Differentially Private Multi-Site Treatment Effect Estimation

Tatsuki Koga
Dept. of Computer Science and Engineering
University of California, San Diego
La Jolla, CA, USA
tkoga@ucsd.edu

Kamalika Chaudhuri
Dept. of Computer Science and Engineering
University of California, San Diego
La Jolla, CA, USA
kamalika@cs.ucsd.edu

David Page
Dept. of Biostatistics
Duke University
Durham, NC, USA
david.page@duke.edu

***Abstract*—Patient privacy is a major barrier to healthcare AI. For confidentiality reasons, most patient data remains in silo in separate hospitals, preventing the design of data-driven healthcare AI systems that need large volumes of patient data to make effective decisions. A solution to this is collective learning across multiple sites through federated learning with differential privacy. However, literature in this space typically focuses on differentially private statistical estimation and machine learning, which is different from the causal inference-related problems that arise in healthcare. In this work, we take a fresh look at federated learning with a focus on causal inference; specifically, we look at estimating the average treatment effect (ATE), an important task in causal inference for healthcare applications, and provide a federated analytics approach to enable ATE estimation across multiple sites along with differential privacy (DP) guarantees at each site. The main challenge comes from site heterogeneity—different sites have different sample sizes and privacy budgets. We address this through a class of per-site estimation algorithms that reports the ATE estimate *and* its variance as a quality measure, and an aggregation algorithm on the server side that minimizes the overall variance of the final ATE estimate. Our experiments on real and synthetic data show that our method reliably aggregates private statistics across sites and provides better privacy-utility tradeoff under site heterogeneity than baselines.**

***Index Terms*—Differential Privacy, Average Treatment Effect, Federated Analysis**

I. Introduction

Patient privacy is a major barrier to healthcare AI. Patient confidentiality reasons prevent hospitals and healthcare providers from freely sharing data; consequently, valuable data often remains in silo in separate sites, preventing the development of healthcare AI systems that can learn from large volumes of patient data to make effective decisions. A potential solution to this challenge is collaborative privacy-preserving learning across multiple sites through federated learning with differential privacy. While this has been well-explored for statistical estimation and machine learning problems, these are quite different from the causal inference-related problems that arise in healthcare applications.

This work takes a fresh look at federated learning with differential privacy, and applies it to causal inference—specifically to average treatment effect (ATE) estimation. Here, we are given data (X_i, Y_i, W_i) for patient i, where W_i corresponds to a treatment (for example, surgery or not), Y_i to an outcome (for example, recovery or not), and X_i to some covariates or features that describe the patient. The goal is to find the average treatment effect or ATE, which measures if treatment results in a different outcome than non-treatment. While this is easy if the treatments are randomly assigned (that is, under randomized control trials or RCTs), the problem is more challenging with observational data where the assignment of the treatment might depend on the covariates. For example, sicker patients may be denied surgery, which may make surgery look like a more appealing option.

Specifically, we consider the problem of ATE estimation from multiple sites, with differential privacy [1], [2], which has emerged as the gold standard in privacy-preserving data analysis. We ensure that each site calculates a DP statistic on its data to ensure the confidentiality of its patients; these statistics are then aggregated by a central server to form an effective ATE estimate.

There are three main challenges in multi-site DP causal inference. First, for observational studies, where the treatment assignment is not controlled, designing even a single-site DP ATE estimator is not straightforward, and little is known about the problem. In particular, the matching estimator, one of the most standard estimators [3], is hard to sanitize since it can significantly depend on a single individual's data in the worst case. Second, the estimate quality can vary across sites due to varying sample sizes and privacy budgets. Therefore, each site needs to report not only the ATE estimate but also a quality measure—which sets the problem apart from standard differentially private federated learning estimation solutions. Third, given the ATE estimates and their quality measures, the central server needs to aggregate them appropriately into a final accurate estimate.

We address the first challenge by proposing a smooth-sensitivity-based DP matching algorithm, SmoothDPMatching. Our algorithm adds significantly less noise for typical real-world datasets than the naive global sensitivity baseline, achieving a better privacy-utility tradeoff. To deal with the second challenge, we let each site send its ATE estimate variance as a quality measure. Since a site estimates the variance with sensitive data, it publishes the private variance estimate to the server to guarantee privacy. To address the third challenge, we propose a minimum-variance aggregation algorithm, MVAgg. MVAgg chooses a subset of sites to aggregate so that the variance of the final ATE estimate is minimum. Combining these three key components gives us a complete method for multi-site DP causal inference.

We evaluate our method on real and synthetic randomized trial and observational study datasets and find that our algorithms lead to significant gains in privacy-accuracy tradeoffs. Specifically, MVAgg automatically adopts estimates from high-quality sites and outperforms baselines, while reliably aggregating the per-site estimates with varying privacy budgets. We also see that SmoothDPMatching considerably reduces the noise variance, and achieves an improved privacy-accuracy tradeoff on both real and synthetic datasets.

A. Related Work

The most closely related model to our work is the distributed differential privacy model, e.g., [1], [4]–[6], where clients report differentially private output to the untrusted central server. There has been a body of work on the combination of distributed differential privacy and secure aggregation [4], [7]–[9]. Secure aggregation ensures that the server obtains the aggregate result but never sees the individual values. To prevent privacy leakages due to the aggregate result, the clients output locally differentially private (LDP) statistics. The fact the server only sees the aggregate result generally amplifies the final central DP guarantee. Shuffling model is another model of distributed differential privacy [5], [6], [10]–[15]. The model assumes an entity called shuffler, which receives LDP outputs from clients, uniformly permutes them, and sends the shuffled one to the central server. Shuffling further amplifies the privacy guarantee by making it harder for the server to identify individual information. While these two distributed differential privacy models mainly address privacy amplification on the final central DP guarantee, our work focuses on how to aggregate client statistics with different qualities to obtain a more accurate final output by the server.

Another line of related work is causal inference under privacy guarantees. The main focus of such papers is to carry out causal inference with privacy in a central DP model, i.e., at a single site, whereas we investigate how causal inference can be done with multiple sites while preserving privacy at each site. [16] provide a private version of the inverse probability weighting (IPW) method for observational study data. [17] study a private procedure to determine whether X causes Y or Y causes X under an additive noise model by privatizing the statistical dependence scores such as Spearman's ρ and Kendall's τ. [18] address private causal graph discovery for categorical and numerical data. More recently, [19] propose a DP meta-algorithm which estimates conditional ATE (CATE). [20] investigate how introducing DP impacts the identification of statistical models. Note that, to the best of our knowledge, no work has addressed the matching estimator under DP even for a single site setting, which is one of our contributions.

Apart from the privacy literature, there has been a line of work discussing how ATE estimation can be done in multisite random trials under site variation in treatment effect [21]–[27]. As for learning from data with variate quality, [28] provide a theory for choosing an appropriate set of data sources with variable qualities. [29] study how heterogeneous noise impacts the performance of stochastic gradient descent (SGD).

II. Preliminaries & Problem Setting

A. Differential Privacy & Federated Learning/Analytics

Differential privacy is a strong cryptographically-motivated definition of individual-level privacy. It guarantees that the participation of a single individual in a dataset does not change the probability of any outcome by much. In particular, suppose we have two datasets D and D', each consisting of private data from n individuals. We say that D and D' are neighboring if they differ in a single individual's private data, i.e., $d(D, D') = |\{i : D_i \neq D'_i\}| = 1$. The output distribution of a differentially private (randomized) algorithm is guaranteed to be close on neighboring datasets.

Definition 1 ((ϵ, δ)-Differential Privacy [1])**.** *A randomized algorithm M satisfies (ϵ, δ)-differential privacy if for any two neighboring datasets D, D' and for any $S \subseteq \mathrm{range}(M)$,*

$$\Pr[M(D) \in S] \leq \exp(\epsilon) \Pr[M(D') \in S] + \delta.$$

The most common differentially private mechanism is the Global Sensitivity method, where we compute a function f on a dataset D, and add noise that is calibrated to the global sensitivity of the function. Specifically, the global sensitivity of a function f is the maximum difference between the outputs of f on *any two* neighboring datasets. The standard instances of the global sensitivity method are the Laplace mechanism [2], which guarantees $(\epsilon, 0)$-DP, and the Gaussian mechanism [30], which guarantees (ϵ, δ)-DP.

a) Global Sensitivity & Laplace [2] and Gaussian [30] mechanism.: The global sensitivity of a scalar function $f : \mathcal{X}^n \to \mathbb{R}$ is

$$\Delta_f = \max_{D, D'} |f(D) - f(D')|,$$

where D and D' are neighboring datasets.
Let $\epsilon > 0$ be arbitrary and $f : \mathcal{X}^n \to \mathbb{R}$ be a function. Then, the algorithm M: $M(D) = f(D) + \xi$ satisfies $(\epsilon, 0)$-DP, where $\xi \sim \mathrm{Lap}(\Delta_f/\epsilon)$.
Furthermore, let $\epsilon, \delta \in (0, 1)$ be arbitrary and $f : \mathcal{X}^n \to \mathbb{R}$ be a function. Then, for $c^2 > 2\ln(1.25/\delta)$, the algorithm M: $M(D) = f(D) + \xi$ satisfies (ϵ, δ)-DP, where $\xi \sim \mathcal{N}(0, \sigma^2)$ and $\sigma \geq \frac{c\Delta_{2,f}}{\epsilon}$.

