Multiple Privacy Regimes Mechanism for Local Differential Privacy

Yutong Ye[1,3], Min Zhang[1,2], Dengguo Feng[2(✉)], Hao Li[1], and Jialin Chi[1]

[1] Trusted Computing and Information Assurance Laboratory, Institute of Software,
Chinese Academy of Sciences, Beijing, China
{yeyutong,mzhang,lihao,chijialin}@is.iscas.ac.cn

[2] State Key Laboratory of Computer Science, Institute of Software,
Chinese Academy of Sciences, Beijing, China
feng@is.iscas.ac.cn

[3] University of Chinese Academy of Sciences, Beijing, China

Abstract. Local differential privacy (LDP), as a state-of-the-art privacy notion, enables users to share protected data safely while the private real data never leaves user's device. The privacy regime is one of the critical parameters balancing between the correctness of the statistical result and the level of user's privacy. In the majority of current work, authors assume that the privacy regime is totally determined by the service provider and dispatched to all users. However, it is inelegant and unpromising for all users to accept the same privacy level in real world. In this paper, we propose a new LDP estimation method MLE which is applicable for the scenario of multiple privacy regimes. MLE uses the idea of parameter estimation to merge the results generated by users of different privacy levels. We also propose an extension of MLE to handle the situation when all users' regimes are in a continuous distribution. We also provide an Adapt estimator which assigns users to use different LDP schemes based on their regimes, and it performs better than the estimator with only one fixed LDP scheme. Experiments show that our methods provide a higher level of accuracy than previous proposals in this multiple regimes scenario.

Keywords: Local differential privacy · Multiple privacy regimes · Frequency estimation

1 Introduction

With the rapid penetration of Internet and Smartphone through the crowded, large-scale collection of user data is already a necessary daily activity for companies. User data has become one of the most important asset, which can give support to data scientists to discover new patterns and provide training examples for machine learning models. However, this comes with huge risks–can these companies protect users' sensitive data from malicious access? Disclosure may violate the users' privacy and lead to scandal.

G. Li et al. (Eds.): DASFAA 2019, LNCS 11447, pp. 247–263, 2019.
https://doi.org/10.1007/978-3-030-18579-4_15

Anonymization techniques are one common method to protect user privacy by blurring the personalized or identifiable information, but it's vulnerable to the de-anonymization attack as shown in the case of Netflix Price [15]. For the above scenario, Differential Privacy [6,7] successfully achieved releasing sanitized datasets, but not at the client level.

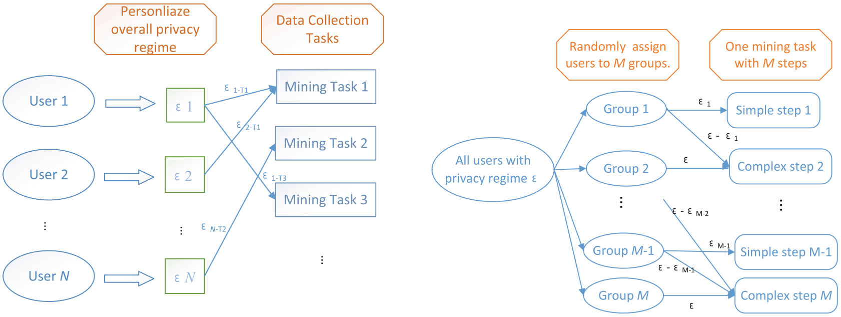
Local Differential Privacy (LDP) [13] is a branch of DP, which gives the benefits of population-level statistics without the collection of raw private data. With LDP, service provider can get statistical information on all users by just aggregating users' noised report. This feature makes LDP widely used in real-world scenarios. For example, Google's use LDP scheme RAPPOR to constantly collect the home-page that all users like to set up; Apple announced its implementation of LDP in iOS 10 and MacOS in WWDC 2016; Microsoft also deploys a LDP-enabled data collection mechanism in Windows Insiders program to collect application usage statistics.

LDP's security parameter (privacy regime ϵ) represents the security level of its randomization process. The bigger (or smaller) the security parameter, the more (or less) availability of the noisy report that the user shares. However, most LDP schemes assume that each user has the same ϵ, hence each user uses the exact same randomized procedure to generate a noisy report from their own data. There have been complaints that deployed LDP schemes use higher values of ϵ while users are not given any choice. So it is questionable that ϵ is entirely determined by the service provider who wants more availability of the data.

Multiple and Personalized Privacy Regimes. In order to meet the privacy demands of different people, we argue that users should be allowed to set their overall privacy levels (e.g., low/moderate/high) independently. Here, we assume that this personalization of privacy regimes does not mean the user should set a new ϵ every time he shares data. Instead, users will set an infrequently changing ϵ, which will be consumed a fixed percentage every time users share data with LDP. This assumption is derived from a study [17], which suggests that Apple's deployment [2] for LDP has an overall privacy regime as high as 16 everyday and sets privacy consumption to 1 or 2 each time shares data while there is no transparency.

In this paper, we consider that simple LDP mining task only contains one data collection activity, and the common simple mining task includes frequency estimation, mean value estimation, heavy-hitters identification and so on. So even if users have the same overall privacy regime ϵ, multiple ϵ may still appear in one mining task because the number of times that data is shared is different. For example, it could involve the real-time sharing of trajectory data and is not the focus of our analysis. We present the above personalization process in Fig. 1(a).

