Differentially Private Parameter Estimation from Distributed Sources

Wei-Ning Chen and Janet Sung

Abstract

We consider the problem of parameter estimation with data collected from M parties, each party holding N copies of identically and independently distributed (i.i.d.) samples. In many circumstances, parties are not willing to reveal their private information, so the released data will be ϵ-*differential private*. In this paper, we propose two algorithms to estimate parameter: for general dataset with the number of samples satisfying $M = o\left(N^2\right)$ and the privacy parameter satisfying $\epsilon = \Omega\left(N^{1/6}M^{-1/3}\right)$, we proposed *subsample-and-aggregate* estimator, which is asymptotically efficient (i.e. its MSE is asymptotically equal to Fisher information). On the other hand, if the underlying distributions are from exponential family and if $\epsilon = \Omega\left(1/N\right)$, we showed that *sufficient-statistic-averaging* is always efficient as N goes to infinity. Our results showed that as long as the number of parties does not grow too fast, one can guarantee differential privacy *for free*, without sacrificing the performance of estimation.

I. Introduction

Efficiently extracting information from large-scale dataset is one of the key factors to the success in recent data science. Many parametric approaches from classical statistical inference and non-parametric machine learning algorithms have been proposed, and with the increasing availability of data, these algorithms showed their great power in real-world application. However, with the explosive growth of available data in both cyber and physical world, almost every piece of them carries someone's fingerprint or sensitive personal information., and existing research have shown that such information could be utilized to identify one's identity and invade individual privacy. For example, [1] exploited the public relevant database to de-anonymize dataset released by netflix, proving that even removing sensitive entries from a dataset, such as name or SSN, is not enough to protect personal privacy.

To address the challenge, Dwork proposed *differential privacy* [2] to quantitatively measure the level of privacy protection in a statistical database. An intuitive approach that satisfies the differentially private constraint is to perturb data with random noise, which is mostly generated from a Laplacian distribution with fixed variance. Though the proposed perturbation mechanism works in single database, it increases the noise of the uncertainty of statistical inference, making maximum likelihood estimator (MLE) no longer statistically efficient (meaning its mean square error is asymptotically equal to *Fisher information*). In other words, the mechanism preserves the privacy at the price of deteriorating the utility. To address such difficulty, in [3], Smith proposed an *subsample-and-aggregation*

W.-N. Chen and Janet Sung are with the Electrical Engineering department, National Taiwan University, Taiwan. This research is under the supervision of Professor I-Hsiang Wang

estimator that achieves Fisher information while preserving differential privacy. However their approach only holds when only one database is of interest, which may not capture the general real-world situation since in practice data is usually stored and collected from multiple sources.

In this paper, we investigate the problem of parameter estimation with data collected from M parties, each party holding N copies of identically and independently distributed (i.i.d.) samples. We proposed a strategy that, under some mild regularity conditions, computes an *statistical efficient estimator* of all MN samples, and simultaneously achieves ϵ-*differential privacy* with respect to each party. We also proved that if the underlying distributions of samples further belong to exponential family, then the regularity conditions can be removed, and as long as N goes to infinity, the proposed estimator is also asymptotically efficient. In addition, the proposed algorithm is computational efficient and the multi-party setting can be easily adopted into real world scenario.

A. Related Work

Prior to our work, Smith proposed a mechanism to obtain differentially private and efficient estimator [3]. It is possible to repeatedly used his mechanism in each of our party. However, the result is not optimal and the added noise might be redundant.

Beside point esimator for parametric model, another possible application of the database is the *classifier* for machine learning problem. In [4] Hamm et al. design an ensemble voting classifier using the empirical risk minimization approach to ensure differential privacy. However, in their setting there is a trusted third party. If so, one can simply let the trusted party train the classifier. Furthermore, the accuracy of the mechanism dropped significantly at high privacy level.

Another way to protect individual data in distributed setting is via cryptography approach [5]. However, in such way, the privacy constraint must be relaxed into ϵ-δ differential privacy and all data distance is assumed to be within a unit ball.

II. Problem Formulation

Consider M parties $\mathcal{P}_1, ..., \mathcal{P}_M$ with the i-th party $\mathcal{P}_i$ holds dataset $S_i = \{x_{i1}, ..., x_{iN}\}$, $i = 1, ..., M$. Each sample x_{ij} is drawn identically and independently from distribution P_θ, where $\theta \in \Theta$ is a compact set in $\mathbb{R}^n$ with diameter Λ. Our goal is to estimate θ privately from the distributed dataset $S_1, ..., S_M$. We will use the convention that capital letters (X, T , etc) refer to random variables, and the lower cases refer to certain realizations.

