Expectation-Maximization Estimation for Key-Value Data Randomized with Local Differential Privacy

Hikaru Horigome, Hiroaki Kikuchi and Chia-Mu Yu

Abstract This paper studies the local differential privacy (LDP) algorithm for key-value data that are pervasive in big data analysis. One of the state-of-the-arts algorithms, PrivKV, randomizes key-value pairs with a sequence of LDP algorithms. However, maximum likelihood estimation fails to estimate the statistics accurately when the frequency of the data for particular rare keys is limited. To address the problem, we propose the expectation-maximization-based algorithm designed for PrivKV. Instead of estimating continuous values $[-1, 1]$ in key-value pairs, we focus on estimating the intermediate variable that contains the encoded binary bit $\in \{1, -1\}$. This makes the problem tractable to estimate because we have a small set of possible input values and a set of observed outputs. We conduct some experiments using some synthetic data with some known distributions, e.g., Gaussian and power-law and well-known open datasets, MoveLens and Clothing. Our experiment using synthetic data and open datasets shows the robustness of estimation with regards to the size of data and the privacy budgets.

1 Introduction

A key-value is a primitive data structure used for many applications and is pervasive in big data applications such as mobile app activity analysis. If we can collect daily usage data of smartphone apps, the data can be applied for optimizing battery management, personalized services, digital contents

Hikaru Horigome and Hiroaki Kikuchi
Meiji University, 4-21-1 Nakano, Nakano, Tokyo, 164-8525 Japan e-mail: cs212030@meiji.ac.jp, e-mail: kikn@meiji.ac.jp

Chia-Mu Yu
National Yang Ming Chiao Tung University, 1001 University Rd., Hsinchu, 300, Taiwan.

delivery and prediction diseases for healthcare. However, daily active usage is confidential data, and many people deny access to their personal data.

Local differential privacy (LDP) is a state-of-the-art private data anonymization mechanism. Erlingsson et al. at Google proposed a LDP algorithm [6]. It has been deployed by major platformers including Apple[7], Google[6], and Microsoft [8]. Variation of LDP schemes have been studied in order to expand the domain of LDP applications. Ye et al. [4] proposed the key-value data collection mechanisms that can satisfy LDP and estimate the key frequencies and the mean values from the sophisticatedly randomized key-value pairs. Their algorithm combines two LDP mechanisms known as primitives, the Randomized Response (RR) for keys [2] and the Value Perturbation Primitive (VPP) [3] for values. PCKV [14] is also proposed to collect the key-value pairs in an LDP manner.

The estimation accuracy is a one of current issues in LDP schemes. The size of users is also known as one of the factors to determine the estimation accuracy. The more perturbed key-value pairs, the more accurate the estimate. Most LDP algorithms, e.g., RAPPOR and PrivKV, estimate the statistics by solving the expected relationship between the observed randomized value and the true statistics. It is estimated *by means of a single point of expected value of randomized output*, which is a kind of Maximum Likelihood Estimation (MLE). Therefore, it suffers lower estimation accuracy when many values are randomized far from the theoretical expected value.

In order to address the lower estimation accuracy when the frequency is limited, we propose an iterative approach to improve the estimation accuracy of perturbed data in the LDP algorithm. Our idea is based on Bayes' theorem and the Expectation-Maximization (EM) algorithm[10]. The iterative process updates the posterior probabilities so that the all elements are consistent with the given observed data. Hence, it is more stable and more robust than the data that contain values for rear keys. However, it is not trivial to apply the EM algorithm to PrivKV because of its sequential combination of randomized key and continuous value $v \in [-1, 1]$. Instead of naïvely estimating v, we attempt to estimate the probability of intermediate values $v^+ \in \{1, -1\}$ in randomizing process of PrivKV. It makes the problem simple and tractable.

We conducted some experiments using synthetic data with some known distributions, e.g., Gaussian and Power-law. Then, we compared our proposed algorithm with PrivKV and PrivKVM (three iterations with responders)[4] to explore the accuracy improvement in terms of privacy budget ϵ and the number of responders n. We also evaluate the estimation accuracy using some open datasets, MovieLens and Clothing. It demonstrates that the proposed scheme performs well in general cases.

Our contribution are as follows.