In addition, when LDP handles complex mining task like "Frequent Itemset Mining" [20] which contains several rounds of data collection activities, it is common to randomly assign all users to several groups, and then different groups finish each step of the mining task respectively with the same ϵ. However, some groups of users only need to pay a small amount of privacy to complete easy step (e.g., frequency estimation in small candidate space), the remaining can be assigned to challenging steps by segmentation as shown in Fig. 1(b). Combining the above two points, it's urgent to deal with multiple privacy regimes under LDP.



(a) Users set privacy regimes independently and the collection activity that each user may participate in is different.

(b) Multiple regimes in complex mining tasks, users participate in different steps with different regimes

Fig. 1. Application scenarios for multiple privacy regimes.

In this paper, we assume that users may have different privacy acceptances for personalization: Once a user sets his overall privacy regime ϵ, he does not change this value frequently and his participation in any tasks will automatically consume a certain proportion of this ϵ until it's used up; Since the collection behavior is usually long-term and the user's privacy data may change (e.g., web pages visited), users' specific choices of privacy regimes are not related to the value of the private data.

As far as we know, frequency estimation is the most basic LDP mining task, so it is meaningful to apply it under the mechanism of multiple privacy regimes. An obvious method of frequency estimation in this scenario, is to divide users with the same ϵ into the same group, and estimate frequency for different groups separately. In the end, service provider aggregates each group's estimated value by weighting. This is discussed in Sect. 3.

Contributions

- We propose MLE method which applies the idea of parameter estimation to obtain an optimal estimate from user groups with different privacy regimes. Our theoretical analysis shows that the accuracy of MLE can be equivalent to tradition method which forces all users to choose one specific privacy regime, and this equivalence shows that MLE is practical.

- We propose S-MLE method to handle the situation when all users' privacy regimes are subject to continuous distribution in one mining task. It uses a predefined β parameter to segment the contiguous ϵ on the basis of MLE. The experiments show that the β parameter of S-MLE may greatly influences the accuracy under certain conditions.
- We propose Adapt-MLE method which encompasses different LDP schemes for multiple privacy regimes. And the performance of Adapt-MLE is better than that of MLE, especially when the discrete values of ϵ of different users span a wide range.

2 Preliminaries and Notations

We assume that all users are willing to share their information to help service provider update its statistical information. For the sake of privacy, each user perturbs his own data by advanced technique (via LDP) with different demands for security. The service provider aims to find out the frequencies of values among the population. Such a process involves the following preliminaries.

2.1 Local Differential Privacy

Definition 1 *(Local Differential Privacy). An algorithm A satisfies ϵ-local differential privacy (ϵ-LDP) where $\epsilon > 0$, if and only if for any input v_1 and v_2, we have $\forall y \in Range(A)$,*

$$\frac{Pr(A(v_1) \in y)}{Pr(A(v_2) \in y)} \leq e^{\epsilon},$$

where Range(A) denotes the set of all possible outputs of the algorithm A.

When privacy regime ϵ is small, the adversary can't identify the true value from the noise version reliably. The basic core of algorithm A is Randomized response (RR) [21], which is a statistical technique used for collecting social embarrassing questions.

2.2 Frequency Estimation in LDP

Let $\boldsymbol{f} = (f_1, ..., f_k)$ be a probability distribution on a set containing k candidate values, f_1 is the true frequency of v_1 and the sum of $\boldsymbol{f}$ is 1. We can consider that $\boldsymbol{f}$ is the proportion of N users choosing different values. Using LDP allows the server to obtain an estimate $\widehat{\boldsymbol{f}} = (\widehat{f_1}, ..., \widehat{f_k})$ without obtaining the user's original data.

Each of N users holds a value v_j taken from the above k candidates and shares this v_j to the service provider in a LDP manner. In the beginning, user need to encode the v_j to a specific format, $x_j = E(v_j)$, then select a parameter ϵ t to obtain y_j by randomization. Finally, service provider aggregates N data records $(y_1, ..., y_N)$ and figures out $\widehat{\boldsymbol{f}}$.

2.3 General LDP Schemes

Randomization mechanisms that satisfy LDP have been widely studied in recent years, our work involves the following very common schemes.

kRR (k-ary Randomized Response) [12] is a generalization of binary randomized response (RR) which performs well in low privacy level.

Base RAPPOR is simplest configuration of RAPPOR [9] which has been widely accepted. It has a output alphabet $\upsilon = \{0, 1\}^k$ of size 2^k. It first maps $v_i(1 \leq i \leq k)$ onto $e_i \in 0, 1^k$, where $x = e_i$ is the i-th standard basis vector. At last a length-k binary vector y is generated from x by $y = RR(x)$, RR here can be seen with a pair of alterable probability p and q.

$$Pr(y[i] = 1|x[i] = 1) = p; \quad Pr(y[i] = 1|x[i] = 0) = q \tag{1}$$

Base Rappor sets p to $e^{\epsilon/2}/(e^{\epsilon/2}+1)$ and q to $1-p$. For the fact that Base RAPPOR is classical and is the basis of many other schemes, we mainly use it to analyze the scenario of multiple privacy regimes in Sects. 3 and 4.

