ_2/ϵ_3, for instance, $\epsilon_2/\epsilon_3 = 1$, under which both high fairness and low estimate error on mean estimation can be achieved.

6.2.3 Impact of θ. In the second phase of our proposed Adaptive scheme, we introduced the parameter θ to control the degree of non-uniformity of random sampling. Fig. 3(a) shows the variation in mean squared error (MSE(m)) as θ increases from 0 to 12. We can see that both the average and maximum MSE(m) initially decrease and subsequently increase with rising θ. With small values of θ such as $\theta = 0$, the sample size for infrequent keys is minimal, leading to a high estimation error. On the other hand, for moderate values of θ, specifically, $\theta \in [1, 4]$, the sample size for infrequent keys increases while maintaining a relatively high sample size for frequent keys, thus reducing the estimation error. Beyond a certain point (e.g., $\theta > 6$), the sample size for frequent keys decreases, thereby increasing the estimation error. The minimal MSE(m) remains fairly constant as it is contingent upon the key with the most samples, which remains largely unaltered across different θ values.

Fig. 3(b) shows the estimation fairness in relation to θ. The fairness initially increases as the number of samples for infrequent keys rises, while the sample size for frequent keys remains high, yielding a fairly similar and high estimation accuracy across all keys. However, as θ continues to increase, the sample size for frequent keys diminishes and that of infrequent keys escalates, leading to a decline in fairness. Furthermore, the sampling probabilities do not significantly fluctuate under larger θ values, resulting in a consistent and reduced fairness.



(a) MSE(m) vs. s (b) Fairness vs. s

Figure 4: MSE and Fairness versus s.

6.2.4 Impact of s. Fig. 4 shows the accuracy and fairness of mean estimates under different datasets, distinguished by different key frequencies' discrepancy, s, ranging from 0.01 to 10. Notably, Fig. 4(a) shows an increase in the estimation error as s increases. This can be attributed to the fact that a larger s implies a greater disparity in key frequencies between high-frequency and infrequent keys. Given that the total number of key-value pairs remains fixed, this suggests that more users possess the high-frequency keys, thus yielding smaller estimation errors, while fewer users possess the infrequent keys, leading to larger estimation errors. However, the accuracy boost provided by the high-frequency keys is insufficient to offset this increase, resulting in a higher overall MSE(m).

Furthermore, we can see that the estimation error under the PCKV-GRR is lower than that under the proposed Adaptive method when $s < 1$. This is anticipated, as the Adaptive method is explicitly designed to address imbalances in estimation accuracy among keys with significantly different frequencies. Therefore, it does not offer a substantial advantage when the key frequencies are similar, especially at smaller s values. However, the proposed method achieves a comparable estimation accuracy to the PCKV-GRR method when $s \geq 1$. On the other hand, Adaptive $\theta = 0$ results in the highest MSE(m) because it relies on uniform sampling in the second phase without taking advantage of adaptive sampling.

Fig. 4(b) shows the fairness of the mean estimate with respect to varying s values. Initially, fairness starts close to 1, gradually decreases, and eventually stabilizes near 0. The initial high fairness results from the similar key frequencies at lower s values.

As s increases, the discrepancy between the frequencies of high-frequency and infrequent keys widens, resulting in reduced fairness. As s increases further (e.g., $s \geq 5$), estimating infrequent keys becomes virtually impossible, while the estimation accuracy for high-frequency keys remains high, yielding the lowest fairness. Importantly, the proposed Adaptive method exhibits greater fairness than the PCKV-GRR method for reasons previously discussed. By considering both Figs. 4(a) and 4(b), we can conclude that the proposed method not only achieves estimation accuracy comparable to the PCKV-GRR method but also enhances fairness across a broad spectrum of datasets with varying key frequency distributions.

6.3 Summary of Simulation Results

We summarize the simulation results as follows.

- Compared with PCKV-GRR, Adaptive achieves the same frequency estimation accuracy, superior accuracy and fairness for mean estimation under different privacy budgets.
- With an appropriate privacy budget allocation setting, e.g., $\epsilon_2/\epsilon_3 = 1$, Adaptive can achieve both high fairness and low estimation error on mean estimation.
- With a moderate value of θ, e.g., $\theta \in [1, 4]$, Adaptive can ensure more samples for infrequent keys while maintaining sufficient samples for frequent keys, resulting in a high accuracy and fairness of mean estimation.
- Adaptive outperforms PCKV-GRR for datasets with imbalance key frequency, specifically when $s \geq 1$.

7 CONCLUSIONS

In this paper, we have presented the design and evaluation of Adaptive, a novel locally differentially private and fair key-value aggregation scheme. By decoupling frequency estimation and mean estimation into two phases, Adaptive utilizes non-uniform random sampling to adjusting the sampling probabilities associated with different keys based on their estimated frequencies that ensures a greater number of samples for infrequent keys and fewer samples for those more frequently appearing. As a result, Adaptive achieves a consistently high level of accuracy in estimating the value mean across all keys. The efficacy of Adaptive has been substantiated through detailed theoretical analysis and simulation studies utilizing synthetic datasets.

ACKNOWLEDGMENTS

We would like to thank anonymous reviewers for their insightful comments that have helped improve the quality of this work. This work was supported in part by the US National Science Foundation under grants CNS-2325564 and CNS-1933047.