- We propose a new algorithm to estimate the key frequencies and the mean values in key-value data that randomized in local differential private algorithm PrivKV (Section 3).

- The experimental results using synthetic data with three major probability distributions Gaussian, power-low and linear, and well-known open datasets, MovieLens and Clothing, demonstrate that our proposed algorithm overperforms state-of-the-art LDP schemes in estimation of frequency and mean value. Our experiment using synthetic data shows the robustness of estimation with regard to the size of data and the privacy budgets (Section 4).

Our paper is organized as follows. In Section 2, we provide some necessary fundamental definitions of LDP and the baseline estimation algorithms. In Section 3, we propose our algorithm and prove useful property in estimation of frequency and mean values. We report our experiments using synthetic and open data in Section 4. Our experiments show that the performance and the efficiency of the proposed algorithm in comparison to the existing LDP schemes. Section 5 shows some related works in this study. We conclude our study in Section 6.

2 Local Differential Privacy

2.1 Fundamental Definition

Suppose that users periodically submit their location data to a service provider. Differential privacy guarantees that the randomized data do not reveal any privacy disclosure from them. By contrast, LDP needs no trusted party. LDP is defined as follows.

Definition 1. A randomized algorithm Q satisfies ϵ-local differential privacy if for all pairs of values v and v' of domain V and for all subset S of range Z ($S \subset Z$), and for $\epsilon \geq 0$, $Pr[Q(v) \in S] \leq e^{\epsilon} Pr[Q(v') \in S]$.

2.2 PrivKV

Multidimensional data are one of the big challenges for perturbations. Several randomization mechanisms with LDP have been proposed so far.

Ye et al. [4] addressed the issue using two variables that are perturbed accordingly in their proposed LDP algorithm, PrivKV. PrivKV takes inputs in the form key-value data, two-dimensional data structure of discrete (key) and continuous (value) variables, and estimates the key frequencies and the mean values. Their idea combines two LDP protocols, RR for randomizing keys and VPP for perturbing values. The dimension is restricted to two, but the key-value is known as a primitive data structure commonly used for several applications. For example, a movie evaluation dataset consists of ratings for

movies, which are stored in a key-value database in which keys are movie titles and the values are ratings for titles. In a smartphone survey, users indicate their favorite apps such as $\langle \texttt{YouTube}, 0.5\rangle$, $\langle \texttt{Twitter}, 0.1\rangle$, $\langle \texttt{Instagram}, 0.2\rangle$, by stating their total time using those apps.

Let S_i be a set of key-value tuple $\langle k, v\rangle$ owned by i-th user. In PrivKV, the set of tuples is encoded as a d-dimensional vector, where d is the cardinality of the domain of keys K and missing key is represented as $\langle k, v\rangle = \langle 0, 0\rangle$. For instance, a set of key-value $S_i = \{\langle k_1, v_1\rangle, \langle k_4, v_4\rangle, \langle k_5, v_5\rangle\}$ is encoded as $d = 5$ dimensional vector $\boldsymbol{S}_i = (\langle 1, v_1\rangle, \langle 0, 0\rangle, \langle 0, 0\rangle, \langle 1, v_4\rangle, \langle 1, v_5\rangle)$ where keys k_1, k_4 and k_5 are specified implicitly with 1 at the corresponding location.

Perturbation in PrivKV is performed by random sampling one element $\langle k_a, v_a\rangle$ from $\boldsymbol{S}_i$. It has two proceeding steps, perturbing values and keys. It uses VPP used in Harmony[3] for the chosen tuple. A value of the tuple $\langle 1, v_a\rangle$ is replaced by ${v^+}_a = VPP(v_a, \epsilon_2)$, where ϵ_2 is a privacy budget for values. A value of the "missing" tuple $\langle 0, 0\rangle$ is replaced by ${v^+}_a = VPP(v'_a, \epsilon_2)$, where v'_a is chosen uniformly from $[-1, 1]$.