Optimal Scheme [22] **and OLH** [18] are two similar schemes, and they are all obtained by optimizing the probability of RR in the Base RAPPOR (Change p, q in Eq. 1), the difference is that the latter is evolved from SH [3] and reduces communication cost by hash method.

At Last, frequency estimation formulas of these schemes are all

$$\hat{f}_i = \frac{C(i) - q * N}{(p-q) * N}$$

where $C(i)$ is the count of reported vector which has the i'th bit being 1. From OLH, we get the following properties.

Lemma 1. *For LDP scheme which uses RR with probability p and q, the frequency estimation $\hat{f}_i$ is an unbiased estimate of f_i, and its variance is*

$$var(\hat{f}_i) = \frac{(f_i * p + (1-f_i) * q) * (1 - f_i * p - (1-f_i) * q)}{N(p-q)^2}$$

Employing Base rappor's settings for p and q and taking $e^{\epsilon/2} > 1 >> f_i$ into account, Base rappor's variance is as follows:

$$var(\hat{f}_i) = \frac{e^{\epsilon/2}}{N(e^{\epsilon/2}-1)^2} \tag{2}$$

3 Problem Formulation

In this paper, we consider the specific LDP problem that users specific choices of privacy level will lead to multiple ϵ in one mining task. In this section, we first

Table 1. Notations

Notations	Explanations	Notations	Explanations
v	Raw data in frequency estimation	p, q	Defined in equation (1)
x	Encoded data	N	Total number of users in the collection
y	Sanitized data	n_m	The size of the m-th group
$\boldsymbol{f}$	Frequency distribution	k	The domain of value
f_i	The true frequency of v_i	$\hat{f}_i$	An estimate of f_i

introduce a primary method called Raw-PCE (Personalized Count Estimation) which can be considered a simplified version of PCE [5] proposed for the handling spatial data aggregation. Compared with the original PCE scheme, only the customization of privacy regime is preserved in Raw-PCE while the customization of other factors are omitted. Then we analyze the estimate of Raw-PCE and get a new probability model (Table 1).

3.1 A Multiple Privacy Regime Scheme: Raw-PCE

Raw-PCE is an intuitive and primitive method handling this scenario. Suppose there are totally M different privacy regimes namely ϵ_1, ϵ_2, ...,ϵ_M.

Firstly, the service provider groups the users according to their personalized privacy regimes which results in totally M groups.

Then, each group with the same privacy regime independently generate its frequency estimate vector, the estimate vector generated by group m is denoted as $\boldsymbol{f(\hat{m})} = (f(\hat{m})_1, f(\hat{m})_2, ..., f(\hat{m})_k)$. Totally M estimates are generated. Without loss of generality, here we consider only one candidate value v to simplify the problem. The M estimates for value v can be denoted as $\hat{f_{(1)}}, \hat{f_{(2)}}, ..., \hat{f_{(M)}}$.

Finally, if there is no other auxiliary information, the way in which the estimated value $\hat{f}$ is calculated by Raw-PCE is as follows:

$$\hat{f} = \sum_{m=1}^{M} \hat{f_{(m)}} \alpha_{(m)} \tag{3}$$

where $\alpha(\cdot)$ represents the weights of each group's estimation. Raw-PCE here ignores the fact that every estimate's accuracy is different, and it just takes the size of each group as the weights where $\alpha_m = n_m/N$. Combining Eqs. 2 and 3, we have:

Lemma 2. *In Base RAPPOR scheme, estimation of Raw-PCE has the variance as follows,*

$$var(\hat{f}) = \sum_{m=1}^{M} \frac{n_m^2 e^{\epsilon_m/2}}{(e^{\epsilon_m/2} - 1)^2 * N^3}$$

The above formula shows that the error is cumulative, whenever there is an estimation with large error among M estimations, final errors are greatly increased which is rather unacceptable.

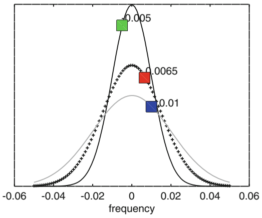


Fig. 2. The true frequency of $v_i(i \in [k])$ is 0, and there are three estimations of f_i which are drawn from three normal distribution.

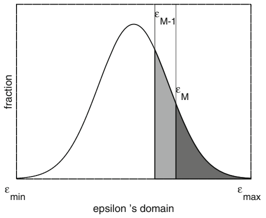


Fig. 3. S-MLE: dividing continuous ϵ into discrete values and the size of each shadow area is β.

3.2 Probabilistic Model of Multiple Regimes Setting

Although users are divided into different groups with different ϵ, their choices of privacy regimes can be considered as irrelevant to the choices of their favourite items, which means user's possibility of choosing item v is the same in all M groups.

After the randomization process, the reported number C ($C(i)$ is introduced in Sect. 2.3 and here we omit i) is a random variable from a binomial distribution, namely $C\tilde{}\mathbf{B}(N, pf + q(1-f))$. We use p' to denote $pf + q(1-f)$. Furthermore, N is usually big enough to ensure normal approximation and we have $C\tilde{}\mathbf{N}(Np', Np'(1-p'))$. $\hat{f}_{(m)}$ is a normalization of $C_{(m)}$ and follows a Gaussian distribution.

