Utility Analysis of Horizontally Merged Multi-Party Synthetic Data with Differential Privacy

Bingyue Su and Fang Liu
{bsu1, fliu2}@nd.edu
Applied and Computational Mathematics and Statistics, University of Notre Dame
Notre Dame, IN 46556

***Abstract*—A large amount of data is often needed to train machine learning algorithms with confidence. One way to achieve the necessary data volume is to share and combine data from multiple parties. On the other hand, how to protect sensitive personal information during data sharing is always a challenge. We focus on data sharing when parties have overlapping attributes but non-overlapping individuals. One approach to achieve privacy protection is through sharing differentially private synthetic data. Each party generates synthetic data at its own preferred privacy budget, which are then released and horizontally merged across the parties. The total privacy cost for this approach is capped at the maximum individual budget employed by a party. We derive the mean squared error bounds for the parameter estimation in common regression analysis based on the merged sanitized data across parties. We identify through theoretical analysis the conditions under which the utility of sharing and merging sanitized data outweighs the perturbation introduced for satisfying differential privacy and surpasses that based on individual party data. The experiments suggest that sanitized HOMM data obtained at a practically reasonable small privacy cost can lead to smaller prediction and estimation errors than individual parties, demonstrating benefits of data sharing while protecting privacy.**

***Index Terms*—data synthesis, differential privacy, multi-party, mean squared error, horizontally merged, regression, utility**

I. Introduction

A. Background and Motivation

Big data has enabled extensive and efficient applications of statistical models and machine learning algorithms to solve practical problems. One way to obtain the necessary data volume to train a model with confidence is to share and combine data across multiple parties; for example, patient data collected by different hospitals and education data collected from various sources and platforms. On the other hand, there are always privacy concerns when it comes to data sharing. Differential privacy (DP) is a popular concept for privacy protection in recent years[1]. In this paper, we focus on applying DP to sharing data among multiple parties, where each party has overlapping attributes but non-overlapping individuals with other parties. Being able to share and merge small data sets into large ones is important especially if the cost on data collection is high for each individual party and aggregating data across multi-parties becomes necessary to obtain a sizable data with improved analytical potentials.

This research is funded by NSF Award #1717417 and the University of Notre Dame Blockchain Project Grant.

978-1-7281-5628-6/20/$31.00 ©2020 IEEE

B. Our Approach and Contributions

We assume the individuals from different parties come from the same underlying population, so that it makes sense to share and combine the information from multiple sources. Information sharing can take different forms, as presented in the related work in Sec I-C, we focus on sharing and merging individual-level data horizontally with DP. Specifically, differentially private synthetic data are first generated given the original information in each party, and they are released and combined to obtain the HOrizontally Merged Multi-parties (HOMM) data.

The overall privacy cost for sharing and releasing the sanitized HOMM data is the maximum budget employed by a party across all the parties per the parallel composition principle [2] in DP, given that each party has non-overlapping individuals with other parties. This implies that the sanitized HOMM can involve as many parties as possible without having to increase the overall privacy cost. As more parties join the sharing efforts, the size of the sanitized HOMM data will grow, and the amount of information will outweigh the noise injected to achieve DP. As a result, the trained models and algorithms will become more efficient and robust, surpassing those based on the original individual-party data.

We derive the error bounds for parameter estimation and outcome prediction based on the sanitized HOMM data in regression settings. We examine how the relevant factors, including sample complexity and privacy budget, affects the errors, and the conditions under which the errors will be smaller than those based on the original individual-party data. Both our theoretical analysis and empirical studies suggest that the sanitized HOMM data helps to improve the robustness and stability of learning and estimation tasks in regression settings. To the best of our knowledge, our work provides the first in-depth look at the utility of sanitized HOMM data and the theoretical results on the error bounds on the queries results in the regression setting.

C. Related Work

An alternative way to combine information from multiple data-parallel parties while protecting privacy is to train models locally and then aggregate the local models to obtain a composite model in a differentially private manner [3, 4, 5, 6, 7, 8, 9]. Differentially private federated learning [10] falls under this category. The downside of this approach is that each training of an optimization algorithm or machine learning procedure on the original data will cost a certain amount of privacy, the cost accumulates per the sequential composition

theorem in DP [2] until all the pre-set privacy budget is used up and no more training is allowed. The sanitized HOMM data approach does not suffer from this as the data are individual-level and has the same structure as the original; data users can perform analysis on their own without having to worry about running out of privacy budget. This paper deals with data merging from data-parallel parties. Another line of work on merging information from multiple parties with DP focuses on vertical combination of feature-parallel data sets [11, 12, 13], where the parties possess non-overlapping sets of attributes but overlapping individuals.

II. Preliminaries

A. Definitions

Definition 1 (**ϵ-differential privacy (DP)** [1])**.** A randomization mechanism $\mathcal{M}$ is ϵ-differentially private if for all data sets D_1 and D_2 that differ by one element and all $S \in \text{Range}(\mathcal{M})$, $e^{-\epsilon} \leq \frac{\Pr[\mathcal{M}(D_1)\in S]}{\Pr[\mathcal{M}(D_2)\in S]} \leq e^{\epsilon}$.

Definition 2 (**parallel composition** [2])**.** Let $\{D_k\}_{k=1,\ldots,g}$ represent a sequence of disjoint subsets of data D. The *Parallel Composition* states that releasing the g query results from the sequence of randomization mechanisms $M_k\left(D \cap D_k\right)$ of ϵ_k for $k = 1, \ldots, g$ satisfies $\max_k \epsilon_k$-DP.

Definition 3 (**Laplace mechanism** [1])**.** For a given query f on data D, the Laplace mechanism that satisfies ϵ-differential privacy is defined as $\mathbf{f}^*(D)=\mathbf{f}(D)+\text{Lap}(\Delta_{\mathbf{f}}/\epsilon)$, where $\Delta_{\mathbf{f}} = \max_{D_1,D_2} \|\mathbf{f}(D_1) - \mathbf{f}(D_2)\|_1$ is the l_1 global sensitivity of query f, for all D_1, D_2 differing by one element.

B. Problem Setting

Suppose there are g parties; each owns a data set D_k for $k = 1, ..., g$. We assume all g data sets have at least one common attributes, and the individuals from different data sets are independent from each other; in other words, there is minimal overlapping information across different parties. Merging horizontally the g sets leads to a larger data set, providing opportunities for more efficient and robust training of learning algorithms with higher accuracy rates in prediction and more efficient statistical inferences. To mitigate privacy concerns, before data merging, each party generates synthetic data in a differentially private manner given the original data. Denote by ϵ_k the privacy budget party k chooses to sanitize its data, then the privacy cost of merging and sharing the g parties of data is $\epsilon = \max_k \epsilon_k$ per the parallel composition principle (Def 2).