It uses RR with privacy budget ϵ_1. A tuple $\langle 1, v_a\rangle$ is randomized as

$$\langle k^*_a, v^+_a\rangle = \begin{cases} \langle 1, v^+_a\rangle & w/p\ p_1 = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}, \\ \langle 0, 0\rangle & w/p\ q_1 = \frac{1}{1+e^{\epsilon_1}}, \end{cases}$$

where v^+_a is perturbed as mentioned. A "missing" tuple $\langle 0, 0\rangle$ is randomized as

$$\langle k^*_a, v^+_a\rangle = \begin{cases} \langle 0, 0\rangle & w/p\ p_1 = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}, \\ \langle 1, v^+_a\rangle & w/p\ q_1 = \frac{1}{1+e^{\epsilon_1}}. \end{cases}$$

Responder in PrivKV submits the perturbed tuple $\langle k^*_a, v^+_a\rangle$ with the index a of the tuple.

3 Proposed Algorithm

3.1 Idea

The drawback of PrivKV and PrivKVM is their low estimation accuracy. Because PrivKV uses the MLE of frequencies and means, the estimate accuracy reduces when the values are not uniformly distributed or sparse data are given. An iterative approach like PrivKVM consumes a privacy budget for every iteration, and the optimal assignment is not trivial.

MLE used in PrivKV works well for some cases but has low estimate accuracy for a biased distribution. Instead, we attempt to address this limitation by using an iterative estimate approach known as the EM (EM) algorithm. Because EM estimates posterior probabilities iteratively so that the estimated probabilities are more consistent with all observed values, it can

improve accuracy when the number of users n increases and many observed data are given. However, it is challenging to estimate exact continuous value $v \in [-1, 1]$. Instead of estimating v directly, we focus to estimate the encoded v in binary, $v^+ \in \{1, -1\}$. It is tractable to estimate because we have a small set of possible value of input as $X = \{\langle 1, 1\rangle, \langle 1, -1\rangle, \langle 0, 1\rangle, \langle 0, -1\rangle\}$ and a set of observed output $Z = \{\langle 1, 1\rangle, \langle 1, -1\rangle, \langle 0, 0\rangle\}$. Estimated marginal probability of X allows the mean values to be estimated accurately.

3.2 EM Algorithm for PrivKV

EM algorithm performs an iterative process for which posterior probabilities are updated through Bayes' theorem[10]. Each iteration estimates the best probabilities θ^j for all possible values in a domain. First, we show the EM algorithm generally and then modify it for PrivKV.

Let $X = \{x_1, \ldots, x_d\}$ be a set of input values and $Z = \{z_1, z_2, \ldots, z_{d'}\}$ a set of output values. A responder owning private value $x_i \in X$ uses a randomized algorithm to output $z_i \in Z$. Given n observed values $z_1, \ldots, z_n$, we iterate estimating posterior probabilities for $x_1, \ldots, x_d$ as $\Theta^{(t)} = ({\theta_1}^{(t)}, {\theta_2}^{(t)}, \ldots, {\theta_d}^{(t)})$ until converged. We start iteration with the initialized values assigned to probabilities uniformly as $\Theta^{(0)} = (\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d})$.

The conditional probability of input x_i given output z_j is given as $Pr[z_j|x_i] = \frac{Pr[z_j, x_i]}{Pr[x_i]}$. Bayes' theorem gives the posterior probability of $X = x_i$ given z_j as $Pr[x_i|z_j] = \frac{Pr[z_j|x_i]Pr[x_i]}{\sum_{s=1}^{|X|} Pr[z_j|x_s]Pr[x_s]}$. By letting $\theta_i^{(t-1)} = Pr[x_i]$ be the $(t-1)$-th estimate of marginal probability of $x_i \in X$, we have the t-th estimate of conditional probability for the u-th responder who responds $z_u \in Z$ as

$$\hat{\theta}_{u,i}^{(t)} = Pr[x_i|z_u] = \frac{Pr[z_u|x_i]{\theta_i}^{(t-1)}}{\sum_{s=1}^{|X|} Pr[z_u|x_s]{\theta_s}^{(t-1)}}, \tag{1}$$

which follows the t-th estimate of marginal probability by aggregating all n estimates of responders as, $\theta^{(t)} = \frac{1}{n}\sum_{u=1}^{n} \hat{\theta}_u^{(t-1)}$. This process iterates until updating converges as $|{\theta_i}^{(t)} - {\theta_i}^{(t-1)}| \leq \eta$, where η is predetermined precision.