Then all M groups follow M different normal distributions, which share the same expectations but have different variances due to the different values of ϵ. Therefore, as Fig. 2 shows, the M estimates generated by M groups, respectively $\hat{f}_{(1)}, \hat{f}_{(2)}, ..., \hat{f}_{(M)}$, can be regarded as M random samplings of the actual user proportion f_v, each of which follows a unique normal distribution.

The problem of multiple privacy regimes can be regarded as equivalent to the problem of obtaining the best estimate of f_v from M group estimations $\hat{f}_{(1)}, \hat{f}_{(2)}, ..., \hat{f}_{(M)}$, which are random samples separately drawn from M normal distributions. Our target is simplified as to give the optimal estimate of the expectation with the help of parameter estimation methods.

4 Estimate Methods

In this section, we present two types of frequency estimate for multiple privacy regimes. In Sect. 4.1, we discuss a new MLE method and give theoretical proof.

Then we prove MLE's accuracy is much better than the basic Raw-PCE (given in Sect. 3.1). In Sect. 4.2, we discuss the situation when there exists large number of groups after grouping operation and propose S-MLE method on the basis of MLE.

4.1 MLE

Maximum likelihood is an effective method in parameter estimation, which is adopted here to generate unbiased expectation f. Once the service provider gets M estimations $(\hat{f_{(1)}}, \hat{f_{(2)}}, ..., \hat{f_{(M)}})$ and their variances as well, the Maximum likelihood estimation (MLE) $\hat{f}$ is defined as follows:

Theorem 1. *Given the M estimate $(\hat{f_{(1)}}, \hat{f_{(2)}}, ..., \hat{f_{(M)}})$, the MLE for the multiple privacy regimes scenario is*

$$\hat{f} = (\sum_{m=1}^{M} \frac{\hat{f_{(m)}}}{var(\hat{f_{(m)}})}) / (\sum_{m=1}^{M} \frac{1}{var(\hat{f_{(m)}})})$$

The proof is in Appendix A.

It's interesting to observe that the expectation we derived for MLE is in a weighted-sum manner. The weights become the reciprocal of the variances. Bring it to Eq. 3 for demonstration,

$$\alpha_m = \frac{1/var(\hat{f_{(m)}})}{\sum_{m=1}^{M} 1/var(\hat{f_{(m)}})}$$

Estimation Accuracy Analysis. Due to MLE is unbiased and contains grouping operation, its accuracy can be somehow equivalent to a traditional LDP method with a special privacy regime ϵ. Assuming there are two service providers doing the same collection on a population. One allows user to choose different ϵ and groups them, finally there are M groups whose size and privacy regime are $(n_1, \epsilon_1), (n_2, \epsilon_2), ..., (n_M, \epsilon_M)$; The other makes all users in the same privacy regime. So in which condition they achieve the same level of accuracy or their estimations have the same variance. The latter obliges all users to have the same ϵ, and here we call this reckless method traditional estimation (TE).

Theorem 2. *The variance of the estimates obtained by the MLE is $var(\hat{f}) = 1/(\sum_{m=1}^{M} \frac{1}{var(\hat{f_{(m)}})})$.*

The proof is in Appendix B. It can be inferred from the formula that if the data collector only uses the estimate with the least error, its effect is not as good as that of MLE which combines all the estimates.

When substituting the variance generated by Base RAPPOR into the Theorem 2, we obtain a new lemma.

Lemma 3. *In Base RAPPOR, if any* $e^{\epsilon_m/2} >> 1$*, there exists a privacy regime* $\epsilon' \approx 2*ln\frac{\sum n_m*exp(\epsilon_m)}{\sum n_m}$ *such that makes the accuracy of directly using TE method with* ϵ' *equals to MLE with multiple* $\epsilon_m (m \in [M])$*.*

The proof is in Appendix C.

Comparing Theorem 2 with Lemma 2, we find that the overall variance of MLE is much lower than that of Raw-PCE method. The explanation can be that Theorem 2 implies the final estimate will be accurate as long as at least one of the estimates has low variance, while Lemma 2 has high variance if just one group has high variance. Our experiments also show MLE is much more accurate than Raw-PCE.

4.2 S-MLE

When the user are allowed to choose any value as their overall privacy regimes from a considerate large set, there will be too many groups and some groups inevitably contain too few users. In this scenario, the above MLE method may be inapplicable because large errors are introduced into $\hat{f}$. And one extreme situation is that ϵ is continuous function over real number field as shown in Fig. 4. In this section, we start by analyzing why the minimum size of the group (βN) should be set and explaining what factors will affect the value, then we provide a supplement method called S-MLE for this scenario.

$\boldsymbol{\beta N}$**: Minimum Size of the Group.** From the perspective of sampling theory, the essence of grouping process in MLE is that the users are randomly sampled into M groups, and the frequency estimation result of each group is equivalent to the result of sampling scheme without replacement. And what we'll find is that the sample size of this M group is different and there may exist invalid sample group because sample survey with low sample size introduces lots of sampling error and would not represent the whole. Specifically, small group's unbiased estimation $\hat{f_{i(m)}}$ on value v_i differs greatly from the actual results of the population. Namely, $|E(\hat{f_{i(m)}}) - f_i| < \sigma$ can't hold where σ is tolerable sampling error.