While there are various approaches to generate differentially private synthetic data (model-based v.s. model-free, Bayesian v.s. Frequentist, etc; readers may refer to [14] for an overview if interested more on this topic), we focus on generating synthetic data from differentially private empirical distributions estimated by histograms, due to a couple of reasons. First, the approach is straightforward to implement; and second, differentially private histograms converge to the true underlying distributions at the rate of $O(n^{-2/(2+p)})$ [15], where n is the sample size and p is the dimension of the histogram. In this approach, party k ($k = 1, ..., g$) first forms a full-dimensional histogram over D_k with m_k bins with bin counts n_{kj} for $j = 1, \ldots, m_k$. It then sanitizes the histogram as in $\tilde{n}_{kj} = \max\{n_{kj} + e_{kj}, 0\}$, where $e_{kj} \sim \text{Lap}(0, \epsilon_k^{-1})$ independently. Finally, it draws $\tilde{n}_{kj}$ samples from $\text{Unif}(c_{kj,0}, c_{kj,1})$, where $c_{kj,0}, c_{kj,1}$ are the cutoff points that bound the j-th bin for $j = 1, \ldots, m_k$. The collection of the $\sum_j \tilde{n}_{kj}$ samples from all m_k bins makes the synthetic data $\tilde{D}_k$ in party k. After each party releases its synthetic data, they are combined horizontally to form one set of sanitized HOMM data.

The totality of the information contained in the sanitized HOMM data relates to its sample size $\tilde{n} = \sum_{k=1}^{g} \sum_{j=1}^{m_k} \tilde{n}_{kj}$, as well as the amount of injected noise to guarantee DP. As more parties join the sharing efforts, the accumulated information will eventually outweigh the amount of injected noise and surpass the non-private unshareable original information in at least some individual parties if not in all parties. In other words, we conjecture $\exists k' \in \{1, \ldots, g\}$ that $\tilde{I} \geq I_{(k')}$, where $\tilde{I}$ is the information contained in the sanitized HOMM data and $I_{(k')}$ is k'-th highest amount of information among the g individual parties, for a given g, $\epsilon > 0$, and the party size configuration $\mathbf{n} = (n_1, \ldots, n_g)$. To facilitate testing of the conjecture, we focus on some commonly seen analyses and learning procedures (generalized linear models including linear and logistic regression, kernel regression, and tree-based learning procedures) and present both the theoretical results and empirical conclusions in these setting. We will continue to explore ways to test the conjecture in general settings, such as using the information-theoretic framework.

III. Theoretical Analysis

In this section, we present the mean squared error (MSE) bounds based on the sanitized HOMM data for mean estimation and prediction, parameter estimation and prediction in linear regression and generalized linear models, and prediction from kernel regression with the Box kernel.

Denote the number of parties by g, and the number of histogram bins in party k by m_k; and $m = \sum_{k=1}^{g} m_k$ is the total number of histogram bins across all the parties. n denotes the sample size of the HOMM data, and n_{kj} is the original count in the j-th bin of the histogram in party k, $n_k = \sum_j n_{kj}$ is the data sample size in party k, and $n = \sum_{k,j} n_{kj}$ is the sample size of the original HOMM data. We assume that the Laplace mechanism (Def 3) is used to obtain the differentially private histograms in each party; that is, the bin size after sanitization is $\tilde{n}_{kj} = n_{kj} + e_{kj}$, where $e_{kj} \sim \text{Lap}(0, \epsilon^{-1})$, and $\tilde{n} = \sum_{k,j} \tilde{n}_{kj}$.

A. Expected Value

Theorem 1 (**MSE bound for sample mean**)**.** *Assume $Y \sim [\mu, \sigma^2]$ and $|Y| \leq B_Y$ (that is, B_Y is the global bound that does not depend on the local data). Let $\tilde{\mu}$ denote the sample mean, an estimate of μ, based on the sanitized HOMM data. MSE$_{\tilde{\mu}} = \mathbb{E}(\tilde{\mu} - \mu)^2 =$*

$$O\left(\left(\frac{6m}{\bar{\epsilon}} + \overline{n^2} + 2\overline{nn'}\right)\frac{B_Y^2}{n^2} + \frac{\sigma^2}{n}\right), \text{ where} \quad (1)$$

$$\overline{n^2}=\sum_{k,j}\frac{n_{kj}^2}{m_k^2};\ \overline{nn'}=\sum_k\sum_{j\neq j'}\frac{n_{kj}n_{kj'}}{m_k^2}+\sum_{k\neq k'}\sum_{j,j'}\frac{n_{kj}n_{k'j'}}{m_km_{k'}}. \quad (2)$$

$\bar{\epsilon}=m\left(\sum_{k=1}^{g}\frac{m_k}{\epsilon_k}\right)^{-1}$ *is the weighted harmonic mean of* $\{\epsilon\}_{j=1}^{g}$.

Proof. The estimates of μ based on the original and the sanitized HOMM data are respectively $\hat{y}=n^{-1}\sum_{k,j}n_{kj}\bar{y}_{kj}$ and $\tilde{y}=\tilde{n}^{-1}\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}$, where $\bar{y}_{kj}$ and $\tilde{y}_{kj}$ are the means of Y in bin j of the original and sanitized histograms in party k. The MSE of $\tilde{y}$ is

$$\mathbb{E}(\tilde{y}-\mu)^2\leq 2\mathbb{E}(\tilde{y}-\hat{y})^2+2\sigma^2/n. \quad (3)$$

$\mathbb{E}(\tilde{y}-\hat{y})^2$ in Eq (3) $=$

$$\mathbb{E}\left(\frac{\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}}{\tilde{n}}-\frac{\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}}{n}+\frac{\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}}{n}-\frac{\sum_{k,j}n_{kj}\bar{y}_{kj}}{n}\right)^2$$

$$\leq 2\mathbb{E}\left[\left(\frac{1}{\tilde{n}}-\frac{1}{n}\right)\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}\right]^2+2\mathbb{E}\left[\frac{\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}-\sum_{k,j}n_{kj}\bar{y}_{kj}}{n}\right]^2$$

$$=2\mathbb{E}\left[\left(\frac{n-\tilde{n}}{n}\right)\frac{\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}}{\tilde{n}}\right]^2+2\mathbb{E}\left[\frac{\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}-\sum_{k,j}n_{kj}\bar{y}_{kj}}{n}\right]^2$$

$$\leq\frac{2B_Y^2}{n^2}\sum_{i=1}^{g}\frac{m_i}{\epsilon_i}+2\mathbb{E}\left[\frac{\sum_{k,j}\tilde{n}_{kj}\tilde{y}_{kj}-\sum_{k,j}n_{kj}\bar{y}_{kj}}{n}\right]^2$$

$$=\frac{2mB_Y^2}{n^2\bar{\epsilon}}+\frac{2A}{n^2}+\frac{2B}{n^2},\ \text{where} \quad (4)$$