We begin by introducing some context about differential privacy.

A. Differential Privacy

We say two fixed datasets $\mathcal{D}$, $\mathcal{D}'$ are adjacent if $\mathcal{D}$ and $\mathcal{D}'$ differs in one item if for some i,

$$\mathcal{D} = x_1, x_2, ..., x_i, ..., x_N$$

$$\mathcal{D}' = x_1, x_2, ..., x_i', ..., x_N$$

A (randomized) query mechanism is differential privacy if for all possible pairs of neighboring datasets, the outputs of query have similar distribution:

Definition 2.1 (Differential Privacy): A randomized algorithm $T(\cdot)$ is ϵ-differentially private if for all neighboring pairs of datasets $\mathcal{D}$and $\mathcal{D}'$, and for all measurable subsets of outputs (events) $\mathcal{S}$:

$$\frac{\mathcal{P}(\mathcal{T}(\mathcal{D}) \in \mathcal{S})}{\mathcal{P}(\mathcal{T}(\mathcal{D}') \in \mathcal{S})} \leq e^{\epsilon}$$

This condition states that even the adversary holds as much side information as almost the whole database $\mathcal{D}$ (except for a single data), he or she cannot infer much about the unknown item.

Definition 2.2 (Sensitivity): For a deterministic function $f : \mathcal{X}^n \to \mathbb{R}^k$, we define the sensitivity of f as

$$S(f) = \max_{\mathcal{D}_1, \mathcal{D}_2} |f(D_1) - f(D_2)|.$$

Theorem 2.1 (Output Perturbation): Let f be defined as before and $\epsilon > 0$. Define randomized algorithm $\mathcal{A}$ as

$$\mathcal{A}(\mathcal{D}) = f(\mathcal{D}) + Lap(\frac{S(f)}{\epsilon}),$$

where the one-dimensional (zero mean) Laplace distribution $Lap(\lambda)$ has density $P_\lambda(x) = \frac{1}{2\lambda} e^{\frac{-|x|}{\lambda}}$, and $Lap(\lambda)^k = (l_1, ..., l_k)$ where each $l_i \overset{iid}{\sim} Lap(\lambda)$. Then $\mathcal{A}$ is ϵ-differential private.

B. The MLE and Efficiency

To evaluate performances of estimators, we propose the *mean square error*(MSE) to evaluate it, which is defined as following:

$$J_T(\theta) \overset{\text{def}}{=} \mathbb{E}_\theta((T(X) - \theta)^2)$$

The notation $\mathbb{E}_\theta$ means that X is drawn i.i.d. from distribution $f(\cdot, \theta)$. Note that if $T(\cdot)$ is unbiased, the MSE is simply the variance of $T(X)$. It is also well-defined even for randomized estimators: $T(X) = t(X, R)$, where R is externel random source.

Definition 2.3 (Maximium likelihood estimator): Let $L(\theta) \overset{\text{def}}{=} f(\mathbf{x}, \theta)$. The maximium likelihood ratio test is defined as

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L(\theta)$$

if such maximum exsits.

It is a classic result that, for well-behaved parametric families, the $\hat{\theta}_{MLE}$ exists with high probability and is asymptotically normal, centered around the true value θ. Moreover, its expected square error is given by the inverse of Fisher information at θ, where the Fisher information is defined as following:

$$I_f(\theta) \overset{\text{def}}{=} \mathbb{E}_\theta((\frac{\partial}{\partial \theta} \ln f(X, \theta))^2) \tag{1}$$

Lemma 2.1: Under appropriate regularity conditions, the MLE converges in distribution to a Gaussian centered at θ that is, $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{\mathcal{D}} N(0, \frac{1}{I_f(\theta)})$. Moreover, $J_{\hat{\theta}_{\text{MLE}}}(\theta) = \frac{1+o(1)}{nI_f(\theta)}$.

Fisher information upperbounds the MLE of estimators. If an estimtor matches this bound, we say it is *efficient*.