For certain functions, such as the median [31], the global sensitivity may be too high, which may lead to a poor privacy-accuracy tradeoff. In these cases, [31] propose calibrating the noise instead to the smoothed sensitivity, which is a smoothed version of the local sensitivity. Adding the Laplace noise calibrated to the smooth sensitivity still guarantees DP with a slight overhead in the δ term.

b) Local and Smooth Sensitivity & Laplace mechanism [31].: The local sensitivity of a function $f : \mathcal{X}^n \to \mathbb{R}$ at D is

$$\mathrm{LS}_f(D) = \max_{D' : d(D, D') = 1} |f(D) - f(D')|.$$

For $\beta > 0$, the β-smooth sensitivity of f is

$$S^*_{f,\beta}(D) = \max_{D' \in \mathcal{X}^n} \mathrm{LS}_f(D') \cdot \exp(-\beta d(D, D')).$$

If $\beta \leq \epsilon/2\ln(\frac{2}{\delta})$ and $\delta \in (0, 1)$, the algorithm $M : \mathcal{X}^n \to \mathbb{R}$:

$$M(D) = f(D) + \frac{2S^*_{f,\beta}(D)}{\epsilon} \cdot \eta,$$

where $\eta \sim \mathrm{Lap}(0, 1)$, satisfies (ϵ, δ)-DP.

Federated Learning/Analytics (FL/FA) [32] is an emerging paradigm for collaborative learning across multiple devices or sites, which allows a server to learn a model or some target statistics over sensitive client data, without directly acquiring raw data from the clients. However, it is well-known that FL/FA by itself does not directly offer privacy, since the client updates themselves can be reverse-engineered to extract user data [33]–[37]. Hence, we will be considering FL/FA with differential privacy. Additionally, we consider FL/FA over a small number of clients, each of which holds data from a certain number of individuals.

B. Average Treatment Effect

Suppose we have a group of people who are given a treatment, and our goal is to determine whether the treatment is effective. This is done through estimating the Average Treatment Effect (ATE). In particular, for an individual i, we assume two potential outcomes $Y_i(1)$ and $Y_i(0)$, where $Y_i(1)$ is under treatment and $Y_i(0)$ is under control. The average treatment effect (ATE) is then measured by:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)].$$

In practice, estimating ATE is not straightforward since we get to observe only one of $Y_i(1)$ and $Y_i(0)$ and cannot directly compute *individual* treatment effect, $Y_i(1) - Y_i(0)$. Instead, we observe the treatment indicator W_i (1 when treated, 0 under control), the corresponding outcome Y_i^{obs}, and some other covariates X_i. Then, we aim to estimate the ATE given a set of observed individuals' data, $\{W_i, Y_i^{\text{obs}}, X_i\}_{i=1}^N$.

We follow the standard causal inference literature [38] to make following three assumptions on these variables.

1) Stable Unit Treatment Value Assumption (SUTVA): the potential outcomes $(Y_i(1), Y_i(0))$ do not depend on treatments assigned to other individuals
2) Unconfoundedness:

$$\Pr[W_i = 1 | X_i, Y_i(1), Y_i(0)] = \Pr[W_i = 1 | X_i]$$

3) Positivity:

$$\forall x. \quad 0 < \Pr[W_i = 1 | X_i = x] < 1$$

1) Randomized trial and Difference-in-means Estimator: In a randomized trial, where treatment assignment is completely random, we estimate ATE via the difference-in-means estimator. In a randomized trial, where treatment assignment is completely random and independent of individual data, we estimate the ATE via the difference-in-means estimator. Specifically, the numbers of treated and control individuals, N_t and N_c ($N = N_t + N_c$), are determined in advance, and the treatment indicators $W_1, \ldots, W_N$ are drawn from the following distribution:

$$\Pr[W_1, \ldots, W_N] = \begin{cases} \binom{N}{N_t}^{-1} & \text{if} \quad \sum_i W_i = N_t \\ 0 & \text{o.w.} \end{cases}.$$

Then, the ATE estimate is:

$$\hat{\tau} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{\text{obs}}.$$

2) Observational Studies and Matching Estimator: In observational studies, where the treatment assignment is not controlled, standard practice is to use a matching estimator. This estimator first imputes the unobserved outcome of an individual by the observed outcome of a *similar* individual who has the opposite treatment status, and then outputs the average of the individual treatment effects.

Among its variants, we use exact single matching under the assumption that there always exists a *similar* individual, i.e., for an individual i, there exists at least one individual j s.t. $W_i \neq W_j$ and $X_i = X_j$. Let $m : [N] \to [N]$ be a matching function s.t. $m(i) = j \implies W_i \neq W_j \wedge X_i = X_j$. Then, ATE is estimated by:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_i(1) - \hat{Y}_i(0)),$$

where $\hat{Y}_i(1) = W_i Y_i^{\text{obs}} + (1 - W_i) Y_{m(i)}^{\text{obs}}$, and $\hat{Y}_i(0) = W_i Y_{m(i)}^{\text{obs}} + (1 - W_i) Y_i^{\text{obs}}$. One of $\hat{Y}_i(1)$ and $\hat{Y}_i(0)$ is exactly Y_i^{obs} and the other is imputed outcome by the matched individual.

C. Problem Setting

Our goal is to estimate the ATE of a specific binary treatment, where data about the effect of this treatment is distributed across a small number of sites. We would like to ensure that raw data stays on the site, and only differentially private estimates leave a particular site. Specifically, the ATE computation is done by an untrusted server, which receives private statistics from J sites. We demonstrate the figure for this framework in Figure 1.

a) Some Basic Notation.: We assume that site j requires (ϵ_j, δ_j)-DP, and use the notation $\hat{\tau}_{j-\mathrm{DP}}$ to denote the ATE at site j. The final ATE estimated at the server is denoted by $\hat{\tau}_{\mathrm{DP}}$. Furthermore, site j has a dataset D_j of size N_j. Note that the sample sizes, N_j's, are public information since we are interested in the site-level privacy guarantee. i-th element in D_j is a tuple $(Y_{ij}(1), Y_{ij}(0), X_{ij})$, where potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$ are assumed to be bounded, i.e., $0 \leq Y_{ij}(1) \leq B$

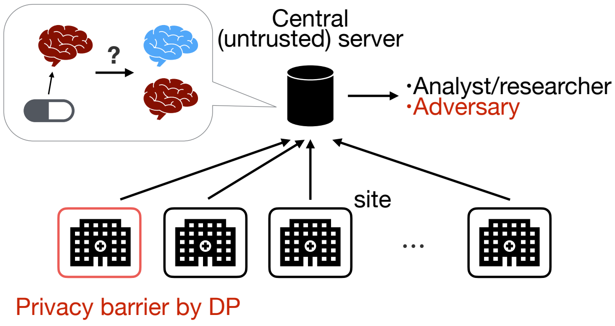

Fig. 1: Our framework on estimating ATE with data from distributed sites

and $0 \leq Y_{ij}(0) \leq B$ for some B, and X_{ij} is the covariates, if any, of i-th individual at site j. X_{ij} is used for observational studies and is typically a vector of multiple covariates. In this work, we assume that it is an element of some finite set $\mathcal{X}$. In practice, covariates can contain continuous values, e.g., height and weight, but we can often discretize them without losing much precision. Then, a site assigns the treatment, W_{ij}, observes the outcome, Y_{ij}^{obs}, and estimates the ATE with a set of observed individual data, $\{W_{ij}, Y_{ij}^{\text{obs}}, X_{ij}\}_{i=1}^{N_j}$. We note that the neighboring datasets for DP are based on the dataset definition *before the treatment assignment*, namely, they differ by an individual's potential outcomes and/or covariates. As a result, the treatment indicators of the differing individuals in neighboring datasets are the same for randomized trials, but can be different for observational studies because the assignment depends on the covariates.

b) Assumptions.: In addition to the three standard causal inference assumptions [38], we make two other mild assumptions. The first is that the sites are homogeneous. That is, if each individual has potential outcomes, $Y_{ij}(1)$ and $Y_{ij}(0)$, we further assume that a tuple $(W_{ij}, Y_{ij}(1), Y_{ij}(0), X_{ij})$ is drawn i.i.d. from some fixed distribution. This assumption is needed so that the estimand $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ makes sense; without this, the underlying ATE at each site differs and it is no longer clear what estimand we should use. Our second assumption is that an individual cannot belong to more than one site. This ensures the privacy loss does not accumulate by outputs from multiple sites.

III. Method

Our method consists of two interconnected components on a distributed client-server setting—first, a site-level estimation algorithm and second, a server-side aggregation algorithm. Combining these two components gives us a complete method for private and distributed ATE estimation.

A. Per-Site Estimation Algorithm

In a FL/FA setting, a per-site estimation algorithm computes a per-site gradient/target statistic on its local data, adds noise for privacy, and sends it to the server. This simple solution, however, does not directly apply to us. For the server to aggregate the ATEs appropriately, it needs to know a quality measure for the ATEs from each site because the estimation quality can vary across sites due to varying sample sizes and privacy budgets. For example, even though all the sites have the same total privacy budget, they can answer multiple queries on the same data, and the privacy budgets allocated for an ATE estimation query could differ by site. To this end, we calculate a differentially private variance estimate for the ATE, which provides a comprehensive estimate by taking into account non-private estimate variance as well as additive noise for DP. Another difficulty for us is that, in the common observational data case, standard ATE estimators are more involved than the sum or average over individual values. Therefore, it is not obvious how to construct their DP versions. Thus, we propose a new smooth-sensitivity-based DP matching algorithm, SmoothDPMatching, through an analysis of the smooth sensitivity of the matching estimator. Our algorithm significantly reduces the noise variance and hence improves the accuracy for typical datasets compared with the baseline global sensitivity method.