In PrivKV, a randomization of tuple $\langle k, v\rangle$ is performed in sequential algorithms. A value $v \in [-1, 1]$ is encoded into $v^* \in \{-1, 1\}$ in a probability depending on v. Then, it is randomized as v^+ in $RR(v^*, \epsilon_2)$ (a part of VPP) using probabilities $p_2 = (e^{\epsilon_2})/(1 + \epsilon_2)$, and $q_2 = 1/(1 + \epsilon_2) = 1 - p_2$. Finally, it is randomized in $RR(v^*, \epsilon_1)$ using probabilities $p_1 = (e^{\epsilon_1})/(1 + \epsilon_1)$, and $q_1 = 1/(1+\epsilon_1) = 1-p_1$, as a part of key randomization. Hence, if we perturb a given tuple $\langle k, v^+\rangle = \langle 1, 1\rangle$ in PrivKV, the output $\langle k^*, v^*\rangle = \langle 1, 1\rangle$ is observed with probability $p_1 p_2$ as the consequence of VPP and RR. Similarly, another tuple happens

Table 1 Conditional probabilities of observed tuple Z given tuple X

$X = \langle k', v^+\rangle$		$Z = \langle k^*, v^*\rangle$		$Pr[z\|x]$	$X = \langle k', v^+\rangle$		$Z = \langle k^*, v^*\rangle$		$Pr[z\|x]$
1	1	1	1	p_1p_2	0	1	1	1	q_1p_2
1	1	1	−1	p_1q_2	0	1	1	−1	q_1q_2
1	1	0	0	$q_1(p_2+q_2)$	0	1	0	0	$p_1(p_2+q_2)$
1	−1	1	1	p_1q_2	0	−1	1	1	q_1q_2
1	−1	1	−1	p_1p_2	0	−1	1	−1	q_1p_2
1	−1	0	0	$q_1(p_2+q_2)$	0	−1	0	0	$p_1(p_2+q_2)$

$$\langle k^*, v^*\rangle = \begin{cases} \langle 1,1\rangle & \text{w/p } p_1p_2, \\ \langle 1,-1\rangle & \text{w/p } p_1q_2, \\ \langle 0,0\rangle & \text{w/p } q_1(p_2+q_2). \end{cases}$$

Thus, we have the conditional probability $Pr[z_1 = \langle k^*, v^*\rangle = \langle 1,1\rangle | x_1 = \langle k', v^+\rangle = \langle 1,1\rangle]$ is p_1q_1. Other conditional probabilities are given in Table 1.

Using these probabilities with Bayes' theorem, we have the posterior probability of input variable x_1 being $\langle 1,1\rangle$ given observed z_1 as follows:

$$\begin{aligned} Pr[x_1|z_1] &= \frac{Pr[z_1|x_1]Pr[x_1]}{\sum_{s=1}^{4} Pr[z_1|x_s]Pr[x_s]} = \frac{Pr[z_1|x_1]{\theta_1}^{(0)}}{\sum_{s=1}^{4} Pr[z_1|x_s]{\theta_s}^{(0)}} \\ &= \frac{\frac{1}{4}p_1p_2}{\frac{1}{4}p_1p_2 + \frac{1}{4}p_1q_2 + \frac{1}{4}q_1p_2 + \frac{1}{4}q_1q_2} = p_1p_2 = \frac{e^{\epsilon_1}e^{\epsilon_2}}{(1+e^{\epsilon_1})(1+e^{\epsilon_2})}. \end{aligned}$$

With privacy budgets $\epsilon_1 = \epsilon_2 = 1/2$ and $\epsilon = \epsilon_1 + \epsilon_2 = 1$, we estimate $\hat{\theta}^{(1)}_{1,u} \approx 0.387455$. Posterior probabilities for input $x_2 = \langle 1,-1\rangle, x_3 = \langle 0,1\rangle, x_4 = \langle 0,-1\rangle$ can be computed similarly.

3.3 Frequency and Mean Estimation

After the EM algorithm estimates the marginal probabilities for binary vector v_a^+, we need to identify the key frequency and mean values in the original key-value data. To estimate these quantities, we show the following property.