C. bias correction

The asymptotic efficiency of the MLE implies that its bias, $b_{\text{MLE}} = \mathbb{E}_\theta(\hat{\theta}_{\text{MLE}} - \theta)$ goes to zero more quickly than $\frac{1}{\sqrt{n}}$. However, we will need an estimator with much lower bias. This can be obtained via a (standard) process known as bias correction. Under appropriate regularity assumptions, we can describe the bias of MLE precisely, namely

$$\mathbb{E}_\theta(\hat{\theta}_{\text{MLE}} - \theta) = \frac{b_1(\theta)}{n} + O\left(\frac{1}{n^{\frac{3}{2}}}\right)$$

where $b_1(\theta)$ has a uniformly bounded derivative (see, for example, discussions in Cox and Hinkley [7], Firth [22], and Li [26]). Several methods exist for correcting this bias. The simplest is to subtract off an estimate of the leading term, using $b_1(\hat{\theta}_{\text{MLE}})$ to estimate $b1(\theta)$; the result is called the bias-corrected MLE,

$$\hat{\theta_{bc}} = \hat{\theta}_{\text{MLE}} - b_1(\hat{\theta}_{\text{MLE}})/n$$

Lemma 2.2: The bias-correcred MLE $\hat{\theta_{bc}}$ converges at the same rate as MLE, but with lower bias, that is,

$$b_{bc} = \mathbb{E}_\theta(\hat{\theta}_{\text{bc}} - \theta) = O\left(\frac{1}{n^{\frac{3}{2}}}\right)$$

III. Main Result

First, we consider the point estimation on non-distributed data. Adam Smith showed that if the data satisfied some weak regularity, then we can construct an efficient estimator by *subsample-and-aggregation* method. We state as the following theorem:

Theorem 3.1 (Centralized subsample-and-aggregate): Under appropriate regularity conditions, there exists a (randomized) estimator T which is ϵ-differentially private and asymptotically efficient, with mean-square-error $\frac{1}{nI_f(\theta)}\left(1 + O(n^{-1/5}\epsilon^{-6/5}\right)$. Thus one can choose $\epsilon_n = \Omega(n^{1/6})$, so that $\lim_{n\to\infty} \epsilon = 0$.

Proof. Detailed proof can be found in A.Smith's work[1]. ■

Algorithm 1 Centralized subsample-and-aggregate

procedure CSAA($\mathbf{x} = (x_1, ..., x_n), \epsilon$) ▷ Private Efficient Estimator T^*

Arbitrarily divide the input $\mathbf{x}$ into k disjoint sets $B_1, ..., B_k$ of $t = \frac{n}{k}$ points. We call these k sets the blocks of the input.

for each block $B_j = (x_{(j-1)t+1}, ..., x_{jt}))$ **do**

Apply the bias corrected MLE $\hat{\theta}_{\text{bc}}$ to obtain an estimate $z_j = \hat{\theta}_{\text{bc}}(x_{(j-1)t+1}, ..., x_{jt})$
end for

Compute the average estimate: $\bar{z} = \frac{1}{k}\sum z_j$

Draw a random observation R from a Laplace distribution with standard deviation $\sqrt{2}\Lambda/(k\epsilon)$.

return $T^* = \bar{z} + R$

Now, we generalized the result of theoerem to distributed framework. Consider the M parties possess datasets $\mathbf{x}_1, ..., \mathbf{x}_M$ respectively.