1) Randomized Trial and Difference-in-means Estimator: For randomized trials, we use the difference-in-means estimator for ATE—namely, $\hat{\tau} = \sum_{i:W_i=1} Y_i^{\text{obs}}/N_t - \sum_{i:W_i=0} Y_i^{\text{obs}}/N_c$. Its differentially private version can be straightforwardly computed using the global sensitivity method. The global sensitivities of $\sum_{i:W_i=1} Y_i^{\text{obs}}$ and $\sum_{i:W_i=0} Y_i^{\text{obs}}$ are both B. Thus, by using the Laplace mechanism, we have:

$$\begin{aligned}\hat{\tau}_{\text{DP}} &= \frac{1}{N_t}\left(\sum_{i:W_i=1} Y_i^{\text{obs}} + \xi_t\right) - \frac{1}{N_c}\left(\sum_{i:W_i=0} Y_i^{\text{obs}} + \xi_c\right) \\ &= \hat{\tau} + \frac{\xi_t}{N_t} - \frac{\xi_c}{N_c},\end{aligned}$$

where $\xi_t, \xi_c \sim \text{Lap}({}^{B}/_{\epsilon_1})$. Recall that N_t and N_c are predetermined parameters; thus, we treat them as public information. Furthermore, the treatment indicator W_i does not change by changing an individual's data. Therefore, by the parallel composition theorem of DP, the mechanism satisfies ϵ_1-DP. We provide the formal proof in Appendix.

The private variance estimation of $\hat{\tau}_{\text{DP}}$ is also simple because ξ_t and ξ_c are independent from the data distribution. That is, we have

$$\begin{aligned}\mathbb{V}[\hat{\tau}_{\text{DP}}] &= \mathbb{V}[\hat{\tau}] + \mathbb{V}\left[\frac{\xi_t}{N_t}\right] + \mathbb{V}\left[\frac{\xi_c}{N_c}\right] \\ &= \mathbb{V}[\hat{\tau}] + \frac{2B^2(\frac{1}{N_t^2} + \frac{1}{N_c^2})}{\epsilon_1^2},\end{aligned}$$

where the last term is computed only with public information. It remains to estimate the sampling variance term, $\mathbb{V}[\hat{\tau}]$, with sensitive data and sanitize the estimate using the global sensitivity method. In particular, $\mathbb{V}[\hat{\tau}]$ is estimated with $s_t^2/N_t + s_c^2/N_c$ [38], where s_t^2 and s_c^2 are sample variance of $\sum_{i:W_i=1} Y_i^{\text{obs}}/N_t$ and $\sum_{i:W_i=0} Y_i^{\text{obs}}/N_c$. It suffices to privately estimate

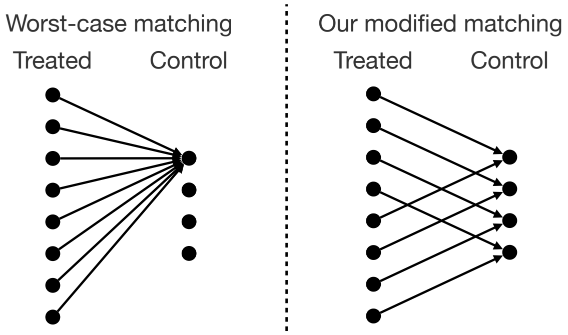

Fig. 2: Worst-case matching (left) and our matching (right) within the same covariate stratum. We omit arrows from control individuals to treated ones for readability.

$\sum_{i:W_i=1}(Y_i^{\text{obs}})^2$ to obtain the private estimate of s_t^2 because $s_t^2 = \sum_{i:W_i=1}(Y_i^{\text{obs}})^2/N_t - (\sum_{i:W_i=1} Y_i^{\text{obs}}/N_t)^2$ and we already know the private estimate of $\sum_{i:W_i=1} Y_i^{\text{obs}}$. The global sensitivity of $\sum_{i:W_i=1}(Y_i^{\text{obs}})^2$ is B^2; thus, we apply the Laplace mechanism with a parameter ϵ_2 and obtain the private estimate of s_t^2. The same argument yields the private estimate of s_c^2. By the parallel composition theorem, this computation satisfies ϵ_2-DP. Consequently, we obtain the private variance estimate of the differentially private difference-in-means estimator. We send this estimation along with $\hat{\tau}_{\text{DP}}$ to the server, which satisfies $(\epsilon_1 + \epsilon_2)$-DP by the sequential composition theorem of DP.

2) Observational Study and Matching Estimator: Things however are more complicated for observational data, since we need to match covariates. To do this privately, we propose a new differentially private approximation to the exact matching estimator in Section II. The main challenge here is that changing one value in the input dataset can alter the final output significantly *in the worst case*. We address this through a smooth-sensitivity-based algorithm, SmoothDP-Matching, which requires much less noise for typical datasets.

We provide the specification to the estimator to make it more amenable to DP. First, recall that the exact matching estimator assumes that every individual in the data has an exact match. If an individual i does not have any matched individual j, then we extend the matching function m to be $m : [N] \to [N] \cup \{\perp\}$ and set $\hat{Y}_i(1) = \hat{Y}_i(0) = Y_i^{\text{obs}}$ when $m(i) = \perp$ so that this term contributes 0 to the ATE. Second, for each covariate stratum $X = x$, the exact matching estimator may match a control individual to many treated individuals while other control individuals have no matches, driving up the sensitivity. We ensure that we balance the number of individuals matched to a particular individual for each covariate stratum as shown in Figure 2. This can be done with a greedy algorithm shown in Algorithm 1. We provide the further details in Appendix.

Even with the specification, unfortunately, it is shown as below that the global sensitivity of the matching algorithm is still a constant. We therefore use a smoothed sensitivity estimator.

Proposition 1. *Let $\hat{\tau}$ be the exact single matching estimator in Algorithm 1, then $\Delta_{1,\hat{\tau}} \geq B$.*

Proof. Consider a pair of neighboring datasets D, D' where $|D| = |D'| = N$, $D_N \neq D'_N$, $D_i = (Y_i(1), Y_i(0), X_i) = (B, B, x)$ for $i = 1, \dots, N-1$, $D_N = (0, 0, x)$, and $D'_N = (0, 0, x')$. Furthermore, consider the treatment assignments yielding $W_i = 1$ for $i = 1, \dots, N-1$, $W_N = 0$, and $W'_N = 1$, where W'_N is the assignment for N-th individual in D'. Here, all individuals in D' are treated; thus, $\hat{\tau}(D') = 0$. Then, $|\hat{\tau}(D) - \hat{\tau}(D')| = |\frac{1}{N}(\sum_{i=1}^{N-1}(Y_i^{\text{obs}} - Y_{m(i)}^{\text{obs}}) + (Y_{m(N)}^{\text{obs}} - Y_N^{\text{obs}})) - 0| = \frac{1}{N}((N-1)B + B) = B$. By the definition of the global sensitivity, the statement is shown. □

We first analyze the smooth sensitivity of our exact single matching estimator, $\hat{\tau}$. For most real datasets, we anticipate the local sensitivity is $\mathcal{O}(1/N)$ whereas the global sensitivity is $\Omega(1)$. Then, the smooth sensitivity, the smooth upper bound of the local sensitivity, is also $\mathcal{O}(1/N)$. This is in fact true as stated in the theorem below.

Theorem 1. *Let $T_x = \{i : W_i = 1 \wedge X_i = x\}$ and $C_x = \{i : W_i = 0 \wedge X_i = x\}$ be the sets of treated and control individuals with the covariate x. Then, the local sensitivity of $\hat{\tau}$ is upper bounded as follows:*

$$\mathrm{LS}_{\hat{\tau}}(D) \leq \frac{1}{N} \max_{x \in \mathcal{X} : |T_x| > 0 \vee |C_x| > 0} \begin{cases} 4(1 + \max(|T_x|, |C_x|))B & |T_x| = 0 \vee |C_x| = 0 \\ 4(1 + \max(\lceil \frac{1+|C_x|}{|T_x|} \rceil, \lceil \frac{1+|T_x|}{|C_x|} \rceil))B & o.w. \end{cases}.$$

Furthermore, let $R_x^{(k)}(D) = \begin{cases} \max(|T_x|, |C_k|) + k & \min(|T_x|, |C_x|) \leq k \\ \lceil \frac{\max(|T_x|, |C_k|) + k + 1}{\min(|T_x|, |C_x|) - k} \rceil & o.w. \end{cases}$. *Then, the β-smooth sensitivity of $\hat{\tau}$ is upper bounded as follows:*

$$S^*_{\hat{\tau},\beta}(D) = \max_{k=0,\dots,N} e^{-k\beta} \frac{4B}{N}(1 + \max_{x \in \mathcal{X}} R_x^{(k)}(D)). \quad (1)$$

In addition, $S^*_{\hat{\tau},\beta}(D)$ can be computed with $\mathcal{O}(\min(|\mathcal{X}|, N))$ space and $\mathcal{O}(N \cdot \min(|\mathcal{X}|, N))$ time.

We provide the proof in Appendix.

We observe that if the dataset is well-balanced in each covariate value x, i.e., $|T_x| \approx |C_x|$, $\mathrm{LS}_{\hat{\tau}}(D) = \mathcal{O}(1/N)$ and also $S^*_{\hat{\tau},\beta}(D) = \mathcal{O}(1/N)$. In contrast, the global sensitivity is $\Omega(1)$ regardless of a dataset. We demonstrate this observation by numerical simulations on a synthetic dataset in Section IV.

For completeness, we present a (ϵ, δ)-DP matching algorithm shown in Algorithm 2, which calibrates the Laplace noise to the analyzed smooth sensitivity.

Next, we turn to estimating the variance of our DP matching estimator privately. The variance estimate is slightly more involved since the additive noise variance now depends on the smooth sensitivity, which is *data-dependent*. Thus, we instead obtain the private estimate of the smooth sensitivity, in addition to the sampling variance, to get the overall variance estimate.