So it makes sense to determine the minimum of the sample size which is also a basis for dividing the group. According to the sampling theory, the sample size is usually determined by the variation degree of the research object, the total number of samples and the demand for accuracy. In our MLE, sample size can be mainly determined by the size of candidate set (k), the total number of users (N), and the complexity of candidate set's frequency (the distribution of $\boldsymbol{f}$). So we get the proposition as follows:

Proposition 1. *In MLE, for any group whose size does not exceed* βN*, it's necessary to ignore its estimation or make users in this group join other higher privacy level group.*

Combining the Proposition 1 and MLE, we need to figure out an empirical value for β and segment privacy regimes and re-merge existing groups.

S-MLE. The S-MLE method can be further extended to the situation when the type of ϵ value is unlimited (continuous distribution).

From Theorem 2 we can see that the entire accuracy is depend on all users' distribution of ϵ. let users who have temp largest ϵ value form a group (size $= \beta * N$) recursively is an efficient segmenting way (segmenting in Fig. 3). Since it's difficult and complexity to figure out sample size β, we give some empirical values here. In the experiment when fixing N to 100000, we find $\beta \in [0.05, 0.1]$ ($[0.1, 0.15]$) is reasonable for the situation when $\boldsymbol{f}$ is generated by zipf's distribution (uniform distribution) and k is ranging from 20 to 200, β should be bigger as k increases.

5 Adapt MLE and Universality of Mining Scenarios

MLE and S-MLE have been able to achieve relatively high accuracy in frequency estimation by Base RAPPOR. Furthermore, they are also applicable for other mining scenarios like heavy-hitter identification, and replace Base RAPPOR with other scheme for better accuracy. In this section, we propose the Adapt MLE method which adaptively selects the most suitable LDP scheme for each group of users to share data, then we briefly show how to apply MLE to other LDP mining tasks.

Adapt MLE. The process of selecting schemes just fits in with some work [11, 18, 22] on how to select LDP schemes based on privacy regime ϵ, the size of k and communication cost.

Here we attach importance to accuracy and use variance as the evaluation criteria to select LDP scheme for frequency estimation. Because uniform distribution is the most difficult to estimate analyzed by minimax [22], we set each value in $\boldsymbol{f}$ to be the same and easily calculate the variance of each scheme through Lemma 1. So by comparing their variances, we can get the following directly:

$$Adapt\ MLE(Group\ m) = \begin{cases} kRR & if\ k < 3e^{\epsilon_m} + 2 \\ OptimalSchemes & otherwise \end{cases}$$

In other scenarios like sampling step in frequent item mining [20], "$3e^{\epsilon} + 2$" might change slightly. So the accuracy of Adapt MLE is higher than that of MLE when the discrete values of ϵ of different users span a wide range. However, accuracy is not the only factor that matters, communication cost and computational complexity are also worth considering in real world.

As far as we know, our multiple privacy setting still can be used for heavy-hitters identification. SH [3] consists of two important steps. First step uses hash function to separate the values into a lot of channels, with high probability each channel has at most one frequent value, then identify whether there is a frequent item in each channel. Referred to Chen [5], multiple privacy setting is fully applicable to this step. The second step employs a frequency oracle to estimate the frequency of those frequent values obtained from first step. And this is similar to the mining task of frequency estimation while obtaining variance from frequency oracle is in another form and the final estimates are slightly biased (still consistent with the goal of heavy-hitters).

6 Experiments

In this section, we evaluate and compare the performance of our proposed personalized approach through extensive experiments. Since there is no existing work for our settings, we mainly verify the correctness of our analysis.

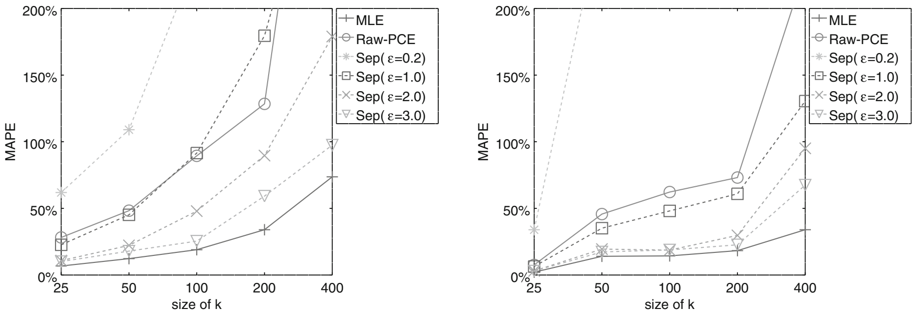
Setup. All experiments are performed 10 times and we plot the Mean Absolute Percentage Error (MAPE) of all frequency estimation. The MAPE is $\frac{1}{k}\sum_{i\ in[k]} \frac{|\hat{f}_i - f_i|}{f_i+\sigma}$, where f_i is the actual fraction of all users taking value i and σ is to prevent the denominator from 0.