$A=\sum_{(k\neq k')\cup(j\neq j')}\mathbb{E}[\tilde{n}_{kj}\tilde{y}_{kj}-n_{kj}\bar{y}_{kj}]\mathbb{E}[\tilde{n}_{k'j'}\tilde{y}_{k'j'}-n_{k'j'}\bar{y}_{k'j'}]$ and $B=\sum_{k,j}\mathbb{E}[\tilde{n}_{kj}\tilde{y}_{kj}-n_{kj}\bar{y}_{kj}]^2$. The term $\mathbb{E}[\tilde{n}_{kj}\tilde{y}_{kj}-n_{kj}\bar{y}_{kj}]$ in A can be written as

$$\mathbb{E}\{\mathbb{E}[\tilde{n}_{kj}\tilde{y}_{kj}-n_{kj}\tilde{y}_{kj}|\tilde{n}_{kj}]\}+\mathbb{E}[n_{kj}\tilde{y}_{kj}-n_{kj}\bar{y}_{kj}]. \quad (5)$$

Since $\mathbb{E}(\tilde{y}_{kj}|\tilde{n}_{kj})=c_{kj}$, where $c_{kj}is$ a constant independent of the actual value of Y (as $\tilde{y}_{kj}$ is the mean of a set of samples from a uniform distribution with fixed bounds), Eq (5) becomes $c_{kj}\mathbb{E}\{\tilde{n}_{kj}-n_{kj}\}+\mathbb{E}[n_{kj}\tilde{y}_{kj}-n_{kj}\bar{y}_{kj}]=\mathbb{E}[n_{kj}\tilde{y}_{kj}-n_{kj}\bar{y}_{kj}]\leq\frac{n_{kj}B_Y}{m_k}$. Similarly, the term $\mathbb{E}[\tilde{n}_{k'j'}\tilde{y}_{k'j'}-n_{k'j'}\bar{y}_{k'j'}]$ in A is $\leq\frac{n_{k'j'}B_Y}{m_{k'}}$. Taken together,

$$A\leq B_Y^2\sum_{(k\neq k')\cup(j\neq j')}\frac{n_{kj}}{m_k}\frac{n_{k'j'}}{m_{k'}}=B_Y^2\overline{nn'},\ \text{and} \quad (6)$$

$$B=\sum_{k,j}\mathbb{E}[\tilde{n}_{kj}\tilde{y}_{kj}-n_{kj}\tilde{y}_{kj}+n_{kj}\tilde{y}_{kj}-n_{kj}\bar{y}_{kj}]^2$$

$$\leq 2\sum_{k,j}\mathbb{E}[(\tilde{n}_{kj}-n_{kj})^2\tilde{y}_{kj}^2+n_{kj}^2(\tilde{y}_{kj}-\bar{y}_{kj})^2]$$

$$\leq\frac{2mB_Y^2}{\bar{\epsilon}}+2B_Y^2\overline{n^2}. \quad (7)$$

Plugging Eqs (6) and (7) in Eq (4), we have $\mathbb{E}(\tilde{y}-\hat{y})^2\leq\frac{2mB_Y^2}{n^2\bar{\epsilon}}+\frac{2B_Y^2}{n^2}\overline{nn'}+\frac{4B_Y^2}{n^2}\left(\frac{m}{\bar{\epsilon}}+\overline{n^2}\right)=\frac{6mB_Y^2}{n^2\bar{\epsilon}}+\frac{2B_Y^2}{n^2}\overline{nn'}+\frac{4B_Y^2}{n^2}\overline{n^2}$, and Eq (3) becomes

$$\text{MSE}_{\tilde{\mu}}\leq\frac{12mB_Y^2}{n^2\bar{\epsilon}}+\frac{2B_Y^2}{n^2}\overline{nn'}+\frac{4B_Y^2}{n^2}\overline{n^2}+2\frac{\sigma^2}{n}$$

$$=O\left(\left(\frac{6m}{\bar{\epsilon}}+\overline{n^2}+2\overline{nn'}\right)\frac{B_Y^2}{n^2}+\frac{\sigma^2}{n}\right). \qquad \square$$

There are several interesting observations from Theorem 1.

- The utility of the sanitized HOMM data for estimating μ is determined by $\bar{\epsilon}$, the weighted harmonic mean of ϵ_k, whereas the overall privacy cost for merging and releasing the synthetic data from g parties is $\max_k\epsilon_k$. As a matter of fact, $\bar{\epsilon}$ tends strongly toward $\min_k\epsilon_k$, representing the largest amount of perturbation among all parties. In other words, the overall privacy cost relates to $\max_k\epsilon_k$, but the utility of the sanitized HOMM data is more influenced by $\min_k\epsilon_k\forall k$. When $\epsilon_k\equiv\epsilon$, both the privacy cost and the utility directly relate to ϵ.
- The MSE error bound in Eq (1) converges to σ^2n^{-1}, the MSE based on the original HOMM data as $\bar{\epsilon}\to\infty$ and $n_k\to\infty$ for $\forall k$. To see this, we notice that the MSE bound comprises 3 components: σ^2n^{-1}; $6mB_Y^2/(\bar{\epsilon}n^2)$, the error introduced due to the DP guarantee; and $\left(\overline{n^2}+2\overline{nn'}\right)B_Y^2n^{-2}$, representing the approximation error of using histograms in place of the known underlying distribution to sample data. As $\bar{\epsilon}\to\infty$, the error term due to privacy guarantee disappears. As $m_k^{-1}\to 0$ and $n_km_k^{-1}\to\infty$, the histogram in each party converges to the underlying distribution, and $\left(\overline{n^2}+2\overline{nn'}\right)n^{-2}\to 0$.
- When more parties join the sharing effort, the overall n increases, $\text{MSE}_{\tilde{\mu}}$ decreases and eventually becomes smaller than the individual MSE ($\sigma^2n_k^{-1}$ for party k) in at least some if not all parties. In other words, it is beneficial to share sanitized data compared to each individual party holding onto their own data, for the purposes of obtaining an estimate for μ with smaller MSE.

Theorem 1 does not impose a specific distribution family on Y, though it does have a global bound on Y, which is perfectly justifiable and in some cases necessary from a DP perspective. In the case of Gaussian Y, we may express the bound B_Y in term of σ and n, the formal result of which is presented in Corollary 1.

Corollary 1 (MSE bound of sample mean for Gaussian Y).

$$\text{MSE}_{\tilde{\mu}}=O\left(\left(\frac{6m}{\bar{\epsilon}}+\overline{n^2}+2\overline{nn'}\right)\left(\frac{\mu+\sigma\sqrt{2\log(n)}}{n}\right)^2+\frac{\sigma^2}{n}\right).$$

Proof. The proof uses the results from the following lemma to replace the global bound B_Y^2 with $\mathbb{E}(B_{\mathbf{y}}^2)$, where $\mathbf{y}$ represents the local data, and the rest of the proof is similar to that for Theorem 1.

Lemma 1. [16, 17]] If $X_1,...,X_n\overset{\text{i.i.d}}{\sim}N(0,\sigma^2)$, then $\mathbb{E}(X_{(n)})\leq\sigma\sqrt{2\log(n)}$, and $V(X_{(n)})\leq\frac{C\sigma^2}{\log(n)}$, where C is a constant independent of σ^2 or n.