More formally, we consider the variance conditioned on X_i's and W_i's as in the literature [38]. Recall we have

Algorithm 1: Non-private Matching at site

```
Data: Observed data: $\{(W_i, Y_i^{\text{obs}}, X_i)\}_{i=1}^N$
Define placeholders $\{(\hat{Y}_i(1), \hat{Y}_i(0))\}_{i=1}^N$
for $x \in \mathcal{X}$ do
    $T_x = \{i : W_i = 1 \land X_i = x\}$
    $C_x = \{i : W_i = 0 \land X_i = x\}$
    if $|T_x| = 0 \lor |C_x| = 0$ then /* when there's no match */
        for $i \in T_x \cup C_x$ do
            $(\hat{Y}_i(1), \hat{Y}_i(0)) \leftarrow (Y_i^{\text{obs}}, Y_i^{\text{obs}})$
    else /* when there are matches */
        for $j = 0$ to $|T_x| - 1$ do
            $i \leftarrow T_x[j]$
            $m(i) \leftarrow C_x[j \bmod |C_x|]$ /* matched individual */
            $(\hat{Y}_i(1), \hat{Y}_i(0)) \leftarrow (Y_i^{\text{obs}}, Y_{m(i)}^{\text{obs}})$
        for $j = 0$ to $|C_x| - 1$ do
            $i \leftarrow C_x[j]$
            $m(i) \leftarrow T_x[j \bmod |T_x|]$ /* matched individual */
            $(\hat{Y}_i(1), \hat{Y}_i(0)) \leftarrow (Y_{m(i)}^{\text{obs}}, Y_i^{\text{obs}})$
Compute non-private ATE: $\hat{\tau} = \frac{1}{N}\sum_{i=1}^N \hat{Y}_i(1) - \hat{Y}_i(0)$
return $\hat{\tau}$
```

Algorithm 2: SmoothDPMatching at site

```
  Data: Observed data: $\{(W_i, Y_i^{\text{obs}}, X_i)\}_{i=1}^N$, Privacy parameters: $\epsilon$, $\delta$
1 $\beta = \frac{\epsilon}{2\ln(\frac{2}{\delta})}$
2 Compute $S^*_{\hat{\tau},\beta}(D)$ as in Eq. (1)
3 Compute non-private ATE: $\hat{\tau} = \frac{1}{N}\sum_{i=1}^N \hat{Y}_i(1) - \hat{Y}_i(0)$ with Algorithm 1
4 $\hat{\tau}_{\text{DP}} = \hat{\tau} + \frac{2S^*_{\hat{\tau},\beta}(D)}{\epsilon} \cdot \eta$, where $\eta \sim \text{Lap}(1)$
5 return $\hat{\tau}_{\text{DP}}$
```

$\hat{\tau}_{\text{DP}} = \hat{\tau} + (2S^*_{\hat{\tau},\beta}(D)/\epsilon) \cdot \eta$, where $\eta \sim \text{Lap}(1)$ (line 4 in Algorithm 2). Then, the variance of $\hat{\tau}_{\text{DP}}$ is:

$$\begin{aligned}\mathbb{V}[\hat{\tau}_{\text{DP}}] &= \mathbb{V}[\hat{\tau}] + \mathbb{V}\left[\frac{2S^*_{\hat{\tau},\beta}(D)}{\epsilon} \cdot \eta\right] \\ &= \mathbb{V}[\hat{\tau}] + \frac{8(S^*_{\hat{\tau},\beta}(D))^2}{\epsilon^2}.\end{aligned}$$

It remains to estimate both terms from data privately. Note that, conditioned on X_i's and W_i's, the smooth sensitivity in eq. (1) is constant.

As for the sampling variance term, $\mathbb{V}[\hat{\tau}]$, we obtain its differentially private estimate by the smooth sensitivity method. We defer the detail to Appendix. As for the second term, $8(S^*_{\hat{\tau},\beta}(D))^2/\epsilon^2$, we need to privately estimate the smooth sensitivity because it is data-dependent. We thus provide an *unbiased* (ϵ, δ)-DP estimator of the β-smooth sensitivity as follows. Since the smooth sensitivity is the *smoothed* version of the local sensitivity, it is designed not to vary a lot by changing a single individual's data. Therefore, we first show that the global sensitivity of $\ln S^*_{\hat{\tau},\beta}$ is β. Then, we apply the Gaussian mechanism to obtain the differentially private smooth sensitivity, $S^*_{\hat{\tau},\beta-\text{DP}}(D)$.

Lemma 1. *Let $S^*_{f,\beta}$ be the β-smooth sensitivity of f. Then,*

$$S^*_{f,\beta-\text{DP}}(D) = \exp(\ln S^*_{f,\beta}(D) + z - \frac{\sigma^2}{2}),$$

where $\sigma = \sqrt{2\ln(1.25/\delta)}\beta/\epsilon$ and $z \sim \mathcal{N}(0, \sigma^2)$, satisfies (ϵ, δ)-differential privacy.
Furthermore, it holds that $\mathbb{E}[S^*_{f,\beta-\text{DP}}(D)] = S^*_{f,\beta}(D)$, where the randomness is over the draws of z.

Proof. By the definition of smooth sensitivity, it holds that for any neighboring datasets D, D', $S^*_{f,\beta}(D) \leq e^{\beta} S^*_{f,\beta}(D')$. Therefore, the global sensitivity of $\ln S^*_{f,\beta}$ is β. The privacy guarantee of the Gaussian mechanism and the post-processing theorem of DP guarantee (ϵ, δ)-DP for $S^*_{f,\beta-\text{DP}}$.

Additionally, by taking the expectation over the draws of z,

```
Algorithm 3: MVAgg at server
Data: Sample sizes: $N_1, \dots, N_J$, Noisy estimates and
      their variances: $\{\hat{\tau}_{j-\mathrm{DP}}, \hat{\sigma}^2_{j-\mathrm{DP}}\}_{j=1}^J$
1 $I^* = \mathrm{argmin}_{I \subseteq [J]} \sum_{j \in I} (\frac{N_j}{N_I})^2 \hat{\sigma}^2_{j-\mathrm{DP}}$
2 $\hat{\tau}_{\mathrm{DP}} = \sum_{j \in I^*} \frac{N_j}{N_{I^*}} \hat{\tau}_{j-\mathrm{DP}}$
3 return $\hat{\tau}_{\mathrm{DP}}$
```

we have:

$$\begin{aligned}\mathbb{E}[S^*_{f,\beta-\mathrm{DP}}(D)] &= S^*_{f,\beta}(D) \cdot \exp(-\frac{\sigma^2}{2}) \cdot \mathbb{E}[\exp(z)] \\ &= S^*_{f,\beta}(D),\end{aligned}$$

where the last equality holds due to $\mathbb{E}[\exp(z)] = \exp(\frac{\sigma^2}{2})$. □

B. Aggregation Algorithm on Server

In the simple FL/FA, the server would simply average the gradients/statistics transmitted by the clients. Unfortunately for us, this solution is not enough—different sites will have different estimation quality, due to varying dataset size and/or varying privacy budgets. We therefore propose a new aggregation procedure that takes this heterogeneity into account.

Since we are interested in the average treatment effect on *an individual*, we consider the weighted average of ATEs from sites with weights proportional to the sample sizes at sites N_j's. Given a set of sites I and a set of DP ATE estimates $\{\hat{\tau}_{j-\mathrm{DP}}\}_{j \in I}$, let $N_I = \sum_{j \in I} N_j$, and the server publishes $\hat{\tau}_{\mathrm{DP}} = \sum_{j \in I} ({N_j}/{N_I}) \cdot \hat{\tau}_{j-\mathrm{DP}}$.

The central problem at the server is then how to choose the set of sites I. When some sites in I have very noisy ATE estimates, the final estimate can be noisy, or has high variance, as well. In such a case, we might want to remove these sites from the set so that the final estimate is less noisy. Therefore, we propose a new aggregation algorithm that embodies this idea by choosing the set of sites that minimizes the variance of the aggregate ATE: minimum-variance aggregation algorithm (MVAgg).

More concretely, the minimum-variance aggregation algorithm shown in Algorithm 3 takes noisy ATEs, noisy variance of ATE, and sample sizes as the inputs. Then, it minimizes the estimated variance over a set of sites. Here, since $\hat{\tau}_{\mathrm{DP}}$ is the weighted average of $\hat{\tau}_{j-\mathrm{DP}}$'s, its variance given the set I is $\mathbb{V}[\hat{\tau}_{\mathrm{DP}}] = \sum_{j \in I} (N_j/N_I)^2 \mathbb{V}[\hat{\tau}_{j-\mathrm{DP}}]$. It finally computes the weighted average of the noisy ATEs over the chosen set of sites. Note that by the post-processing theorem of DP, the privacy guarantee at each site never changes as a result of the aggregation.

Our algorithm is general in the sense that it only requires the noisy estimate and its noisy variance from each site in addition to the publicly known sample sizes, and it does not limit the specific estimator used at each site. On the other hand, our algorithm currently adopts a brute-force search to determine the minimum variance set (line 1 in Algorithm 3). Finding a greedy approximation algorithm for the minimization is a possible direction for our future work.