Theorem 1 (frequency and mean). *Let* $\Theta^{(t)} = (\theta^{(t)}_{\langle 1,1\rangle}, \theta^{(t)}_{\langle 1,-1\rangle}, \theta^{(t)}_{\langle 0,1\rangle}, \theta^{(t)}_{\langle 0,-1\rangle})$ *be marginal probabilities of the binary-encoded tuples* $\langle k_a^*, v_a^+\rangle$ *in PrivKV. Then, the expected values for the frequency for key* k_a *and the mean values are*

$$\hat{f}_a = n\left(\theta^{(t)}_{\langle 1,1\rangle} + \theta^{(t)}_{\langle 1,-1\rangle}\right), \hat{m}_a = \frac{\theta^{(t)}_{\langle 1,1\rangle} - \theta^{(t)}_{\langle 1,-1\rangle}}{\theta^{(t)}_{\langle 1,1\rangle} + \theta^{(t)}_{\langle 1,-1\rangle}}. \tag{2}$$

Algorithm 1 EM algorithm for PrivKV

$S_1, \ldots, S_n \leftarrow$ key-value data for n responders.
for all $u \in [n]$ **do** sample a tuple $\langle k'_a, v'_a \rangle$ from a vector $\boldsymbol{S}_i$
  $v_a^+ \leftarrow VPP(v'_a, \epsilon_2)$ and $k_a^* \leftarrow RR(k'_a, \epsilon_1)$
end for
$\Theta^{(0)} \leftarrow$ a uniform probability for $X = \{\langle 1,1\rangle, \langle 1,-1\rangle, \langle 0,1\rangle, \langle 0,-1\rangle\}$.
repeat(E-step)
  $t \leftarrow 1$
  Estimate posterior probability $Pr[V_i = 1|Z_i]$ in Eq. (1).
  (M-step) Update marginal probability $\theta_i^{(t+1)}$.
until $|\theta_i^{(t+1)} - \theta_i^{(t)}| \leq \eta$
for all $a \in K$ **do** estimate
  Estimate $\hat{f}_a$ and $\hat{m}_a$ in eq. 2.
end for return $\hat{f}_1, \hat{m}_1, \ldots, \hat{f}_d, \hat{m}_d$

If we have an accurate estimation of marginal probabilities via the EM algorithm, the theorem means that we can estimate the frequency and the mean as well.

Algorithm 1 shows the overall processes in the proposed EM algorithm for estimating frequency and mean for key-value data.

4 Experiment

4.1 Objective

The objective of the experiment is to explore the accuracy improvement in terms of privacy budget ϵ and the number of responders n. Using synthetic data with some common distributions, we compare our proposed algorithm with some conventional ones.

4.2 Data

We use some synthetic data and open datasets for our analysis. For synthetic data, we generate keys and values according to three known probability distributions, Gaussian ($\mu = 0, \sigma = 10$) power-law ($F(x) = (1 + 0.1x)^{-\frac{11}{10}}$), and linear ($F(x) = x$). Table 2 shows the mean and the variance of frequency f_k for key and mean f_k for values, where the number of users is $n = 10^5$.

Table 3 shows the specifications of open datasets used for our analysis. Both datasets have a large number of items, e.g., movie titles and clothe brands. Hence, the use-item matrices are sparse. The values of ratings are

Table 2 Statistics of the synthetic data (# responders $n = 10^5$, and # keys d)

model	$E(f_k/n)$	$Var(f_k/n)$	$E(m_k)$	$Var(m_k)$
Gaussian	0.49506	0.10926	−0.00987	0.43702
Power-low	0.20660	0.062901	−0.58681	0.25160
Linear	0.51	0.08330	0	0.34694

Table 3 Open datasets

item	MoveiLens[15]	Clothing[16]
# ratings	10,000,054	192,544
# users	69,877	9,657
# items	10,677	3,183
value range	0.5 – 5	1 – 10

Table 4 $MSE(f)[\times 10^{-4}]$ with regard to ϵ

	Gauss			Power-Law			Linear		
ϵ	EM	PrivKV	PrivKVM	EM	PrivKV	PrivKVM	EM	PrivKV	PrivKVM
0.1	756.682	1921.743	1472.772	671.251	2170.253	1851.214	602.837	1885.284	1462.740
0.5	63.996	84.629	75.394	55.478	84.833	62.403	70.346	92.988	82.795
1	18.076	22.588	26.213	18.579	19.274	23.183	16.023	20.174	18.440
3	2.018	2.324	2.508	1.591	2.587	2.420	2.523	2.790	2.597
5	1.147	1.320	1.173	0.973	1.019	0.992	1.283	1.429	1.280

distribute normally and the frequency of items follows power-law distributions. Therefore, the synthetic data are models of the real open data.