Therefore, $\mathbb{E}(B_{\mathbf{y}}^2)=\mathbb{E}^2(B_{\mathbf{y}})+V(B_{\mathbf{y}})=\left(\mu+\sigma\sqrt{2\log(n)}\right)^2+\frac{C\sigma^2}{\log(n)}$, and $\text{MSE}_{\tilde{\mu}}$

$$=O\left(\left(\frac{6m}{\bar{\epsilon}}+\overline{n^2}+2\overline{nn'}\right)\frac{\left(\mu+\sigma\sqrt{2\log(n)}\right)^2+\frac{C\sigma^2}{\log(n)}}{n^2}+\frac{\sigma^2}{n}\right)$$

$$= O\left(\left(\frac{6m}{\bar{\epsilon}} + \overline{n^2} + 2\overline{nn'}\right) \frac{\left(\mu + \sigma\sqrt{2\log(n)}\right)^2}{n^2} + \frac{\sigma^2}{n}\right) \quad \square$$

B. Linear Regression

Linear models are commonly used regression models to quantify the effects of independent variables $\mathbf{X}$ on Gaussian outcome Y and to predict Y given $\mathbf{X}$: $Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + e_i$, where $e_i \sim N(0, \sigma^2)$. Let $\tilde{D} = (\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ denote the combined synthetic data from g parties generated from the differentially private histograms in each party. Theorem 2 shows the MSE bound for the least-squared (LS) estimates for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ in the linear model given $\tilde{D}$.

Theorem 2 (MSE bounds for LS regression coefficients). *Assume $|X_j| \le B_j$ for $j = 1, \ldots, p$, and party k has privacy budget ϵ_k. The MSE of the LS estimate $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}$ of $\boldsymbol{\beta}$ based on the sanitized HOMM data is*

$$MSE_{\tilde{\boldsymbol{\beta}},k} = \mathbb{E}||(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{\beta}||^2 = O\left(\frac{B_X^2}{n^2}\left(\frac{m}{\bar{\epsilon}} + \frac{\overline{n^2}}{m} + \overline{nn'}\right)\left(\gamma^2 + \sigma^2\log(n)\right)^2 + \frac{\sigma^2}{n}\right), \quad (8)$$

where $\gamma = \sum_{j=1}^{p}\beta_j B_j$, $B_X^2 = \sum_{j=1}^{p} B_j^2$, $\overline{n^2} = \sum_{k,j}\frac{n_{kj}^2}{m_k^2}$, and the other notations are the same as in Theorem 1.

Proof. $\mathbb{E}||(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{\beta}||^2 \le$

$$2\mathbb{E}||(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}||^2 + 2\mathbb{E}||(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y} - \boldsymbol{\beta}||^2$$
$$\le 4A + 4B + +2\sigma^2\mathrm{tr}((\boldsymbol{x}^T\boldsymbol{x})^{-1}), \quad (9)$$

where $A = \mathbb{E}||(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}||^2$ and $B = \mathbb{E}||(\boldsymbol{x}^T\boldsymbol{x})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}||^2$. We bound A and B, respectively. First, $A =$

$$\mathbb{E}\left\{[(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}]^T[(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}]\right\}$$
$$= \mathbb{E}\left\{\mathrm{tr}\left([(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}]\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}\right)^T\left([(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}]\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}\right)\right\}$$
$$= \mathbb{E}\left\{\mathrm{tr}\left(\boxed{[(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}]^T[(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}]}\,\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{x}}\right)\right\}.$$

If matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are symmetric and positive definite, then $\mathrm{tr}(\boldsymbol{M}_1\boldsymbol{M}_2) \le \lambda_1(\boldsymbol{M}_1)\mathrm{tr}(\boldsymbol{M}_2)$ [18], where $\lambda_1(\boldsymbol{M}_1)$ is the largest eigenvalue of $\boldsymbol{M}_1$. Now let $\boldsymbol{M}_1$ be the boxed term above and $\boldsymbol{M}_2$ be $\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{x}}$, then $A \le \mathbb{E}\left\{\lambda_1(\boldsymbol{M}_1)\mathrm{tr}(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{x}})\right\} = \mathbb{E}\left\{\tilde{n}^2\lambda_1(\boldsymbol{M}_1)\frac{\mathrm{tr}(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{x}})}{\tilde{n}^2}\right\}$. λ_1 associated with $\boldsymbol{M}_1$ is equal to the square of the maximum eigenvalue of $(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}$. Also noted is that

$$2\sigma^2\mathrm{tr}((\boldsymbol{x}^T\boldsymbol{x})^{-1}) = O(n^{-1}\sigma^2). \quad (10)$$

Since every element in $(\boldsymbol{x}^T\boldsymbol{x})^{-1}$ is $O(n^{-1})$, then $\lambda_1 = O\left(\frac{1}{\tilde{n}} - \frac{1}{n}\right)^2$. Therefore, $\mathbb{E}(\tilde{n}^2\lambda_1) = O\left(\mathbb{E}\left(\frac{\tilde{n}-n}{n}\right)^2\right) = O\left(\frac{m}{n^2\bar{\epsilon}}\right)$ and $\mathbb{E}\left(\frac{\mathrm{tr}(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{x}})}{\tilde{n}^2}\right) = \frac{1}{\tilde{n}^2}\mathbb{E}\left(\sum_{j=1}^{p}(\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{x}}_j)^2\right) = \sum_{j=1}^{p}\mathbb{E}\left(\frac{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{x}}_j}{\tilde{n}}\right)^2 \le B_X^2\mathbb{E}(B_Y^2) = O\left(\left(\gamma + \sigma\sqrt{2\log(n)}\right)^2 B_X^2\right)$, where $B_X^2 = \sum_{j=1}^{p} B_j^2$ and $\gamma = \sum_{j=1}^{p}\beta_j B_j$. Taken together,

$$A = O\left(\frac{m}{n^2\bar{\epsilon}}\sum_{j=1}^{p} B_j^2\left(\gamma + \sigma\sqrt{2\log(n)}\right)^2\right). \quad (11)$$

The bound on B in Eq (9) can be derived in a similar manner as on A. Specifically, $B =$

$$\mathbb{E}\left\{[(\boldsymbol{x}^T\boldsymbol{x})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}]^T[(\boldsymbol{x}^T\boldsymbol{x})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}]\right\}$$
$$= \mathbb{E}\{\mathrm{tr}[(\boldsymbol{x}^T\boldsymbol{x})^{-2}(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{x}^T\boldsymbol{y})(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{x}^T\boldsymbol{y})^T]\}$$
$$\le \mathbb{E}\{\lambda_1\mathrm{tr}((\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{x}^T\boldsymbol{y})(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{x}^T\boldsymbol{y})^T)\},$$

where $\lambda_1((\boldsymbol{x}^T\boldsymbol{x})^{-2})$ is the largest eigenvalue of $(\boldsymbol{x}^T\boldsymbol{x})^{-2}$. Since every element in $(\boldsymbol{x}^T\boldsymbol{x})^{-2}$ is $O(n^{-2})$, $\lambda_1((\boldsymbol{x}^T\boldsymbol{x})^{-2}) \sim O(n^{-2})$, and $|(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{x}^T\boldsymbol{y})^T(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{x}^T\boldsymbol{y})| = \sum_{i=1}^{p}(\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{x}}_j - \boldsymbol{y}^T\boldsymbol{x}_j)^2$, thus

$$B \sim O\left(B_X^2\left(\frac{m}{n^2\bar{\epsilon}} + \frac{\overline{n^2}}{n^2m} + \frac{\overline{nn'}}{n^2}\right)\left(\gamma + \sigma\sqrt{2\log(n)}\right)^2\right). \quad (12)$$