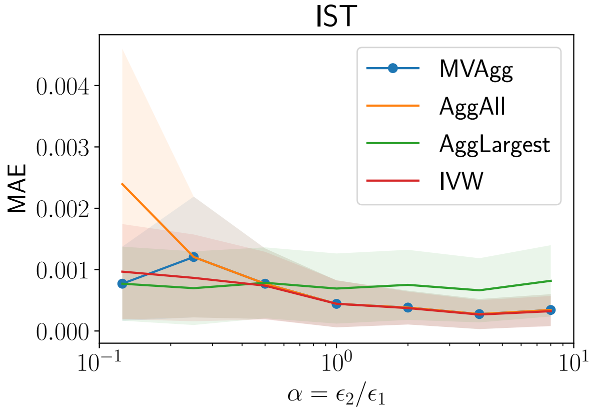

Fig. 3: Mean MAEs and standard deviations of MVAgg, AggAll, AggLargest, and IVW on IST dataset under two-site setting

IV. Experiment

We now empirically investigate how putting together our estimation algorithms at the client sites with the aggregation process on the server side works. Specifically, we ask the following questions:

1) How does the smooth-sensitivity-based DP matching algorithm (Algorithm 2) improve the privacy-utility tradeoff on observational study data at each site?
2) How does our aggregation algorithm (Algorithm 3) impact the final ATE estimation on the server on randomized trial and observational study data?
3) How do site-level privacy parameters affect the overall performance of the algorithms?

We answer the first question with real and synthetic observational study data. We then answer the rest of the questions with real randomized trial data as well as those observational data.

A. Methodology

a) Datasets.: For randomized trial data, we use two real datasets. The International Stroke Trial (**IST**) [39] is a dataset with $N = 18995$ individuals, where $N_t = 9705$ are randomly treated by the aspirin allocation and $N_c = 9703$ are controlled. The outcome measures whether the recurrent ischemic stroke occurs within 14 days after treatment. Tennessee's Student Teacher Achievement Ratio (**STAR**) dataset [40] contains the trial results from $N = 10331$ students, who are randomly assigned into either a small class ($N_t = 2643$) or regular-size class ($N_c = 7688$). The data is collected from 80 schools. We use the four kinds of school urbanity (rural, suburban, urban, inner city) to determine which site the student belongs to, i.e., $J = 4$. The sample sizes result in $N_1 : N_2 : N_3 : N_4 \approx 5 : 3 : 3 : 1$.

For observational study data, we use a synthetic, a semi-real, and a real dataset. We generate the synthetic dataset (**Synth**) by first sampling X_i's uniform randomly from a discrete set $\mathcal{X} = \{0, 1/(|\mathcal{X}| - 1), \dots, 1\}$. Then, we sample W_i from the

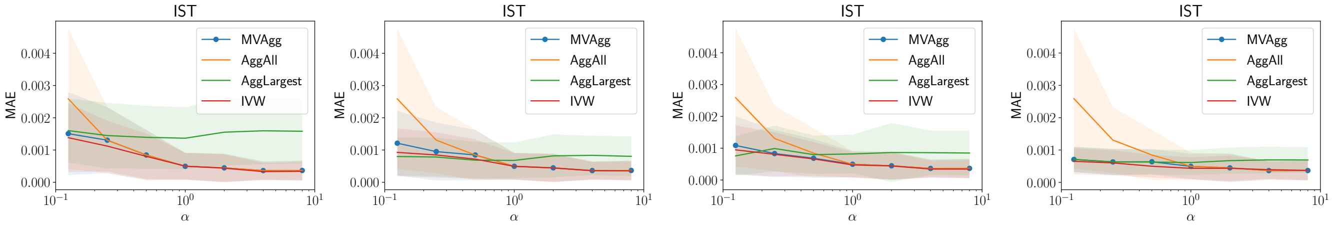

Fig. 4: Mean MAEs and standard deviations of MVAgg, AggAll, AggLargest, and IVW on IST dataset under three-site setting. $N_1 : N_2 : N_3 = 1 : 1 : 1$ (left most), $3 : 2 : 1$ (middle left), $9 : 9 : 2$ (middle right), and $18 : 1 : 1$ (right most).

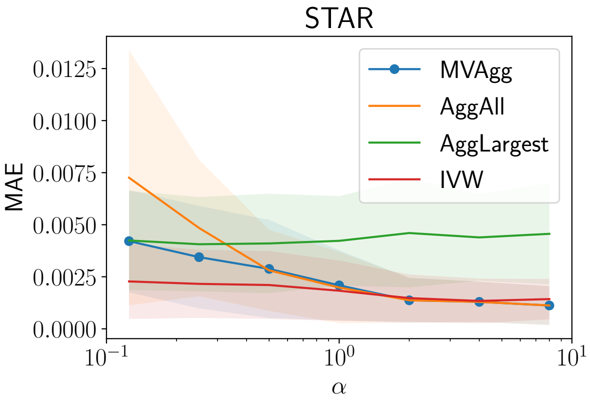

Fig. 5: Mean MAEs and standard deviations of MVAgg, AggAll, AggLargest, and IVW on STAR dataset ($J = 4$)

Bernoulli distribution with a parameter $\mathrm{sigmoid}(a \cdot (2X_i - 1))$ with some $a \in \mathbb{R}$, where a parameter a is drawn from $\mathrm{Uniform}([-1, 1])$ if not mentioned or set as constant. This ensures that the treatment variable W_i has some dependence on X_i. Note that as a gets larger, the dataset gets more imbalanced, i.e., $|T_x| \gg |C_x|$ or $|T_x| \ll |C_x|$. Finally, given the underlying true ATE $\tau = 0.5$, we set $Y_i = b \cdot X_i + \tau \cdot W_i + e_i$, where $b \sim \mathrm{Uniform}([0, 0.4])$ is a parameter and $e_i \sim \mathrm{Uniform}([0, 0.1])$ is observation noise. This generation process ensures that Y_i depends on both X_i and W_i and that $0 \leq Y_i \leq 1$. The semi-real dataset we use is the Infant Health and Development Program (**IHDP**) dataset [41], where only the outcome value is simulated. It has $N = 747$ individuals comprised of $N_t = 139$ treated and $N_c = 608$ controlled individuals. The treatment is specialist home visits to the children and the outcome is future cognitive test scores. We choose 3 discrete covariates out of 25 covariates in the original dataset to ensure the exact matching. The real dataset we use is **Lalonde** [42], which is composed of $N = 722$ individuals where $N_t = 297$ are treated and $N_c = 425$ are controlled. The treatment is job training and the outcome is earning in 1978. We choose age as the only covariate to ensure the exact matching. For all datasets, we preprocess them so that $0 \leq Y_i \leq 1$.

b) Algorithms.: The per-site estimation algorithms used in the experiments are as follows. We use the DP version of the difference-in-means estimator presented in Section III-A1 for randomized trial data. For observational study data, we compare two DP matching algorithms whose additive noises are calibrated to the global sensitivity (**GlobalDPMatching**) and smooth sensitivity (Algorithm 2; **SmoothDPMatching**) respectively.

We compare our **MVAgg** (Algorithm 3) with three baseline aggregation algorithms on the server. The first algorithm computes the weighted average of the ATEs from all sites with weights proportional to the sample sizes (**AggAll**). The second one publishes the result of the largest site (**AggLargest**). The last one aggregates the site ATE estimates with inverse-variance weights (**IVW**). In particular, the weight assigned to j-th site is $w_j = c/\hat{\sigma}^2_{j-\mathrm{DP}}$, where $c = 1/\sum_j(1/\hat{\sigma}^2_{j-\mathrm{DP}})$. IVW is known to be optimal if the true variances are given from sites. However, we only have the DP estimates of the variances; thus, it is not necessarily optimal in our setting.

c) Experiment Setup.: We consider two-site ($J = 2$) and three-site ($J = 3$) settings on the datasets except for STAR dataset where the sites are pre-assigned ($J = 4$). For the two-site setting, we randomly assign individuals to each site while keeping the sample sizes equal, $N_1 = N_2$ [1]. For the three-site setting, we consider different sample size proportions as follows: $N_1 : N_2 : N_3 = 1 : 1 : 1$, $3 : 2 : 1$, $9 : 9 : 2$, and $18 : 1 : 1$.

We fix the privacy parameter for the first site, ϵ_1, and sweep the others, $\epsilon_2, \ldots, \epsilon_J$. In particular, for each $\alpha \in \{1/8, 1/4, 1/2, 1, 2, 4, 8\}$, let $\epsilon_j = \alpha^{(j-1)/(J-1)}\epsilon_1$. As α gets larger, the second to J-th sites are expected to send more accurate statistics. We use $\epsilon_1 = 1$ for IST, STAR, and Synth, with $N = 10000$ and $|\mathcal{X}| = 100$, and $\epsilon_1 = 5$ for IHDP and Lalonde due to their small sample sizes. We further fix $\delta_1 = \cdots = \delta_J = 10^{-5}$. For the per-site estimation, we evenly split the privacy budget into multiple DP algorithms, e.g., we assign $(\epsilon/3, \delta/3)$ separately for obtaining the ATE estimate, the sampling variance estimate, and the private smooth sensitivity.

The evaluation metric is the mean absolute error (MAE) between the non-private and private ATE estimates. For Synth data, we measure the MAE between the true underlying ATE and the private estimate. We repeat the algorithms 100 times and report the mean and standard deviation of MAE.

[1] We assume AggLargest always chooses the first site as the largest site.

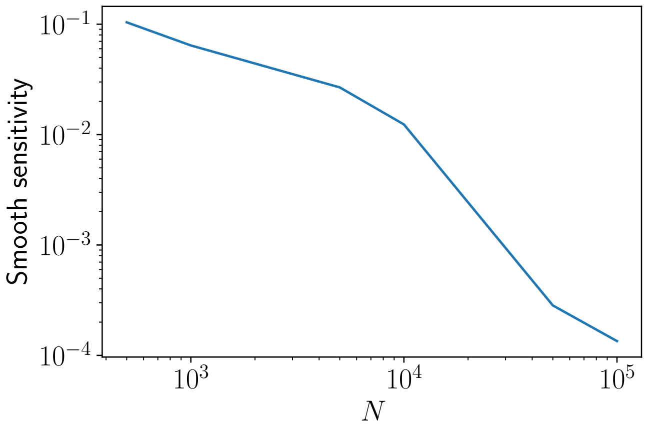

Fig. 6: Smooth sensitivity of Synth dataset with varying sample sizes (N) under $|\mathcal{X}| = 100$.