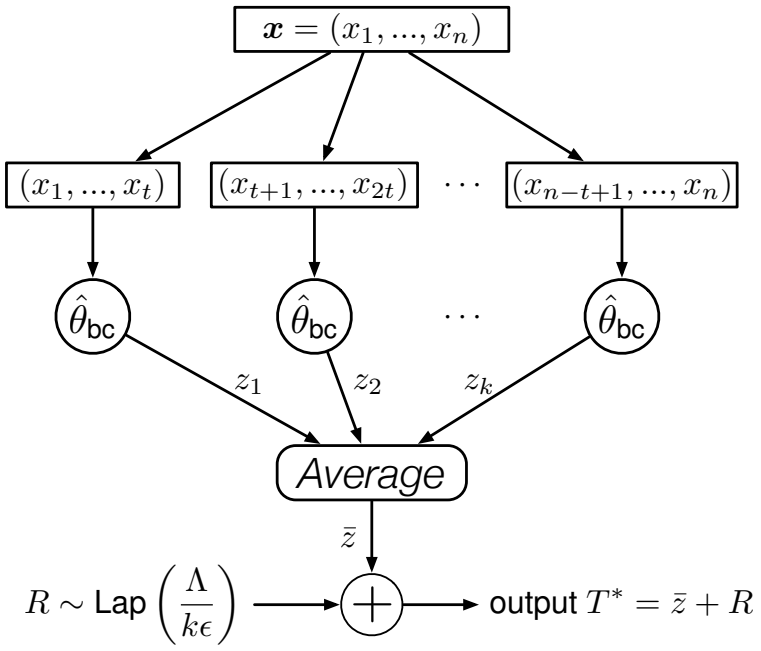

Fig. 1: The estimator T^*. When the number of bins k is $o(n^{2/3})$ and ϵ is not too small, T^* is asymptotically efficient

Algorithm 2 Distributed subsample-and-aggregate

procedure DSAA($(\mathbf{x}_1, ..., \mathbf{x}_M), \epsilon$) ▷ Private Efficient Estimator T^*

  for each party P_j, run CSaA as in Algorithm 1 , $\mathbf{x}_j = (x_{j1}, ..., x_{jN})$) **do**

    $T_j^* \leftarrow CSaA(\mathbf{x}_j, \epsilon)$

  end for

  Compute the average: $\bar{T} = \frac{1}{M}\sum T_j^*$

  return $\bar{T}$

Theorem 3.2 (Distributed subsample-and-aggregate): For M parties with datasets $\mathbf{x}_1, ..., \mathbf{x}_M$ respectively, if M not to large, say, $M = o(N^2)$, then the estimator $\bar{T}$ discribed in Algorithm 2 has MLE

$$J_{\bar{T}}(\theta) = \frac{1}{MN}\left(\frac{1+o(1)}{I_f(\theta)} + \frac{N\Lambda^2}{k^2\epsilon^2} + \frac{k^2 M}{N^2}\right),$$

where k is the number of subblocks in ALgorithm 1.

By appropriate choosing k, one can get the following result:

Corollary 3.1: If $M = o(N^2)$, and $\epsilon = \Omega(N^{\frac{1}{6}}M^{-\frac{1}{3}})$, $\bar{T}$, given by DSaA, is $\epsilon-$differential private and asymptotically efficient.

Subsample-and-aggregate method provides a brilliant idea to estimate parameters privately. However, in some case, our dataset may come from some well-behaved distribution, for example, the exponential family. If we utilize properties of the distribution, we can obtain more improvement. One of the best improvement assures the better convergence rate, which allows us to get rid of the constraint $M = O(N^2)$.

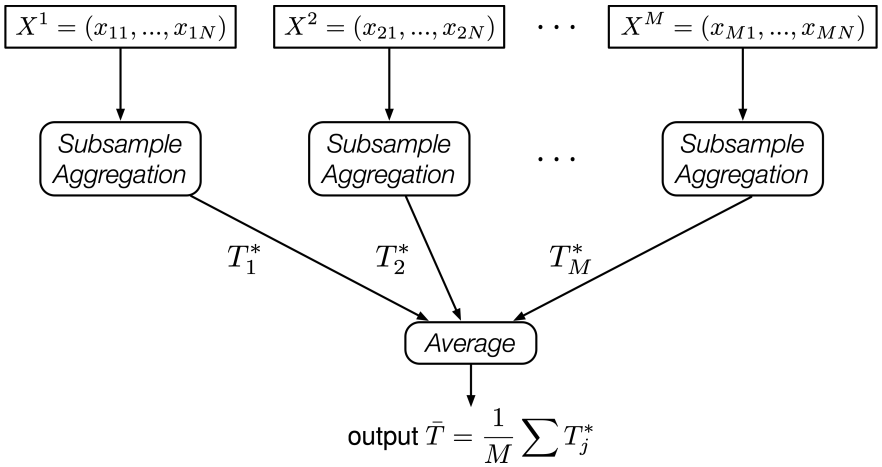

Fig. 2: The estimator $\bar{T}$. When the number of parties not to large , such that $M = o(N^2)$ and ϵ is not too small, $\bar{T}$ is asymptotically efficient

Definition 3.1 (exponential family): An distribution has the following form

$$f(x, \theta) = h(x) e^{\theta^T T(x) - \alpha(\theta)}$$

is called in *exponential family*. We introduces some basic property of exponential family which we may use in our scheme.