Eqs (10), (11), and (12) taken together, the MSE in Eq (9)

$$\le 4A + 4B + 2\sigma^2\mathrm{tr}((X^TX)^{-1})$$
$$= O\left(\left(\frac{m}{n^2\bar{\epsilon}} + \frac{\overline{n^2}}{n^2m} + \frac{\overline{nn'}}{n^2}\right)B_X^2\left(\gamma^2 + \sigma^2\log(n)\right)^2 + \frac{\sigma^2}{n}\right). \square$$

The MSE values of the LS estimates of $\boldsymbol{\beta}$ based on the original HOMM data and the data from an individual party k are

$$\mathrm{MSE}_{\beta} = \sigma^2\mathrm{diag}(\boldsymbol{x}^T\boldsymbol{x})^{-1} = \sigma^2\mathrm{tr}(\boldsymbol{x}^T\boldsymbol{x})^{-1} = O(n^{-1}\sigma^2), \quad (13)$$
$$\mathrm{MSE}_{\beta,k} = \sigma^2\sum\mathrm{diag}_k(\boldsymbol{x}_k^T\boldsymbol{x}_k)^{-1} = \sigma^2\mathrm{tr}(\boldsymbol{x}_k^T\boldsymbol{x}_k)^{-1} = O(n_k^{-1}\sigma^2).$$

Similar to the expected value case in Sec III-A, the MSE based on the sanitized HOMM data in Eq (8) has two additional terms compared to Eq (13) due to the sanitization error and the histogram approximation error. As $\bar{\epsilon} \to \infty$, and $m_k^{-1} \to 0$ and $n_k m_k^{-1} \to \infty\ \forall k$, the MSE in Eq (8) converges to the MSE in Eq (13). Compared to the per-party MSE, as more parties join the sharing effort, the MSE based on the sanitized HOMM data will keep decreasing and become smaller than the per-party MSE for more and more individual parties.

Theorem 2 can be easily extended when a global bound B_Y on Y is used in lieu of the expected local bound.

Corollary 2 (MSE bounds for linear regression coefficients with global bound on Y). Provided that $|Y| \le B_Y$, whereas all the other settings are the same as in Theorem 2, then

$$\mathbb{E}||(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}})^{-1}\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{y}} - \boldsymbol{\beta}||^2 = O\left(\left(\frac{m}{n^2\bar{\epsilon}} + \frac{\overline{n^2}}{n^2m} + \frac{\overline{nn'}}{n^2}\right)B_X^2B_Y^2 + \frac{\sigma^2}{n}\right). \quad (14)$$

The error bound for the predicted $\hat{y}_h$ at a given set of predictor $\boldsymbol{x}_h$ can also be easily derived from Theorem 2.

Corollary 3 (error bound on prediction). Given a set of predictors $\boldsymbol{x}_h$, the MSE of the predicted outcome $\tilde{y}_h$ is

$$\mathbb{E}(\tilde{y}_h - \mathbb{E}(Y_h))^2 = ||\boldsymbol{x}_h||^2 O\left(\left(\frac{m}{n^2\bar{\epsilon}} + \frac{\overline{n^2}}{n^2m} + \frac{\overline{nn'}}{n^2}\right)B_X^2(*) + \frac{\sigma^2}{n}\right).$$

where $(*) = \left(\gamma^2 + \sigma^2\log(n)\right)^2$ if the expected local bound on Gaussian $\mathbf{y}$ is used; $(*) = B_Y^2$ if a global bound is used.

C. Kernel Regression

Kernel regression is a non-parametric technique that estimates nonlinear relation f between predictor X and outcome Y: $Y = f(X) + \epsilon$, where $\epsilon \overset{i.i.d}{\sim} [0, \sigma^2]$. Given a data set (x_i, y_i) for $i = 1, \ldots, n$, if the Box kernel is employed with bandwidth h, then $f(x)$ at any point x can be estimated as

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} \mathbb{I}(|x_i - x| \le h) y_i}{\sum_{i=1}^{n} \mathbb{I}(|x_i - x| \le h)} = \sum_{i=1}^{n} w_i(x) y_i. \quad (15)$$

Theorem 3 derives the MSE bound on $\hat{f}(x)$ in Eq (15) based on the sanitized HOMM data.

Theorem 3 (MSE bound on predicted Y from Box-kernel regression). *WLOG, Assume $x \in [0, B_X]$ and that $f(x)$ is continuous and is locally bounded. The MSE of the estimator $\tilde{f}(x)$ for a given x in Eq (15) based on the sanitized HOMM data is*

$$\mathbb{E}(\tilde{f}(x) - f(x))^2 \le O\left(\frac{\sigma^2}{n_0} + \left(\frac{m}{n_0^2 \bar{\epsilon}} + \frac{\overline{n_0^2}}{n_0^2 m} + \frac{\overline{n_0 n_0'}}{n_0^2}\right) B_w^2 B_Y^2\right),$$

where $n_0 = B_X^{-1} 2nh$ for uniformly distributed $X \in [0, B_X]$.

Proof. Denote the set of observations in the neighborhood of $(x-h, +h)$ by $S(x)$ for a given x, the size of which is $2\tilde{n}h/B_X$ and $2n_k h/B_X$ in the sanitized HOMM data and in the data of party k, respectively.

$$\mathbb{E}(f(x) - \tilde{f}(x))^2 \le 2\mathbb{E}(f(x) - \hat{f}(x))^2 + 2\mathbb{E}(\hat{f}(x) - \tilde{f}(x))^2$$
$$= \frac{2\sigma^2}{n_0} + \left(f(x) - \frac{2}{n_0} \sum_{i \in S(x)} f(x_i)\right)^2 + 2\mathbb{E}(\hat{f}(x) - \tilde{f}(x))^2. \quad (16)$$

A comparison between Eq (16) and Eq (3) suggests that the former has one additional term $(f(x) - 2\sum_{i \in S(x)} f(x_i)/n_0)^2$, which is a constant and ignorable as $f(x)$ is continuous and locally bounded. The bound on the second term $\mathbb{E}(\hat{f}(x) - \tilde{f}(x))^2$ in Eq (16) can be derived in a similar manner as in the linear regression case by noting that $\hat{f}(x) = \sum_i w_i(x) y_i$ and $\tilde{f}(x) = \sum_i w_i(\tilde{x}) y_i$ per Eq (15), both of which are linear functions of y as in the linear regression case. □