In RAPPOR [9] with h = 1, their value for epsilon is actually set to $2ln3$ and the number of users is a million level. And Apple also set epsilon to 1 or 2. So we assume that 100 thousand users participates the collection, and set M options in most experiments to $\boldsymbol{\epsilon} = [0.2, 1.0, 2.0, 3.0]$ and the proportion of users in each group is $G = [0.2, 0.3, 0.3, 0.2]$ where G denotes the corresponding proportion.

For better verification of correctness, we generate two synthetic data which are from Zipf's distribution (parameter a = 2.5) and uniform distribution. The schemes used in each experiment contain KRR, Base Rappor and Optimal Scheme. We change the distribution of $\boldsymbol{f}$ by controlling the size of k.

6.1 Accuracy of MLE Method

Whatever the distribution of $\boldsymbol{f}$ is in Fig. 4, MAPE curves generated by four groups shows that group with higher ϵ generates better estimates; Raw-PCE's accuracy has been greatly affected by the group with low variance, namely $\epsilon =$



(a) uniform distribution of $\boldsymbol{f}$

(b) Zipf's distribution of $\boldsymbol{f}$

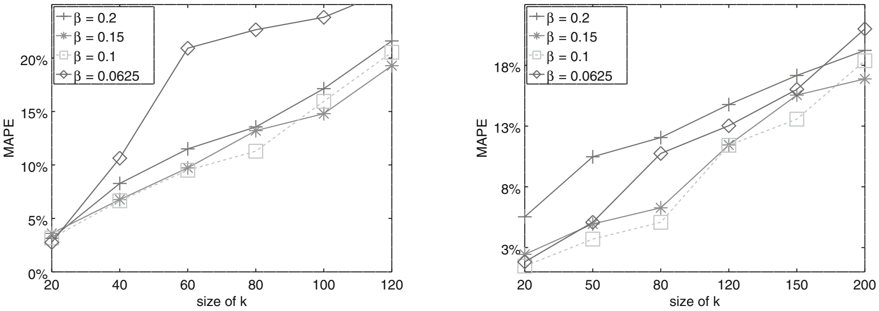
Fig. 4. Comparing MLE and Raw-PCE, varing k: estimates generated by each group are marked as Sep

0.2; From Theorem 2 and this figure, we know the variance of MLE is always smaller than the variance estimated by each group.

In uniform distribution (Fig. 4(a)), MAPE increases as k increases, roughly the same multiple because the denominator of MAPE's calculation is actually $1/k$. In fact, the size of the variance is independent of k. So in zipf's distribution (Fig. 4(a)), the change in MAPE will not be obvious when k does not exceed 200. This also the reason why uniform distribution is difficult to estimate.

6.2 S-MLE Method on Continuous ϵ

When there are too many options of M or all users' ϵ is continuously distributed, the value of β can help to divide all users into $\lfloor 1/\beta \rfloor$ groups as described in Sect. 4.2. Small β will increase sampling error, but it can also reduce the overall variance from Theorem 2. A balance between overall variance and sampling error is reasonable.



(a) uniform distribution of $\boldsymbol{f}$

(b) zipf's distribution of $\boldsymbol{f}$

Fig. 5. the influence of β 's value on MAPE, varying k.

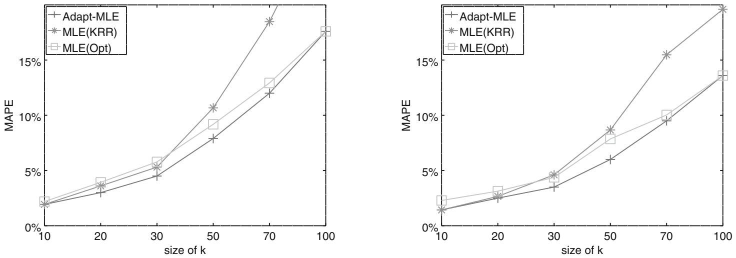
In this part, we make the proportion of users choosing different ϵ obey $N(2.5, 0.8)$ and ϵ's contiguous interval be [1.0,4.0], namely G obeys $N(2.5, 0.8)$ and $\boldsymbol{\epsilon}$ is continuous in [1.0,4.0]. It can be observed from Fig. 5(a) that when k is small, the MAPE with small β is acceptable, namely, the sampling error has little effect. On the contrary, when k becomes larger than 60 for $\beta = 0.0625$, the sampling error even exceeds the error generated by estimator. But for Zipf's distribution, the effect of k on sampling error is not obvious until $k > 150$. So for our settings above concerning the number of users and ϵ ratio, $\beta \in [0.05, 0.1]$ for Zipf's distribution and $\beta \in [0.1, 0.15]$ for uniform distribution are appropriate choices.

6.3 Multiple Schemes for Different Groups

In Sect. 5, we claim that users in different groups can use different LDP schemes to achieve better accuracy. From Sect. 5, kRR performs better when $k < 3e^{\epsilon} + 2$.

We divide 100 thousand users into 4 groups with $\epsilon = [1.0, 2.0, 3.0, 3.2]$ and $G = [0.3, 0.35, 0.3, 0.05]$.

Learning from Fig. 6, “Adapt-MLE” performs better because users of these 4 groups use kRR if k are less than 10, 25 and 62 and 76 respectively and otherwise use Optimal Scheme.