4.3 Method

We perform the proposed and the conventional algorithms PrivKV and PrivKVM to estimate the key frequency $\hat{f}_k$ and the mean values $\hat{m}_k$ of given n-responder synthetic data. The mean errors for key and value are evaluated by Mean Square Error (MSE) defined as $MSE(f) = \frac{1}{|K|}\sum_{i=1}^{|K|}\left(\frac{\hat{f}_i}{n} - \frac{f_i}{n}\right)^2$, $MSE(m) = \frac{1}{|K|}\sum_{i=1}^{|K|}\left(\hat{m}_i - m_i\right)^2$, where f_i and m_i are true statistics. We repeat the measurements for 10 times and take the mean.

4.4 Results

4.4.1 Privacy Budget ϵ

Table 4 shows the MSE of frequency estimation of synthetic data generated in Gaussian, Power-low and Linear distributions. We use the MSE with regard to the privacy budget ϵ ranging from 0.1 to 5, the number of responders $n = 10^5$, and the number of keys $d = 50$.

The estimation accuracy of the proposed EM algorithm overwhelms the conventional PrivKV and PrivKVM for all distributions and all privacy budgets. The improvement is significant for $\epsilon = 0.1$, and the MSE of EM algorithm is $602.83 \cdots 10^{-4}$ (41% of PrivKVM).

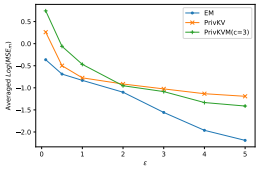

Fig. 1 $MAE(m)$ with regard to ϵ (Gaussian)

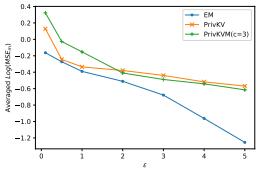

Fig. 2 $MAE(m)$ with regard to ϵ (Power-Law)

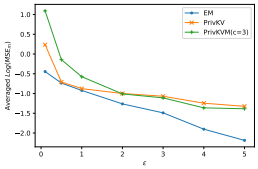

Fig. 3 $MAE(m)$ with regard to ϵ (Linear)

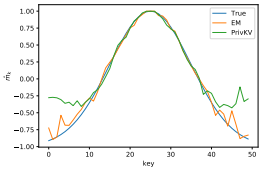

Fig. 4 Estimated distributions ($\epsilon = 4$, Gaussian)

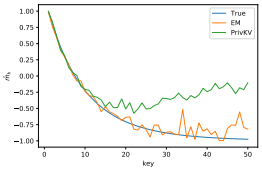

Fig. 5 Estimated distributions ($\epsilon = 4$, Power-law)

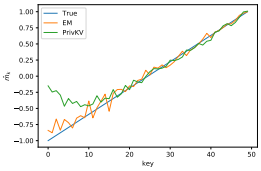

Fig. 6 Estimated distributions ($\epsilon = 4$, Linear)

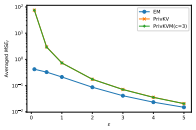

Fig. 7 $MAE(f_{movie})$ with regard to privacy budget (MoveLens)

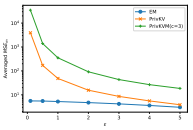

Fig. 8 $MAE(m_{raring})$ with regard to privacy budget (MoveLens)

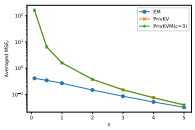

Fig. 9 $MAE(f_{cloth})$ with regard to privacy budget (Clothing)

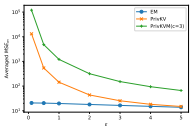

Fig. 10 $MAE(m_{rating})$ with regard to privacy budget (Clothing)

Figures 1, 2, and 3 show the MSE of mean estimations of synthetic data generated in Gaussian, Power-low, and Linear distributions, respectively. The proposed EM algorithm (blue) has the smallest MSE for all algorithms and all cases. The accuracy is improved well as the privacy budget ϵ increases.