Proposition 3.1: If $x_1, ..., x_n \overset{iid}{\sim} X$, where X is a random variable from exponential family defined as above. Then $\frac{1}{n} \sum_{i=1}^{n} T(x_i)$ is a sufficient statistic of θ.

An observation of the original subsample-and-aggregate scheme is that averaging of MLE losses too much information. Instead, if we replace the $\hat{\theta}_{\text{MLE}}$ by the sufficient statitic $T(x_i)$, and apply MLE at the final steps, then we can obtain a better comvergent rate.

Algorithm 3 Centralized sufficient-statistics-averaging

procedure CSSA($\mathbf{x} = (x_1, ..., x_n), \epsilon$) ▷ Private Efficient Estimator T^*

Arbitrarily divide the input $\mathbf{x}$ into k disjoint sets $B_1, ..., B_k$ of $t = \frac{n}{k}$ points. We call these k sets the blocks of the input.

for each block $B_j = (x_{(j-1)t+1}, ..., x_{jt})$) **do**

Compute $t_j = \frac{1}{t} \sum_{i=1}^{t} T_{(j-1)t+i}$
end for

Compute the average estimate: $T = \frac{1}{k} \sum t_j$

Draw a random observation R from a Laplace distribution with standard deviation $\sqrt{2}\Lambda_T/(k\epsilon)$.

return $T^* = T + R$

Algorithm 4 Distributed sufficient-statistic-averaging

procedure DSSA($(\mathbf{x}_1, ..., \mathbf{x}_M), \epsilon$) ▷ Private Efficient Estimator T^*

  for each party P_j, run DSSA as in Algorithm 3 , $\mathbf{x}_j = (x_{j1}, ..., x_{jN})$) **do**

    $T_j^* \leftarrow CSSA(\mathbf{x}_j, \epsilon)$

  end for

  Compute the average: $\bar{T} = \frac{1}{M} \sum T_j^*$

  Compute $\hat{\theta}_{\text{MLE}}(\bar{T})$

  return $\hat{\theta}_{\text{MLE}}(\bar{T})$

Theorem 3.3 (sufficient-statistic-averaging): For M parties with datasets $\mathbf{x}_1, ..., \mathbf{x}_M$ respectively, the estimator $\hat{\theta}_{\text{MLE}}(\bar{T})$discribed in Algorithm 4 has MLE

$$J_{\hat{\theta}_{\text{MLE}}(\bar{T})}(\theta) = \frac{1}{MN}\left(\frac{1+o(1)}{I_f(\theta)} + \frac{1}{\sqrt{N}}\frac{\sqrt[4]{32}\Lambda_T(1+o(1))}{\sqrt{\epsilon I_f(\theta)}} + \frac{\sqrt{2}\Lambda_T}{\epsilon N}\right).$$

Proof. Let $\bar{T}$ be defined as algorithm 4, and $T = \frac{1}{MN}\sum_{i=1}^{MN} T(x_i)$ be the unperterbed sufficient statistcs. First, since T is sufficient, and by lemma 11, we have

$$J_{\hat{\theta}_{\text{MLE}}(T)}(\theta) = \mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(T) - \theta)^2] = \frac{1}{NM}\left(\frac{1+o(1)}{I_f(\theta)}\right).$$

Also, if range(T) is bounded by Λ_T, we can bound $\mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(\bar{T}) - \hat{\theta}_{\text{MLE}}(T))^2]$ by

$$\mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(\bar{T}) - \hat{\theta}_{\text{MLE}}(T))^2] \le \mathbb{E}_\theta[(\hat{\theta}'_{\text{MLE}}(\tilde{T})^2(T-\bar{T})^2], \text{ for some } \tilde{T} \in \text{ range } (T)$$

Also note that

$$\hat{\theta}'_{\text{MLE}} \le \text{var}(T) \le \Lambda_T,$$

and

$$\begin{aligned}\mathbb{E}_\theta[((T-\bar{T})^2] &= \text{var}_\theta[\frac{1}{M}\sum_{i=1}^{M} R_i] \text{ (note that } R_i \overset{iid}{\sim} \text{Lap}(\Lambda/\epsilon)) \\ &= \frac{1}{MN^2}\frac{\sqrt{2\Lambda_T}}{\epsilon}\end{aligned}$$