(a) uniform distribution of $\boldsymbol{f}$

(b) zipf’s distribution of $\boldsymbol{f}$

Fig. 6. Adapt MLE with multiple schemes, “MLE(kRR)” represents only KRR and “MLE(Opt)” represents only Optimal scheme.

7 Related Work

The traditional differential privacy (DP) was developed for interactive query-response on a central database and provides theoretical privacy guarantee by mathematically randomizing the results of statistical queries. However, the major limitation of DP is that all users need to trust a central server. Despite attacks from aggregate queries, individual’s data may also suffer from privacy leakage before aggregation [8].

On the other hand, Local differential privacy (LDP) [13], a variant of DP, guarantees privacy of data without that server. Random Response (RR) [21], where the user responds either true or opposite answer depending on coin flipping, is the most basic technique in LDP schemes.

Widely accepted schemes for frequency estimation under LDP are Rappor by Erlingsson et al. [9] and succinct histogram (SH) by Bassily and Smith [3]. RAPPOR’s key idea is encoding values into Bloom filters and applying RR to each bit of Bloom filters. In order to conquer hash collision problems in Bloom filters, RAPPOR brings in cohorts. In this paper, we use RAPPOR’s no bloom filter version to analyze our multiple setting. SH’s has two important data structures—frequency oracle and succinct histogram, these two work together to estimate those values whose frequencies exceed η, so some applications do heavy hitters [5,16,19] identification referred to SH. Due to the high complexity of SH, Bassily et al. [4] recently developed an efficient way to query the frequency estimation based on SH mechanism.

In addition, discrete distribution estimation under LDP considers all values’ frequency accuracy. Kairouz et al. [11] analyzed several key factors (privacy regimes, discrete distribution) which affect accuracy. Ye et al. [22] came up with Optimal Scheme for discrete distribution estimation. About [11,22], their analysis tools all contained minimax with l_2-norm as loss function which is similar to variance. Wang et al. [18] introduced a framework that can generalize the most LDP schemes by recognizing RR's features. These work all have a prerequisite that the size of k is limited.

In personalization privacy fields, Jorgensen et al. [10] incorporated personalized settings for DP (PDP), and Li et al. [14] proposed a k-partition strategy to improve it. Then Chen et al. [5] first introduced the concept of personalized privacy in LDP (PLDP), it allows users to have two optional privacy regime ϵ and τ. The former has no change, while τ represents a small piece of the candidates list. They assume some users set small size of candidates list and this part of users can greatly improve their performance on heavy-hitter mining task. Obviously, τ makes this personalization process complex for users. Akter et al. [1] borrowed the definition of PLDP to estimate numeric data like average instead of heavy-hitters mining.

8 Conclusion

In this paper, we mainly study frequency estimation under Local Differential Privacy (LDP) in multiple regimes scenarios. We have formulated the problem of multiple privacy levels and proposed a MLE method to deal with this situation. Then, we propose S-MLE and Adapt-MLE to deal with the situation when users' privacy levels are in some special cases.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. U1636216) and National Key R&D Program of China (No. 2016YFB0502302).

9 Appendix

A. Proof of Theorem 1

Proof. $(\hat{f_{(1)}}, \hat{f_{(2)}}, ..., \hat{f_{(M)}})$ are drawn from different normal distributions, normal distribution has probability density function as follows:

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-u)^2}{2\sigma^2})$$

According to probability density function $g(x)$, we know the closer estimation $\hat{f_{(m)}}$ is to the expectation, the greater the $g(\hat{f_{(m)}})$. For ease of calculation, we use Eq. 2 to ignore the effect of f_i on variance. $g(\hat{f_{(m)}})$ actually has only one variable–expectation. Separately bring each $\hat{f_{(m)}}$ into function and multiply these functions according to maximum likelihood, we get the final target function which needs to be maximized.

$$F(f) = \prod_{m=1}^{M} g_m(f)$$

We first turn it to logarithmic function $y = ln(F(f))$, and after derivation, the first derivative and the two derivative of $F(f)$ are obtained sequentially.

$$y' = \frac{\partial ln(F(f))}{\partial f} = -\sum\nolimits_{m=1}^{M} \frac{\hat{f_{(m)}} - f}{\sigma_m^2}$$

$$y'' = \frac{y'}{\partial f} = \sum\nolimits_{m=1}^{M} \frac{1}{\sigma_m^2}$$

Through simple analysis, y'' is always bigger than 0 and y' is a strictly monotone increasing function. So $F(f)$ is a convex function with a max value. Then set the first derivative function to zero, here when $\hat{f} = (\sum_{m=1}^{M} \frac{\hat{f_{(m)}}}{\sigma_m^2})/(\sum_{m=1}^{M} \frac{1}{\sigma_m^2})$, we can get the maximum of the $F(f)$.