D. Generalized Linear Model

Generalized linear models (GLM) are a widely used regression model family to quantify the association between predictors $\boldsymbol{X}$ and non-Gaussian outcome Y and make prediction on Y. They include some of the most popular regression models such as linear, logistic, and poisson regression. With the canonical link, Y follows an exponential family $f(y_i) \propto exp(y_i \eta_i - b(\eta_i) + h(y_i))$, for $i = 1, \ldots, n$, where $\eta_i = \mathbf{x}_i \boldsymbol{\beta}$ is the natural parameter. Parameters $\boldsymbol{\beta}$ in GLM are often estimated through maximum likelihood estimation (MLE) by minimizing the negative log-likelihood function (the loss function). In what follows, we show the expected l_2 distance in the loss function based on the sanitized vs the original HOMM data converges to 0 faster than based on the individual party data vs the original HOMM data; and subsequently, the MLE based on the sanitized HOMM data approaches true parameters $\boldsymbol{\beta}$ faster since the loss function is convex $\boldsymbol{\beta}$ and the MLE based on the original HOMM data is $\sqrt{n}$-consistent for $\boldsymbol{\beta}$.

Theorem 4 (convergence rate of GLM loss function). *Assume $|Y| \le B_Y$ and $|x_j| \le B_j$. Define $B_{X^4} = \sum_{j=1}^{p} B_j^4$, $B_{X\beta} = \sum_{j=1}^{p} \beta_j B_j$, and $D = \sum_{j=1}^{p} \beta_j^2 B_j^2$. Let $\lambda_1(\boldsymbol{\beta}\boldsymbol{\beta}^T)$ denote the largest eigenvalue of $\boldsymbol{\beta}\boldsymbol{\beta}^T$. The expected squared difference of the loss function (negative log-likelihood) based on the sanitized vs. the original HOMM data is*

$$O\Big(n^{-2}\Big[\Big(6m\bar{\epsilon}^{-1} + \overline{n^2} + 2\overline{nn'}\Big) pDB_Y^2 + \lambda_1^2(\boldsymbol{\beta}\boldsymbol{\beta}^T)\Big(6mB_{X^4}\bar{\epsilon}^{-1} + B_{X^4}\overline{nn'} + 2B_{X^4}\overline{n^2}\Big)\Big]\Big), \quad (17)$$

Proof. WLOG, we assume $\mathbf{X}$ is centralized. The first-order Taylor expansion of the loss functions around $\eta_i (:= \mathbf{x}_i \boldsymbol{\beta}) = 0$ based on the sanitized and the original HOMM data are, respectively,

$$l(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y}) = \frac{-1}{n}\left[C + \boldsymbol{\beta}^T \boldsymbol{x}^T \boldsymbol{y} - b'(0) \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\beta}\right] + O\left(\frac{1}{n}\|\boldsymbol{x}\boldsymbol{\beta}\|^2\right), \quad (18)$$

$$l(\boldsymbol{\beta}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \frac{-1}{\tilde{n}}\left[\tilde{C} + \boldsymbol{\beta}^T \tilde{\boldsymbol{x}}^T \tilde{\boldsymbol{y}} - b'(0) \sum_{i=1}^{\tilde{n}} \tilde{\boldsymbol{x}}_i \boldsymbol{\beta}\right] + O\left(\frac{1}{\tilde{n}}\|\tilde{\boldsymbol{x}}\boldsymbol{\beta}\|^2\right), \quad (19)$$

where $C = \sum_{i=1}^{n} h(y_i) - b(0)$ and $\tilde{C} = \sum_{i=1}^{\tilde{n}} h(\tilde{y}_i) - b(0)$ are independent of $\boldsymbol{\beta}$ and irrelevant for the optimization of the loss functions. First, given $\frac{1}{n}\boldsymbol{\beta}^T \boldsymbol{x}^T \boldsymbol{y} = \frac{1}{n}\sum_{j=1}^{p} \boldsymbol{x}_j^T \boldsymbol{y} \beta_j$, then $\frac{1}{\tilde{n}}\boldsymbol{\beta}^T \tilde{\boldsymbol{x}}^T \tilde{\boldsymbol{y}} - \frac{1}{n}\boldsymbol{\beta}^T \boldsymbol{x}^T \boldsymbol{y} = \sum_{j=1}^{p} (\frac{1}{\tilde{n}} \tilde{\boldsymbol{x}}_j^T \tilde{\boldsymbol{y}} - \frac{1}{n} \boldsymbol{x}_j^T \boldsymbol{y}) \beta_j$, $\mathbb{E}\left(\frac{1}{\tilde{n}} \tilde{\boldsymbol{x}}_j^T \tilde{\boldsymbol{y}} - \frac{1}{n} \boldsymbol{x}_j^T \boldsymbol{y}\right)^2 = O\left(\left(\frac{6m}{\epsilon} + \overline{n^2} + 2\overline{nn'}\right)\frac{B_j^2 B_Y^2}{n^2}\right)$, and thus

$$\mathbb{E}||\tilde{n}^{-1} \boldsymbol{\beta}^T \tilde{\boldsymbol{x}}^T \tilde{\boldsymbol{y}} - n^{-1} \boldsymbol{\beta}^T \boldsymbol{x}^T \boldsymbol{y}||^2 \le p \sum_{j=1}^{p} \mathbb{E}\big(\frac{1}{\tilde{n}} \tilde{\boldsymbol{x}}_j^T \tilde{\boldsymbol{y}} - \frac{1}{n} \boldsymbol{x}_j^T \boldsymbol{y}\big)^2$$
$$= O\left(\left(6m\epsilon^{-1} + \overline{n^2} + 2\overline{nn'}\right) pDB_Y^2 n^{-2}\right), \quad (20)$$

where $D = \sum_{j=1}^{m} \beta_j^2 B_j^2$. Second, given $\sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} x_{ij}$, then

$$\mathbb{E}\left\|\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{\boldsymbol{x}}_{\boldsymbol{i}} \boldsymbol{\beta} - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x_i} \boldsymbol{\beta}\right\|^2 = O\left(\left(\frac{6m}{\epsilon} + \overline{n^2} + 2\overline{nn'}\right)\frac{pD}{n^2}\right). \quad (21)$$