REFERENCES

[1] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. 2017. Practical Locally Private Heavy Hitters. In *NIPS'17*. Long Beach, CA, USA, 2285–2293.
[2] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting Telemetry Data Privately. In *NIPS'17*. Long Beach, California, USA, 3574–3583.
[3] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. 2013. Local Privacy and Minimax Bounds: Sharp Rates for Probability Estimation. In *NIPS'13*. Lake Tahoe, Nevada, 1529–1537.
[4] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. 2013. Local Privacy and Statistical Minimax Rates. In *FOCS '13*. Berkeley, CA, 429–438.
[5] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. 2014. Privacy Aware Learning. *J. ACM* 61, 6 (November 2014), 1–57.
[6] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *TAMC'08*. Xi'an, China, 1–19.
[7] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our Data, Ourselves: Privacy via Distributed Noise Generation. In *Proceedings of EUROCRYPT'06* St. Petersburg, Russia, 486-503.
[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC'06*. New York, NY, 265–284.
[9] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (August 2014), 211–407.
[10] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *CCS'14*. Scottsdale, AZ, 1054–1067.
[11] Giulia Fanti, Vasyl Pihur, and Úlfar Erlingsson. 2016. Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries. *Proceedings on PoPETS* 2016, 3 (2016), 43–51.
[12] Xiaolan Gu, Ming Li, Yueqiang Cheng, Li Xiong, and Yang Cao. 2020. PCKV: Locally Differentially Private Correlated Key-Value Data Collection with Optimized Utility. In *USENIX Security 20*. 967–984.
[13] Raj Jain, Dah Ming Chiu, and Hawe WR. 1998. A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems. *CoRR* cs.NI/9809099 (01 1998).
[14] Tejas Kulkarni, Graham Cormode, and Divesh Srivastava. 2018. Marginal Release Under Local Differential Privacy. In *SIGMOD'18*. Houston, TX, USA, 131–146.
[15] Zitao Li, Tianhao Wang, Milan Lopuhaä-Zwakenberg, Ninghui Li, and Boris Škoric. 2020. Estimating Numerical Distributions under Local Differential Privacy. In *SIGMOD '20*. Portland, OR, USA, 621–635.
[16] Thông T. Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. 2016. Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy. *CoRR* abs/1606.05053 (2016).
[17] Zhan Qin, Yin Yang, Ting Yu, Issa Khalil, Xiaokui Xiao, and Kui Ren. 2016. Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy. In *CCS '16*. Vienna, Austria, 192–203.
[18] Zhan Qin, Ting Yu, Yin Yang, Issa Khalil, Xiaokui Xiao, and Kui Ren. 2017. Generating Synthetic Decentralized Social Graphs with Local Differential Privacy. In *CCS '17*. Dallas, Texas, USA, 425–438.
[19] Lin Sun, Jun Zhao, Xiaojun Ye, Shuo Feng, Teng Wang, and Tao Bai. 2019. Conditional Analysis for Key-Value Data with Local Differential Privacy. *CoRR* abs/1907.05014 (2019).
[20] Abhradeep Guha Thakurta, Andrew H Vyrros, Umesh S Vaishampayan, Gaurav Kapoor, Julien Freudiger, Vivek Rangarajan Sridhar, and Doug Davidson. 2017. Learning new words. US Patent 9,594,741.
[21] Abhradeep Guha Thakurta, Andrew H Vyrros, Umesh S Vaishampayan, Gaurav Kapoor, Julien Freudinger, Vipul Ved Prakash, Arnaud Legendre, and Steven Duplinsky. 2017. Emoji frequency detection and deep link frequency. US Patent 9,705,908.
[22] Ning Wang, Xiaokui Xiao, Yin Yang, Ta Duy Hoang, Hyejin Shin, Junbum Shin, and Ge Yu. 2018. PrivTrie: Effective Frequent Term Discovery under Local Differential Privacy. In *ICDE'18*. Paris, France, 821–832.
[23] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. 2019. Collecting and analyzing multidimensional data with local differential privacy. In *ICDE'19*. Macau, China, 638–649.
[24] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally Differentially Private Protocols for Frequency Estimation. In *USENIX Security 17*. Vancouver, BC, 729–745.
[25] Tianhao Wang, Ninghui Li, and Somesh Jha. 2018. Locally Differentially Private Frequent Itemset Mining. In *SP'18*. San Francisco, CA, USA, 127–143.
[26] Tianhao Wang, Ninghui Li, and Somesh Jha. 2019. Locally differentially private heavy hitter identification. *IEEE Transactions on Dependable and Secure Computing* 18, 2 (2019), 982–993.
[27] SL Warner. 1965. Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 309 (March 1965), 63–66.
[28] Qingqing Ye, Haibo Hu, Xiaofeng Meng, and Huadi Zheng. 2019. PrivKV: Key-Value Data Collection with Local Differential Privacy. In *SP'19*. San Francisco, CA, 317–331.
[29] Qingqing Ye, Haibo Hu, Xiaofeng Meng, Huadi Zheng, Kai Huang, Chengfang Fang, and Jie Shi. 2021. PrivKVM*: Revisiting Key-Value Statistics Estimation with Local Differential Privacy. *IEEE Transactions on Dependable and Secure Computing* 20 (2021), 17–35. Issue 1.
[30] Zhikun Zhang, Tianhao Wang, Ninghui Li, Shibo He, and Jiming Chen. 2018. CALM: Consistent Adaptive Local Marginal for Marginal Release Under Local Differential Privacy. In *CCS '18*. Toronto, Canada, 212–229.