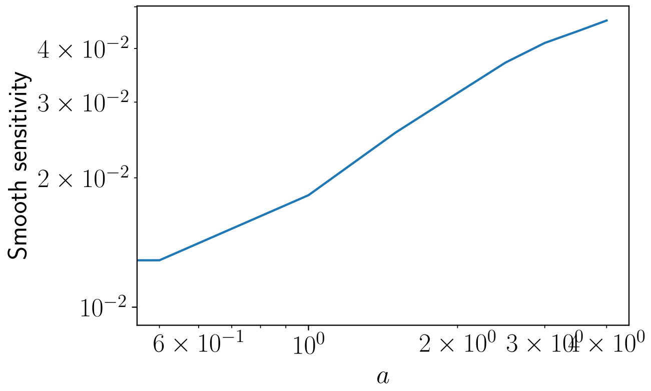

Fig. 7: Smooth sensitivity of Synth dataset with varying extent of imbalance (a; larger a yields more imbalanced dataset) under $N = 10000$ and $|\mathcal{X}| = 100$.

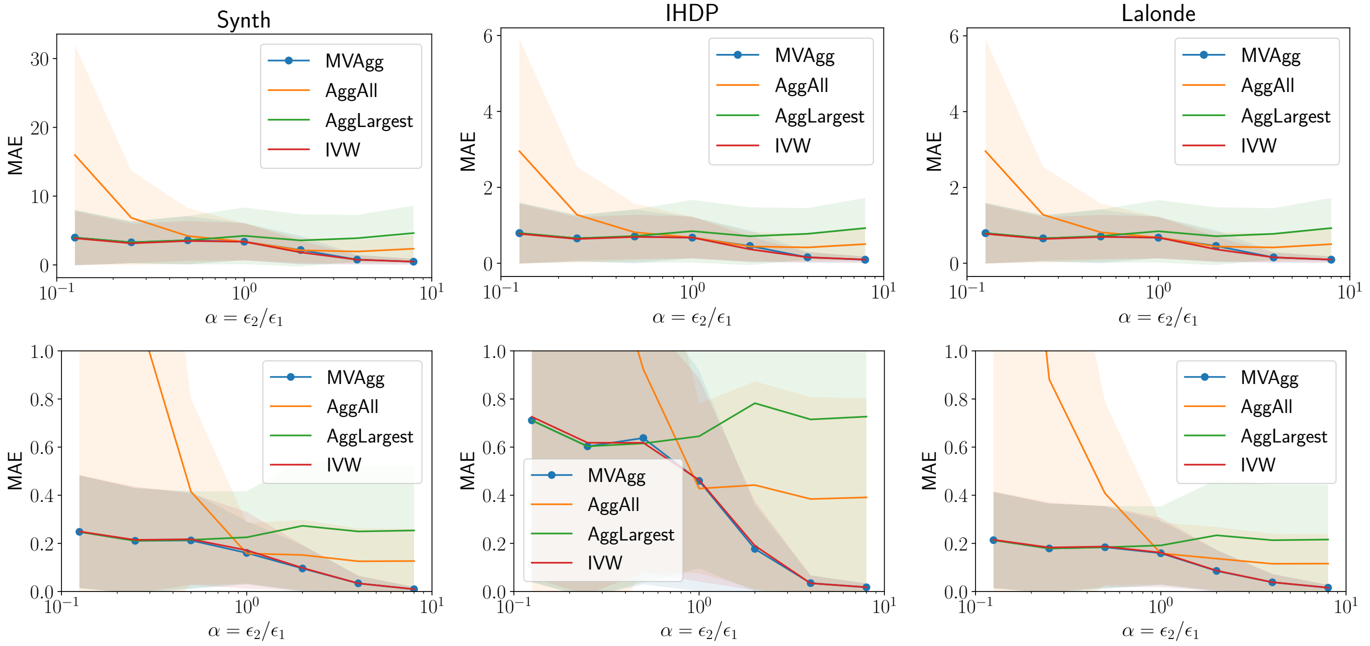

Fig. 8: Mean MAEs and standard deviations of MVAgg, AggAll, AggLargest, and IVW on Synth, IHDP, and Lalonde datasets (from left to right) under two-site setting. Upper row: GlobalDPMatching. Lower row: SmoothDPMatching. **Note that y-axis scales are different between upper and lower rows.**

B. Results

1) Randomized Trial and Difference-in-means Estimator: Figure 3 shows the mean MAEs on IST dataset under the two-site setting ($J = 2$). As α gets larger, the noise variance for the second site gets smaller while the one for the first site remains the same; thus, we generally expect the final ATE estimate to be never less accurate. We confirm this is true for all aggregation algorithms. We see that MVAgg generally achieves the best MAE among the four aggregation methods. For most of $\alpha = \epsilon_2/\epsilon_1$, its MAE matches with the better one of AggAll, AggLargest, and IVW. This suggests that when α is very small meaning the second site sends very noisy statistics, MVAgg discards the noisy site and only uses the results from the first site. On the other hand, when ϵ_2 is relatively large and the statistics from the second site are less noisy, MVAgg uses both sites to reduce a sampling error. We also observe the standard deviations of MVAgg are mostly the smallest, which is because MVAgg aims to minimize the variance of ATE estimate. AggAll performs the worst when $\epsilon_1 \gg \epsilon_2$, which supports our intuition that the noisy site can harm the final ATE estimation. The performance of AggLargest gets relatively worse as α gets larger since it does not utilize the accurate statistics from the second site. IVW performs almost comparably with MVAgg, but the

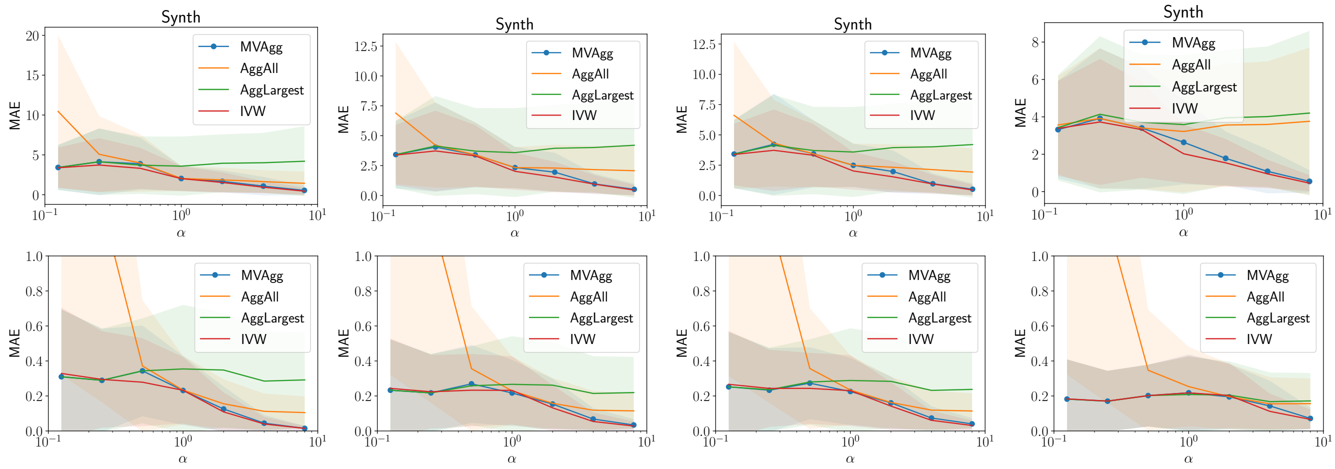

Fig. 9: Mean MAEs and standard deviations of MVAgg, AggAll, AggLargest, and IVW on Synth dataset under three-site setting. Upper low: GlobalDPMatching. Lower row: SmoothDPMatching. $N_1 : N_2 : N_3 = 1 : 1 : 1$ (left most), $3 : 2 : 1$ (middle left), $9 : 9 : 2$ (middle right), and $18 : 1 : 1$ (right most). **Note that y-axis scales are different between upper and lower rows.**

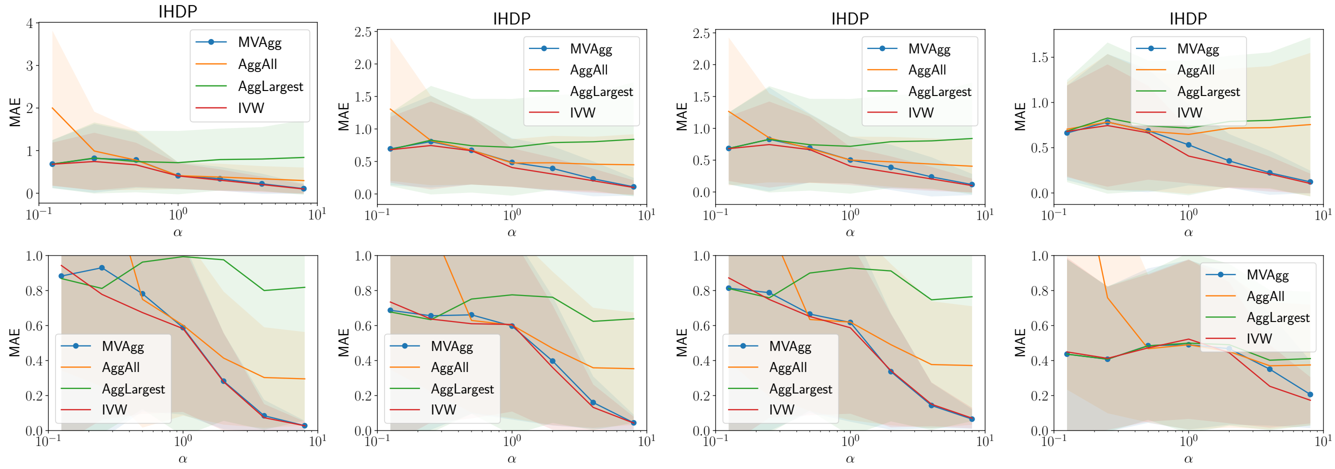

Fig. 10: Mean MAEs and standard deviations of MVAgg, AggAll, AggLargest, and IVW on IHDP dataset under three-site setting. Upper low: GlobalDPMatching. Lower row: SmoothDPMatching. $N_1 : N_2 : N_3 = 1 : 1 : 1$ (left most), $3 : 2 : 1$ (middle left), $9 : 9 : 2$ (middle right), and $18 : 1 : 1$ (right most). **Note that y-axis scales are different between upper and lower rows.**