Third, $O(\tilde{n}^{-1}\|\tilde{\boldsymbol{x}}\boldsymbol{\beta}\|^2 - n^{-1}\|\boldsymbol{x}\boldsymbol{\beta}\|^2) = O(n^{-1}\lambda_1(\boldsymbol{\beta}\boldsymbol{\beta}^T) tr(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}} - \boldsymbol{x}^T\boldsymbol{x}) + (\tilde{n}^{-1} + n^{-1})\|\tilde{\boldsymbol{x}}\boldsymbol{\beta}\|^2)$. Given $\text{tr}(\tilde{\boldsymbol{x}}^T\tilde{\boldsymbol{x}} - \boldsymbol{x}^T\boldsymbol{x}) = \sum_{j=1}^{p} \tilde{\boldsymbol{x}}_{\boldsymbol{j}}^T \tilde{\boldsymbol{x}}_{\boldsymbol{j}} - \boldsymbol{x}_{\boldsymbol{j}}^T \boldsymbol{x}_{\boldsymbol{j}}$, and following the proof in the expected value case, we have

$$O(\mathbb{E}(\tilde{n}^{-1}\|\tilde{\boldsymbol{x}}\boldsymbol{\beta}\|^2 - n^{-1}\|\boldsymbol{x}\boldsymbol{\beta}\|^2)^2)$$
$$= O\left(\lambda_1^2(\boldsymbol{\beta}\boldsymbol{\beta}^T)\left(\frac{6mB_{X^4}}{n^2\epsilon} + \frac{B_{X^4}}{n^2}\overline{nn'} + \frac{2B_{X^4}}{n^2}\overline{n^2}\right)\right), \quad (22)$$

where $B_{X^4} = \sum_{j=1}^{p} B_j^4$, and $B_{X\beta} = \sum_{j=1}^{p} \beta_j B_j$. Eqs (20) to (22) taken together, the difference between the loss functions in Eqs (19) and (18) can be written as

$$\mathbb{E}(l(\boldsymbol{\beta}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) - l(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y}))^2 = O\left(\frac{1}{n^2}\left[\left(6m\frac{1}{\bar{\epsilon}} + \overline{n^2} + 2\overline{nn'}\right) pDB_Y^2 + \lambda_1^2(\boldsymbol{\beta}\boldsymbol{\beta}^T)\left(6mB_{X^4}\bar{\epsilon}^{-1} + B_{X^4}\overline{nn'} + 2B_{X^4}\overline{n^2}\right)\right]\right). \quad \square$$

experiment	data attribute	g	party size	learning task; performance metric	# of repeats
1 (simulated data)	Gaussian $Y \sim N(\mu, 1)$	5	20 $\sim$ 150 (total $n = 370$)	estimation of μ; root mean squared error (RMSE) of $\hat{\mu} = \bar{Y}$	200
2 (simulated data)	Gaussian $Y = \boldsymbol{\beta}\mathbf{X} + \epsilon$, where $\epsilon \sim N(0, 3)$, $\mathbf{X} = (1, X_1, X_2)$, $X_1 \sim N(0, 1)$, $X_2 \sim$ Bern(0.5)	10	50 $\sim$ 275 (total $n = 1000$)	prediction of Y via linear model; prediction RMSE on testing data ($n = 100$)	200
3 (simulated data)	Binary $Y \sim$ Bernoulli$(\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})})$, where $\mathbf{X} = (1, X_1, X_2)$ with $X_1 \sim N(0, 1)$, $X_2 \sim$ Bern(0.5)	10	50 $\sim$ 80 (total $n = 650$)	classification via logistic regression; mis-classification rate of Y on testing data ($n = 100$)	200
4 (medical costs data)[19]	Numerical Y (charges); $\mathbf{X}_{6\times1}$	13	30 $\sim$ 300 (total $n = 900$)	prediction of Y via linear model and via random forest; prediction RMSE on testing data ($n = 438$)	100
5 (social network ads data) [20]	Binary Y (whether purchased products); $\mathbf{X}_{3\times1}$	9	20 $\sim$ 50 (total $n = 300$)	classification via logistic regression; mis-classification rate of Y on testing data ($n = 100$)	100

TABLE I
EXPERIMENT SET-UP

The expected squared difference in the loss function based on the individual party k of size n_k vs the original HOMM data $\mathbb{E}(l(\boldsymbol{\beta}|\boldsymbol{x}_k, \boldsymbol{y}_k) - l(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y}))^2$ can be derived in a similar manner. Specifically, $||n_k^{-1}\boldsymbol{\beta}^T\boldsymbol{x}^T\boldsymbol{y} - n^{-1}\boldsymbol{\beta}^T\boldsymbol{x}^T\boldsymbol{y}||^2 = O(n^{-2}(n-n_k)^2 D B_Y^2)$; $\left|\left|{n_k}^{-1}\sum_{i=1}^{n_k}\boldsymbol{x_i}\boldsymbol{\beta} - n^{-1}\sum_{i=1}^{n}\boldsymbol{x_i}\boldsymbol{\beta}\right|\right|^2 = O(n^{-2}(n-n_k)^2 D)$; $O((n_k^{-1}\|\boldsymbol{x}\boldsymbol{\beta}\|^2 - n^{-1}\|\boldsymbol{x}\boldsymbol{\beta}\|^2)^2) = O(n^{-2}(n-n_k)^2 D^2)$. So $\mathbb{E}(l(\boldsymbol{\beta}|\boldsymbol{x}_k, \boldsymbol{y}_k) - l(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y}))^2 = O(n^{-2}(n-n_k)^2 D^2 B_Y^2)$, which converges to zero at a slower rate compared with with Eq (17) for small n_k.

IV. EXPERIMENTS

We conducted five experiments in simulated and real-life data, examined a range of per-party privacy budget ϵ (set to be the same across the individual parties). We used the Laplace mechanism to perturb the histograms. In the experiments, we also explored different g and various configurations of party sizes. We expect the results from the experiments are sufficient to illustrate the utility of sanitized HOMM data. We will explore the heterogeneous privacy budget case and sanitization via other sanitation mechanisms or with relaxed versions of DP [21, 22]) (e.g., the Gaussian mechanism [23, 24]) in a future extension of this paper.

The set-up of the 5 experiments are listed in Table I. All experiments were run in R on a Macbook Pro with 2.9 GHz Intel Core i7 and 16 GB of memory. The results from the five experiments are presented in Fig. 1. If the solid black curve based on the sanitized HOMM data at a given privacy budget ϵ is lower than those from some individual parties (solid pink lines), then it implies that sharing the sanitized HOMM data benefits these individual parties in inferences or predictions; if it is lower than the average (dashed pink line), then it implies that sharing sanitized HOMM data is more beneficial for inferences or predictions than not sharing on average. In general, the RMSE based on the sanitized HOMM data is smaller than those from at least some individual parties at some practically reasonable small ϵ, demonstrating benefits of sharing data across parties at only a small privacy cost. The smallest ϵ achieving that depends on the specific data and learning tasks. In experiment 1, the RMSE based on the sanitized HOMM data is smaller than the smallest party ($n = 20$) RMSE around $\epsilon = 0.08$, the averaged individual-party RMSE around $\epsilon = 0.2$, and the largest party ($n = 150$) RMSE around $\epsilon = 0.5$. In experiment 2, when $\epsilon \approx 4.5$, the sanitized HOMM data yield a similar RMSE as the original HOMM and smaller than the averaged RMSE from the original individual parties at $\epsilon > 1$. In experiment 3, when $\epsilon \geq 1$, the sanitized HOMM data lead to a lower mis-classification rate than the best rate among the individual parties, and approach the rate from the original HOMM data around $\epsilon > 4.5$. In experiment 4, due to the higher dimensionality of the data (more attributes) compared to experiments 1 to 3, the sanitized HOMM data beats the individual parties in prediction RMSE at a relatively larger ϵ. In experiment 5, the error rates based on the sanitized HOMM data are smaller than the average based on the individual parties when $\epsilon > 2$ and remain fairly constant around 13% for $\epsilon \geq 3.5$.