We show the estimated mean distributions for keys synthesized in Gaussian, Power-law and Linear distribution, in Fig. 4, 5, and 6, respectively. We find that PrivKV suffers large estimation errors for both edges in Fig. 4, where the frequencies of values are less than that of the center. MLE is not robust when not enough samples are given, whereas EM performs well for even small samples by iterative processes.

Figures 7 and 8 shows the distribution of MAE of frequency of items and that of mean values in the MovieLens dataset[15], respectively. Similarly, the MAEs of frequency and mean values in the Clothing datasets[16] are in Figures 9 and 10, respectively.

Table 5 $MSE(f)$ $[(\times 10^{-4})]$ with regard to n

	Gauss			Power-Law			Linear		
$n[10^4]$	EM	PrivKV	PrivKVM	EM	PrivKV	PrivKVM	EM	PrivKV	PrivKVM
1	476.259	527.912	612.903	346.709	543.281	424.728	404.467	538.943	733.468
5	107.263	110.045	137.269	72.823	99.442	95.245	89.922	118.344	113.040
10	36.087	51.430	62.515	39.235	51.116	62.976	54.166	69.999	56.959
100	4.636	4.766	6.249	4.876	5.498	5.254	5.619	7.404	6.158
1000	1.410	1.597	1.556	0.760	1.181	1.094	0.974	1.493	1.041

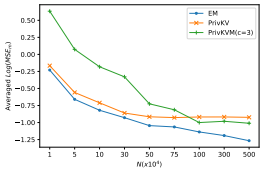

Fig. 11 $MAE(m)$ with regard to number of responders (Gaussian)

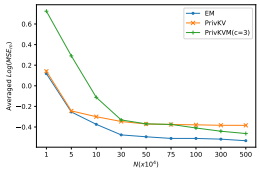

Fig. 12 $MAE(m)$ with regard to number of responders (Power-law)

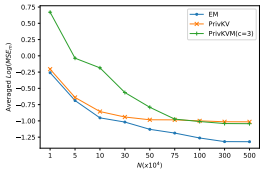

Fig. 13 $MAE(m)$ with regard to number of responders (Linear)

All results show that the proposed ME algorithm estimate the frequencies and the means the most accurately. The improvements of accuracy by the EM estimation are consistent with the results of the synthesized data.

4.4.2 Number of Responders

We evaluate the estimation error with regard to the number of responders (users). We estimate with the fixed privacy budget $\epsilon = 2$ and the number of keys $d = 50$ for the number of users from 10^4 to $30 \cdot 10^4$.

Table 5 shows the MSE for frequency estimation of synthetic data generated in Gaussian, Power-law, and Linear distributions. The EM algorithm has the smallest MSE for all algorithms and all distributions regardless of the number of users. The larger, the less error in general. The estimate improvement is significant when smaller data such as $n = 10^4$ are used. With the result, the EM should be used for the use case where a confidential and rare data are sampled, such as epidemiological study of rare diseases.

Figs. 11, 12, and 13 show the distribution of MAE for means with regard to the number of responders n, of the synthetic data generated in Gaussian, Power-law, and Linear distributions, respectively. Similar to the frequency estimation, the EM algorithm always outperforms than any other algorithm for all estimations and distributions. For example, the accuracy improves 31.6% in the Gaussian distribution where there are many less-frequent keys.

5 Related Works

The idea to preserve the privacy of input with randomization has been studied so far. Agrawal and Srikant [13] proposed a privacy-preserving collaboration filtering and an estimation algorithm based on Bayes' theorem, called reconstruction.

Chen et al. [12] proposed the notion of LDP to provide a privacy guarantee for the user. Compared with the conventional differential privacy studies, LDP has been used for many real-world applications. For example, Erlingsson et al. introduced RAPPOR [6] to use a Bloom filter to encode input as a bit of a vector.

Ren et al. proposed a multidimensional joint distribution estimation algorithm that satisfies LDP[11]. Their proposed method is also based on the EM and Lasso regression. They reported the experimental results on real-world datasets and showed that the proposed algorithm outperforms the existing estimation schemes such as support vector machine and random forest classifications.

Gu et al. [14] proposed a locally differentially private key-value data collection that utilizes correlated perturbation. Their protocol uses an advanced Padding-and-Sampling with two primitives, PCKV-UE (Unary Encoding) and PCKV-GRR (Generalized Randomized Response) to improve the accuracy of mean estimation and does not require further interaction between responders and collector.