performance is relatively worse as α gets smaller. This might be because the estimated DP variance from the second site gets noisier and it adds too much weight on the noisy site. Figure 4 shows the mean MAEs on IST dataset under the three-site setting ($J = 3$) with varying sample size proportions. The results exhibit similar trends to the two-site one. Most notably, MVAgg and IVW outperform AggAll and AggLargest in most of the cases. Comparing the results of different sample size proportions, we see the performance gap between MVAgg and AggLargest is maximized when the sample distribution across sites is uniform, i.e., $N_1 : N_2 : N_3 = 1 : 1 : 1$. This is because AggLargest cannot use large enough sites even when those sites have large enough ϵ's. On the other hand, the gap between MVAgg and AggAll for $\alpha \ll 1$ is largest when $N_1 : N_2 : N_3 = 18 : 1 : 1$. This is because AggAll weighs too much on the largest site, i.e., the first site, even when it has small ϵ_1, leading to noisier results than the ones obtained by removing the first site. This case particularly demonstrates the non-triviality of the problem—more samples do not necessarily help the final ATE estimation in the presence of DP noise. Although the gap is small, IVW performs generally better than MVAgg in this setting. IVW successfully assigns a fine-grained weight to each site.

Figure 5 shows the mean MAEs for STAR dataset, where the assignments to the four sites ($J = 4$) are pre-determined. We observe similar trends for all four aggregation algorithms to the case on IST dataset. Particularly on STAR dataset, MVAgg performs the best for $\alpha \gg 1$. This is because MVAgg

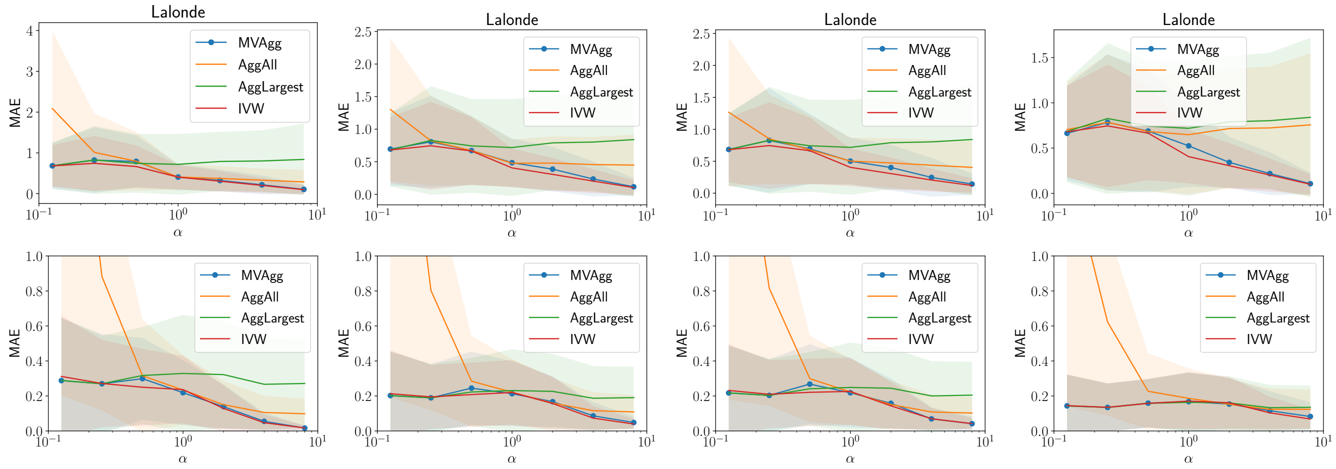

Fig. 11: Mean MAEs and standard deviations of MVAgg, AggAll, AggLargest, and IVW on Lalonde dataset under three-site setting. Upper low: GlobalDPMatching. Lower row: SmoothDPMatching. $N_1 : N_2 : N_3 = 1 : 1 : 1$ (left most), $3 : 2 : 1$ (middle left), $9 : 9 : 2$ (middle right), and $18 : 1 : 1$ (right most). **Note that y-axis scales are different between upper and lower rows.**

has more flexibility to choose the number of sites used, e.g., it can use sites 1 to 3 while AggAll and AggLargest cannot. For $\alpha < 1$, IVW outperforms MVAgg. Under more sites, the performance degradation due to wrongly assigning a large weight to a noisy site by IVW might be alleviated by other sites.

2) *Observational Study and Matching Estimator:* Figures 6 and 7 demonstrate how the smooth sensitivity on the Synth dataset changes along with the sample size N and the extent of imbalance, which is controlled by the parameter a (the larger a is, the more imbalanced the dataset is). Here, we measure β-smooth sensitivity for $\beta = \frac{\epsilon}{2\ln(\frac{2}{\delta})}$, where $\epsilon = 1$ and $\delta = 10^{-5}$, c.f., the smooth-sensitivity-based Laplace mechanism. We observe that the smooth sensitivity actually scales with $\approx \mathcal{O}(1/N)$ for a balanced dataset. We also see that it positively correlates with the extent of imbalance—it is the smallest when the data is well-balanced. Notice that even for imbalanced data, the smooth sensitivity is smaller than the global sensitivity ≈ 1.

Figures 8– 11 show the MAEs on three observational study datasets, Synth, IHDP, and Lalonde, when the ATE estimation algorithm is GlobalDPMatching or SmoothDPMatching under two-site (Figure 8) and three-site (Figure 9– 11) settings. Overall, we observe similar trends for all four aggregation algorithms to the case on the randomized trial datasets. Especially, for both ATE estimation algorithms and for all datasets, we see MVAgg and IVW outperform AggAll and AggLargest in general. Comparing between MVAgg and IVW, we observe that MVAgg performs the best when $\alpha \ll 1$, where there are disparities in site privacy budgets, while IVW achieves the best MAE when $\alpha \approx 1$, where privacy budgets are uniform across sites. Furthermore, we see the standard deviations of MVAgg are as small as the ones of IVW and smaller than those of AggAll and AggLargest. In particular, when $\alpha \gg 1$, they are much smaller than AggAll and AggLargest. These trends suggest that MVAgg discards the sites with small ϵ's, where the additive noises dominate the site outputs, and only uses the sites with large enough ϵ's, where the noises are negligible, to reduce a sampling error.

One main difference from the randomized trial case is the error scale. The global and smooth sensitivities of the matching estimator are larger than the global sensitivity of the difference-in-means estimator, which we use for randomized trials. Therefore, the ATE estimates at each site by the DP matching estimators, GlobalDPMatching and SmoothDPMatching, tend to be noisier, which results in higher MAEs. In such a case, it is more beneficial to use MVAgg or IVW instead of AggAll and AggLargest since the absolute gains in MAE are much larger.

Comparing the ATE estimation algorithms, we observe that SmoothDPMatching achieves much better performance (notice the scales of y-axis). The MAEs of GlobalDPMatching can be around 1 or more which is impermissible considering that $0 \leq \tau \leq 1$ as a result of preprocessing. However, SmoothDPMatching combined with MVAgg or IVW achieves MAEs less than 1 for all cases and even achieves MAEs around 0.1 or less when $\alpha \gg 1$. This indicates that our smooth sensitivity analysis enables us to dramatically reduce an additive DP noise variance and improve the privacy-utility tradeoff.

C. *Discussion*

Our results support the superiority of SmoothDPMatching over GlobalDPMatching, which happens because the smooth sensitivity is much smaller than the global sensitivity in practice. We also anticipate that the advantage of SmoothDPMatching is larger for well-balanced datasets.

Second, we find MVAgg achieves the best final ATE on both randomized trial and observational study data especially when there is a high disparity in the privacy budgets, compared with the other rule-based aggregation algorithms. This is because it reliably adopts the estimate at a site only when the quality is relatively high. The relative quality is hugely dependent on the data through sampling error, which we cannot know in advance. Thus, MVAgg provides a principled way to aggregate the estimates from multiple sites as opposed to some other rule-based aggregation algorithm, e.g., AggAll and AggLargest.

Finally, we find that site-level privacy parameters also have a high impact on performance. In particular, when all sites have comparable privacy, it is best to combine their estimates; on the other hand, if some sites have significantly higher privacy requirements, then it is best not to use those sites. We find that MVAgg reliably does this for a variety of privacy parameters. Furthermore, we note that MVAgg never outputs the impermissible outcome for any combination of privacy parameters across sites. Considering the risk of outputting very noisy final estimates with rule-based algorithms, it is recommended to use MVAgg in general while IVW can also be a good candidate when sites have almost uniform privacy budgets.