To upper bound $J_{\hat{\theta}_{\text{MLE}}(\bar{T})}(\theta)$, we have

$$\begin{aligned}\mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(\bar{T}) - \theta)^2] &= \mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(\bar{T}) - \hat{\theta}_{\text{MLE}}(T) + \hat{\theta}_{\text{MLE}}(T) - \theta)^2] \\ &= \mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(\bar{T}) - \hat{\theta}_{\text{MLE}}(T))^2] \\ &+ 2\mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(T) - \theta)(\hat{\theta}_{\text{MLE}}(\bar{T}) - \hat{\theta}_{\text{MLE}}(T))] + \mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(T) - \theta)^2] \\ &\le \mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(\bar{T}) - \hat{\theta}_{\text{MLE}}(T))^2] + \mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(T) - \theta)^2] \\ &+ 2\sqrt{\mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(T) - \theta)^2]\mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(\bar{T}) - \hat{\theta}_{\text{MLE}}(T))^2]} \text{ (by Cauchy's inequality)} \\ &\le \frac{1}{MN^2}\frac{\sqrt{2}\Lambda_T}{\epsilon} + 2\sqrt{\frac{1}{MN^2}\frac{\sqrt{2}\Lambda_T}{\epsilon}\cdot\frac{1}{MN}\left(\frac{1+o(1)}{I_f(\theta)}\right)} + \frac{1}{MN}\left(\frac{1+o(1)}{I_f(\theta)}\right) \\ &= \frac{1}{MN}\left(\frac{1+o(1)}{I_f(\theta)} + \frac{1}{\sqrt{N}}\frac{\sqrt[4]{32}\Lambda_T(1+o(1))}{\sqrt{\epsilon I_f(\theta)}} + \frac{\sqrt{2}\Lambda_T}{\epsilon N}\right).\end{aligned}$$

■

Corollary 3.2: If $\epsilon = o(N^{-1})$ and the number of data N each parties possess increase to infinity, the $\bar{T}$ given by DSSA is asymptotically efficient.

Proof. By theorem 12, the MSE of $\bar{T}$ is given as

$$J_{\hat{\theta}_{\text{MLE}}(T)}(\theta) = \mathbb{E}_\theta[(\hat{\theta}_{\text{MLE}}(T) - \theta)^2] = \frac{1}{NM}\left(\frac{1 + o(1)}{I_f(\theta)}\right).$$

If $\epsilon = o(N^{-1})$, then

$$J_{\hat{\theta}_{\text{MLE}}(T)}(\theta) = \frac{1}{MN}\left(\frac{1}{I_f(\theta)} + o_N(1)\right)$$

■

IV. Discussion and Related Work

We proposed a mechanism that releases an asymptotically efficient estimator of data collected from multiple party. Furthermore, we explore the relationship of the number of parties M, number of data N, and the privacy measure ϵ while using our mechanism. The MSE of the resulting estimator $\bar{T}$ writing in the form of Fisher information is clearly defined above.

Working on general data, where there is no assumption of data distribution, to output an efficient estimator with our mechanism, the number of parties should be kept under a certain scale related to the number of data retrieved from each party. Furthermore, the privacy level is proportional to the number of party and the number of data. In other words, with more parties participated and each contributes more data, the optimal privacy level is higher as ϵ getting smaller. Less but not least, if the data distribution is assumed to be in the exponential family, we may relax the constraint on the number of participating parties. Besides the relaxation, the privacy level now only concerns the number of data N.

Our mechanism facilitates real life situations such as schools providing students' records to state government or hospitals cooperates in nationwide project. Nevertheless, we sense the need of a hierarchical mechanism. For example, schools provide data to state government and state governments provide either data or calculated information to federal government. Furthermore, in this paper, the number of data contributed by each party is the same while it might not be in practice. These are some improvement we can make to our work.

References

[1] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," *IEEE Symposium on Security and Privacy*, pp. 111–125, 2008.

[2] C. Dwork, "Differential privacy," *Proceedings of the 33rd international conference on Automata, Languages and Programming*, pp. 1–12, 2006.

[3] A. Smith, "Efficient, differentially private point estimators," *arXiv:0809.4794*, 2008.

[4] J. Hamm, P. Cao, and M. Belkin, "Learning privately from multiparty data," *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, vol. 555-563, 2016.

[5] M. A. Pathak, S. Rane, and B. Raj, "Multiparty differential privacy via aggregation of locally trained classifiers," *Advances in Neural Information Processing Systems*, 2010.