B. Proof of Theorem 2

Proof. First use t_m to denote $var(\hat{f_{(m)}})$, the final estimation using maximum likelihood is $\hat{f} = (\sum_{m=1}^{M} \frac{\hat{f_{(m)}}}{t_m})/(\sum_{m=1}^{M} \frac{1}{t_m})$. When we calculate the variance of $\hat{f}$ as follows:

$$var(\hat{f}) = var(\sum\nolimits_{m=1}^{M} \frac{\hat{f_{(m)}}}{t_m} / \sum\nolimits_{m=1}^{M} \frac{1}{t_m})$$

Since the estimations $f_m (m \in [M])$ are independent of each other, and t_m here is actually a constant number.

$$var(\hat{f}) = \sum\nolimits_{m=1}^{M} (\frac{var(\hat{f_m})}{t_m^2})/(\sum\nolimits_{m=1}^{M} \frac{1}{t_m})^2 = 1/\sum\nolimits_{m=1}^{M} \frac{1}{t_m}$$

C. Proof of Lemma 3

Proof. We still judge the accuracy of the final estimation from the perspective of variance. The Lemma 1 shows base rappor's estimation variance is $var(\hat{f_i}) = \frac{e^{\epsilon/2}}{n(e^{\epsilon/2}-1)^2}$, for the sake of simplicity, let's first assume $e^{\epsilon/2} \gg 1$ and use t_m to denote $var(\hat{f_{(m)}})$. So that $t_m = (1/(n_m e^{\epsilon_m/2}))$.

We are clear that the $\hat{f}$'s variance and $\hat{f_{(m)}}$'s variance are the same format, because f is regarded as using Base RAPPOR on the whole population while all users have the same privacy regime ϵ'.

Combining the above equations and Theorem 2 together, we can find $\epsilon' = 2 * ln\frac{\sum n_m * exp(\epsilon_m)}{\sum n_m}$. If $e^{\epsilon/2} \gg 1$ doesn't hold in some situation, the calculation can still be based on the above formula and the result will become a little more complicated.

References

1. Akter, M., Hashem, T.: Computing aggregates over numeric data with personalized local differential privacy. In: Pieprzyk, J., Suriadi, S. (eds.) ACISP 2017. LNCS, vol. 10343, pp. 249–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59870-3_14
2. Apple2017: macos sierra: share analytics information with apple. https://support.apple.com/kb/PH25654?locale=en_US&viewlocale=en_US
3. Bassily, R., Smith, A.: Local, private, efficient protocols for succinct histograms. In: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, pp. 127–135. ACM (2015)
4. Bassily, R., Stemmer, U., Thakurta, A.G., et al.: Practical locally private heavy hitters. In: Advances in Neural Information Processing Systems, pp. 2285–2293 (2017)
5. Chen, R., Li, H., Qin, A.K., Kasiviswanathan, S.P., Jin, H.: Private spatial data aggregation in the local setting. In: IEEE International Conference on Data Engineering, pp. 289–300 (2016)
6. Dwork, C.: Differential privacy. In: International Colloquium on Automata, Languages, and Programming, pp. 1–12 (2006)
7. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
8. Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy. Now Publishers Inc., Hanover (2014)
9. Erlingsson, Ú., Korolova, A., Pihur, V.: RAPPOR: randomized aggregatable privacy-preserving ordinal response. In: ACM SIGSAC Conference on Computer and Communications Security, pp. 1054–1067 (2014)
10. Jorgensen, Z., Yu, T., Cormode, G.: Conservative or liberal? Personalized differential privacy. In: 2015 IEEE 31st International Conference on Data Engineering (ICDE), pp. 1023–1034. IEEE (2015)
11. Kairouz, P., Bonawitz, K., Ramage, D.: Discrete distribution estimation under local privacy. arXiv preprint arXiv:1602.07387 (2016)
12. Kairouz, P., Oh, S., Viswanath, P.: Extremal mechanisms for local differential privacy. In: Advances in Neural Information Processing Systems, pp. 2879–2887 (2014)
13. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S.: What can we learn privately? In: Proceedings IEEE Annual IEEE Symposium on Foundations of Computer Science, vol. 40, no. 3, pp. 793–826 (2008)
14. Li, H., Xiong, L., Ji, Z., Jiang, X.: Partitioning-based mechanisms under personalized differential privacy. In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) PAKDD 2017. LNCS (LNAI), vol. 10234, pp. 615–627. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57454-7_48
15. Narayanan, A., Shmatikov, V.: How to break anonymity of the Netflix prize dataset. Comput. Sci. (2007)
16. Qin, Z., Yang, Y., Yu, T., Khalil, I., Xiao, X., Ren, K.: Heavy hitter estimation over set-valued data with local differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 192–203. ACM (2016)
17. Tang, J., Korolova, A., Bai, X., Wang, X., Wang, X.: Privacy loss in Apple's implementation of differential privacy on macOS 10.12. arXiv preprint arXiv:1709.02753 (2017)

18. Wang, T., Blocki, J., Li, N., Jha, S.: Locally differentially private protocols for frequency estimation. In: Proceedings of the 26th USENIX Security Symposium, pp. 729–745 (2017)
19. Wang, T., Li, N., Jha, S.: Locally differentially private heavy hitter identification. arXiv preprint arXiv:1708.06674 (2017)
20. Wang, T., Li, N., Jha, S.: Locally differentially private frequent itemset mining. In: IEEE Symposium on Security and Privacy, p. 0. IEEE (2018)
21. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. **60**(309), 63–69 (1965)
22. Ye, M., Barg, A.: Optimal schemes for discrete distribution estimation under local differential privacy. In: 2017 IEEE International Symposium on Information Theory (ISIT), pp. 759–763. IEEE (2017)