V. SUMMARY AND DISCUSSION

We have examined the MSE bounds based on sanitized HOMM data in various estimation and regression tasks. We have identified the factors that are associated with the error bounds and the conditions under which bounds will be smaller than those based on the original individual party data. Overall, the results based on our theoretical and empirical utility analysis encourage data sharing in a differentially private manner, especially when there are a large number of small to middle-sized parties, so to achieve higher accuracy and more statistical efficiency in trained models and learning procedures. For future research, we will look into the utility analysis of HOMM data generated via other synthesis approaches than from histograms, and applications to high-dimensional data to examine the scalability of the utility based on sanitized HOMM data. Another interesting topic is the utility of vertically merged synthetic data (feature-parallel) across parties, which can be a challenging topic if the attributes are correlated.

REFERENCES

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[2] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the*

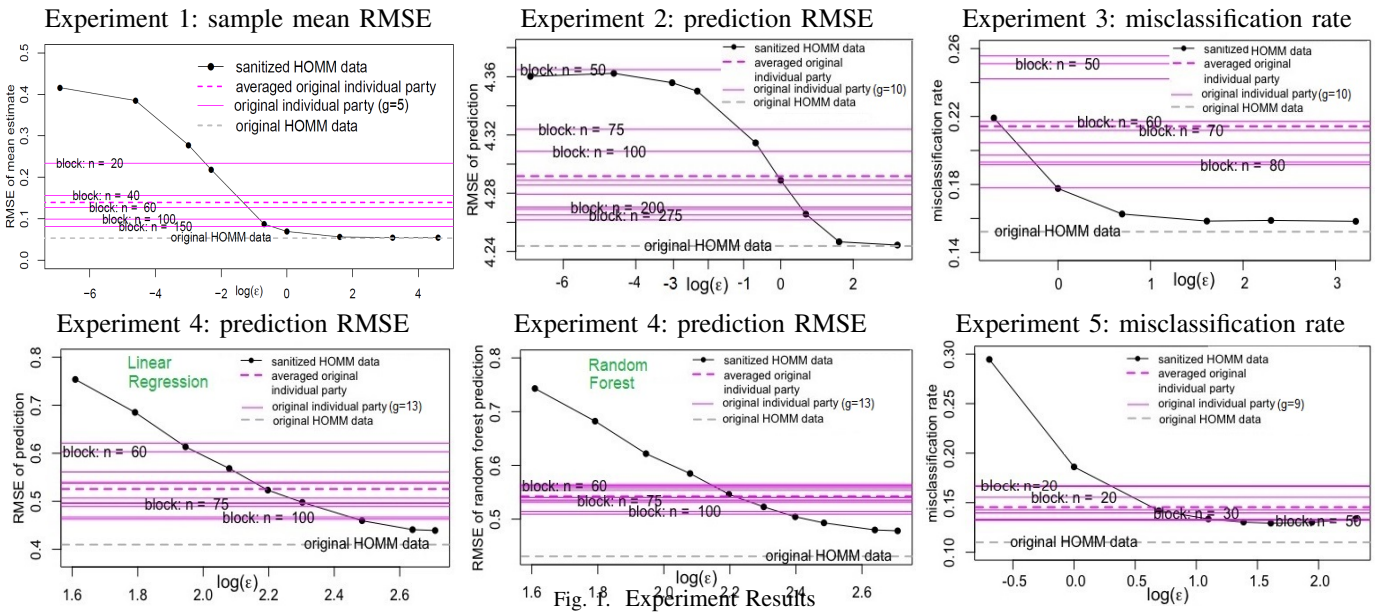

Fig. 1. Experiment Results

2009 ACM SIGMOD International Conference on Management of data. ACM, 2009, pp. 19–30.

[3] M. Pathak, S. Rane, and B. Raj, "Multiparty differential privacy via aggregation of locally trained classifiers," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1876–1884.

[4] J. Hamm, Y. Cao, and M. Belkin, "Learning privately from multiparty data," in *International Conference on Machine Learning*, 2016, pp. 555–563.

[5] M. Heikkilä, E. Lagerspetz, S. Kaski, K. Shimizu, S. Tarkoma, and A. Honkela, "Differentially private bayesian learning on distributed data," in *Advances in neural information processing systems*, 2017, pp. 3226–3235.

[6] T. Zhang and Q. Zhu, "A dual perturbation approach for differential private admm-based distributed empirical risk minimization," in *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*, 2016, p. 129–137.

[7] ——, "Dynamic differential privacy for admm-based distributed classification learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 31, p. 172–187, 2017.

[8] C. Zhang, M. Ahmad, and Y. Wang, "Admm based privacy-preserving decentralized optimization," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, p. 565–580, 2019.

[9] X. Zhang, M. Khalili, and M. Liu, "Improving the privacy and accuracy of admm-based distributed algorithms," in *International Conference on Machine Learning*, 2018, p. 5791–5800.

[10] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[11] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, p. 211–407, 2014.

[12] Y. Hu, D. Niu, J. Yang, and S. Zhou, "Fdml: A collaborative machine learning framework for distributed features," in *Proceedings of KDD '19*. ACM, 2019, pp. 528–544.

[13] Y. Hu, P. Liu, L. Kong, and D. Niu, "Learning privately over distributed features: An admm sharing approach," *arXiv preprint arXiv:1907.07735*, 2019.

[14] C. M. Bowen and F. Liu, "Comparative study of differentially private data synthesis methods," *Statistical Science*, 2020.

[15] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 375–389, 2010.

[16] G. Kamath, "Bounds on the expectation of the maximum of samples from a gaussian," *URL http://www. gautamkamath. com/writings/gaussian max. pdf*, 2015.

[17] S. Boucheron, M. Thomas *et al.*, "Concentration inequalities for order statistics," *Electronic Communications in Probability*, vol. 17, 2012.

[18] Y. Fang, K. A. Loparo, and X. Feng, "Inequalities for the trace of matrix product," *IEEE Transactions on Automatic Control*, vol. 39, no. 12, pp. 2489–2490, 1994.

[19] "Medical cost personal datasets: Insurance forecast by using linear regression," https://www.kaggle.com/mirichoi0218/insurance (accessed in Jan 2020).

[20] "Social network ads: A categorical dataset to determine whether a user purchased a particular product," https://www.kaggle.com/rakeshrau/social-network-ads (accessed in Jan 2020).

[21] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distriuted noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[22] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," *IEEE ICDE IEEE 24th International Conference*, pp. 277 – 286, 2008.

[23] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Theoretical Computer Science*, vol. 9, pp. 211–407, 2013.

[24] F. Liu, "Generalized gaussian mechanism for differential privacy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, pp. 747– 756, 2019.