6 Conclusion

We study the LDP algorithms for key-value data that estimate the key frequencies and the mean values. We propose an algorithm based on the EM algorithms to improve the estimation data accuracy perturbed in the LDP algorithms PrivKV and PrivKVM. Our proposed algorithm estimates the marginal probability of variable X that is a tuple of binary-encoded keys and values and hence can reduce the conditional probabilities needed for iterative processes. With some synthetic data generated in Gaussian, Power-low, and Linear distributions, we conduct experiments that show the proposed estimation has higher accuracy than the PrivKV algorithm. The estimate is robust for privacy budgets. The improvement was 69.5% on average when the number of responders was $n = 10^4$ with $\epsilon = 0.1$.

Major open datasets, MovieLens (10,000,000 records) and Clothing (192,000 records), are used to ensure the performance of the propose algorithm as estimated in the synthetic data. The experimental results confirm that the proposed EM algorithm outperforms any of state-of-the-art LDP schemes, PrivKV and PrivKVM, for all privacy budgets and for both of key frequencies and mean values. The improvement in estimation is especially significant

when small privacy budget is used for randomization (strong privacy level). The EM algorithm estimates the mean well even when the frequency of the key-value record is limited. Hence, we conclude that the proposed EM algorithm is appropriate for private data analysis in epidemiological purposes that requires dealing with rare decease.

For future works, we plan to compare the utility improvement with PCKV[14], the latest version of PrivKV family which improves accuracy without expensive iterations incurred by PrivKVM.

Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number JP18H04099 and JST, CREST Grant Number JPMJCR21M1, Japan.

References

1. P. Kairouz, S. Oh, and P. Viswanat, "Extremal mechanisms for local differential privacy", NIPS, pp. 2879-2887, 2014.
2. S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias", Journal of the American Statistical Association, pp. 63-69, 1965..
3. T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, J. Shin, "Collecting and analyzing data from smart device users with local differential privacy", arXiv:1606.05053, 2016.
4. Q. Ye, H. Hu, X. Meng, H. Zheng, "PrivKV : Key-Value Data Collection with Local Differential Privacy", IEEE S&P, pp. 294-308, 2019.
5. F. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis", SIGMOD, pp. 19-30, 2009.
6. Úlfar Erlingsson, Vasyl Pihur, Aleksandra Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response", ACM Conference on Computer and Communications Security, pp.1054-1067, 2014.
7. Differential Privacy Team, "Learning with privacy at scale", Apple Machine Learning Journal, 1(8), 2017.
8. "Learning with Privacy at Scale". https://machinelearning.apple.com/ (accessed on 2019).
9. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy", Found. Trends Theor. Comput. Sci. 9, 3-4, 211–407, 2014.
10. M. Miyagawa, "EM algorithm and marginal applications", Advanced Statistics, Vol. 16, No. 1, pp. 1-19, 1987 (in Japanese).
11. Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A. McCann, and Philip S. Yu, "LoPub : High-Dimensional Crowdsourced Data Publication With Local Differential Privacy," IEEE Transactions on Information Forensics and Security, vol. 13, no. 9, pp. 2151-2166, Sept. 2018.
12. R. Chen, H. Li, A. K. Qin, S. P. Kasiviswanathan, and H. Jin, "Private spatial data aggregation in the local setting," in Proc. IEEE ICDE, pp. 289–300, 2016.
13. R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining", ACM SIGMOD 2000, pp. 439-450, 2000.

14. Xiaolan Gu, Ming Li, Yueqiang Cheng, Li Xiong and Yang Cao, "PCKV: Locally Differentially Private Correlated Key-Value Data Collection with Optimized Utility", 29th USENIX Security Symposium (USENIX Security 20), pp. 967–984, 2020.
15. MovieLense 10M Dataset, https://grouplens.org/datasets/movielens/ (accessed in 2022).
16. Clothing Fit Dataset for Size Recommendation
17. X. Ren et al., "LoPub : High-Dimensional Crowdsourced Data Publication With Local Differential Privacy," in IEEE Transactions on Information Forensics and Security, vol. 13, no. 9, pp. 2151-2166, Sept. 2018, doi: 10.1109/TIFS.2018.2812146.