V. Conclusion and Future Work

We introduce a multi-site ATE estimation setting with per-site DP guarantees. We then provide a class of per-site ATE estimation algorithms which output both the private ATE estimate and its private variance estimate so that the central server aggregates the estimates from sites properly by looking at their qualities. In particular, for observational study data, we propose a novel DP matching estimator by analyzing the smooth sensitivity. We also propose an aggregation algorithm on the server that minimizes the variance of the final ATE estimate. Our experimental results demonstrate that our method, combining our site and server algorithms, automatically handles the heterogeneity across sites and provides a better privacy-utility tradeoff.

We believe our work is a first step towards enabling causal inference studies across multiple sites with formal privacy guarantees. One of the future directions is to consider how we can combine statistics from sites with different data distributions, e.g., children's hospitals and geriatric hospitals. Another direction would be studying other estimands, e.g., CATE, and other estimators, e.g., IPW.

Acknowledgments

TK and KC would like to thank NSF under 1804829, NSF under 2241100, NSF under 2217058, ARO MURI W911NF2110317, and ONR under N00014-20-1-2334 for research support. Also, TK is supported in part by Funai Overseas Fellowship. We would also like to thank our reviewers for their insightful feedback.

References

[1] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our Data, Ourselves: Privacy Via Distributed Noise Generation," in *Advances in Cryptology - EUROCRYPT 2006*, ser. Lecture Notes in Computer Science, S. Vaudenay, Ed. Berlin, Heidelberg: Springer, 2006, pp. 486–503.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography*, ser. Lecture Notes in Computer Science, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer, 2006, pp. 265–284.

[3] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statistical science : a review journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, pp. 1–21, Feb. 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943670/

[4] E. Shi, T.-H. H. Chan, E. Rieffel, R. Chow, and D. Song, "Privacy-Preserving Aggregation of Time-Series Data," 2011. [Online]. Available: https://www.semanticscholar.org/paper/Privacy-Preserving-Aggregation-of-Time-Series-Data-Shi-Chan/7cc53ef35f8398181bd09755ecc2fa8f52d0da1d

[5] A. Bittau, U. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld, "Prochlo: Strong Privacy for Analytics in the Crowd," in *Proceedings of the 26th Symposium on Operating Systems Principles*, ser. SOSP '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 441–459. [Online]. Available: https://dl.acm.org/doi/10.1145/3132747.3132769

[6] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev, "Distributed Differential Privacy via Shuffling," in *Advances in Cryptology – EUROCRYPT 2019*, ser. Lecture Notes in Computer Science, Y. Ishai and V. Rijmen, Eds. Cham: Springer International Publishing, 2019, pp. 375–403.

[7] G. Ács and C. Castelluccia, "I Have a DREAM! (DiffeRentially privatE smArt Metering)," in *Information Hiding*, ser. Lecture Notes in Computer Science, T. Filler, T. Pevný, S. Craver, and A. Ker, Eds. Berlin, Heidelberg: Springer, 2011, pp. 118–132.

[8] S. Goryczka and L. Xiong, "A Comprehensive Comparison of Multiparty Secure Additions with Differential Privacy," *IEEE transactions on dependable and secure computing*, vol. 14, no. 5, pp. 463–477, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5598559/

[9] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, ser. SIGMOD '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 735–746. [Online]. Available: https://dl.acm.org/doi/10.1145/1807167.1807247

[10] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The Privacy Blanket of the Shuffle Model," in *Advances in Cryptology – CRYPTO 2019*, ser. Lecture Notes in Computer Science, A. Boldyreva and D. Micciancio, Eds. Cham: Springer International Publishing, 2019, pp. 638–667.

[11] L. Chen, B. Ghazi, R. Kumar, and P. Manurangsi, "On Distributed Differential Privacy and Counting Distinct Elements," 2021.

[12] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: from local to central differential privacy via anonymity," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '19. USA: Society for Industrial and Applied Mathematics, 2019, pp. 2468–2479.

[13] B. Ghazi, R. Pagh, and A. Velingker, "Scalable and Differentially Private Distributed Aggregation in the Shuffled Model," Dec. 2019, arXiv:1906.08320 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1906.08320

[14] B. Ghazi, R. Kumar, P. Manurangsi, and R. Pagh, "Private Counting from Anonymous Messages: Near-Optimal Accuracy with Vanishing Communication Overhead," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 3505–3514, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v119/ghazi20a.html

[15] B. Ghazi, N. Golowich, R. Kumar, P. Manurangsi, R. Pagh, and A. Velingker, "Pure Differentially Private Summation from Anonymous Messages," 2020.

[16] S. K. Lee, L. Gresele, M. Park, and K. Muandet, "Privacy-Preserving Causal Inference via Inverse Probability Weighting," *arXiv:1905.12592*

[cs, stat], Nov. 2019, arXiv: 1905.12592. [Online]. Available: http://arxiv.org/abs/1905.12592

[17] M. J. Kusner, Y. Sun, K. Sridharan, and K. Q. Weinberger, "Private Causal Inference," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. PMLR, May 2016, pp. 1308–1317, iSSN: 1938-7228. [Online]. Available: https://proceedings.mlr.press/v51/kusner16.html

[18] D. Xu, S. Yuan, and X. Wu, "Differential Privacy Preserving Causal Graph Discovery," in *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, Aug. 2017, pp. 60–71.

[19] F. Niu, H. Nori, B. Quistorff, R. Caruana, D. Ngwe, and A. Kannan, "Differentially Private Estimation of Heterogeneous Causal Effects," Feb. 2022, arXiv:2202.11043 [cs, econ, stat]. [Online]. Available: http://arxiv.org/abs/2202.11043

[20] T. Komarova and D. Nekipelov, "Identification and Formal Privacy Guarantees," Rochester, NY, Oct. 2022. [Online]. Available: https://papers.ssrn.com/abstract=3635824

[21] J. L. Fleiss, "Analysis of data from multiclinic trials," *Controlled Clinical Trials*, vol. 7, no. 4, pp. 267–275, Dec. 1986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0197245686900346

[22] S. W. Raudenbush and X. Liu, "Statistical power and optimal design for multisite randomized trials," *Psychological Methods*, vol. 5, no. 2, pp. 199–213, Jun. 2000.

[23] H. C. Kraemer, "Pitfalls of Multisite Randomized Clinical Trials of Efficacy and Effectiveness," *Schizophrenia Bulletin*, vol. 26, no. 3, pp. 533–541, Jan. 2000. [Online]. Available: https://doi.org/10.1093/oxfordjournals.schbul.a033474

[24] M. Weinberger, E. Z. Oddone, W. G. Henderson, D. M. Smith, J. Huey, A. Giobbie-Hurder, and J. R. Feussner, "Multisite Randomized Controlled Trials in Health Services Research: Scientific Challenges and Operational Issues," *Medical Care*, vol. 39, no. 6, pp. 627–634, 2001, publisher: Lippincott Williams & Wilkins. [Online]. Available: https://www.jstor.org/stable/3767637

[25] M. J. Weiss, H. S. Bloom, N. Verbitsky-Savitz, H. Gupta, A. E. Vigil, and D. N. Cullinan, "How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials," *Journal of Research on Educational Effectiveness*, vol. 10, no. 4, pp. 843–876, Oct. 2017, publisher: Routledge _eprint: https://doi.org/10.1080/19345747.2017.1300719. [Online]. Available: https://doi.org/10.1080/19345747.2017.1300719

[26] S. E. Robertson, J. A. Steingrimsson, N. R. Joyce, E. A. Stuart, and I. J. Dahabreh, "Center-specific causal inference with multicenter trials: reinterpreting trial evidence in the context of each participating center," *arXiv:2104.05905 [stat]*, Apr. 2021, arXiv: 2104.05905. [Online]. Available: http://arxiv.org/abs/2104.05905

[27] N. Dong, B. Kelcey, and J. Spybrook, "Design Considerations in Multisite Randomized Trials Probing Moderated Treatment Effects," *Journal of Educational and Behavioral Statistics*, vol. 46, no. 5, pp. 527–559, Oct. 2021, publisher: American Educational Research Association. [Online]. Available: https://doi.org/10.3102/1076998620961492

[28] K. Crammer, M. Kearns, and J. Wortman, "Learning from Data of Variable Quality," in *Advances in Neural Information Processing Systems*, vol. 18. MIT Press, 2005. [Online]. Available: https://proceedings.neurips.cc/paper/2005/hash/465636eb4a7ff4b267f3b765d07a02da-Abstract.html

[29] S. Song, K. Chaudhuri, and A. Sarwate, "Learning from Data with Heterogeneous Noise using SGD," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. PMLR, Feb. 2015, pp. 894–902, iSSN: 1938-7228. [Online]. Available: https://proceedings.mlr.press/v38/song15.html

[30] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, Aug. 2014. [Online]. Available: https://doi.org/10.1561/0400000042

[31] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, ser. STOC '07. New York, NY, USA: Association for Computing Machinery, Jun. 2007, pp. 75–84. [Online]. Available: https://doi.org/10.1145/1250790.1250803

[32] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. Nitin Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1-2, pp. 1–210, 2021. [Online]. Available: https://doi.org/10.1561/2200000083

[33] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 603–618. [Online]. Available: https://doi.org/10.1145/3133956.3134012

[34] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html

[35] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning," in *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*. IEEE, 2019, pp. 739–753.

[36] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning," *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 2512–2520, Apr. 2019, conference Name: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications ISBN: 9781728105154 Place: Paris, France Publisher: IEEE. [Online]. Available: https://ieeexplore.ieee.org/document/8737416/

[37] C. Ma, J. Li, M. Ding, H. H. Yang, F. Shu, T. Q. S. Quek, and H. V. Poor, "On Safeguarding Privacy and Security in the Framework of Federated Learning," *IEEE Network*, vol. 34, no. 4, pp. 242–248, Jul. 2020, conference Name: IEEE Network.

[38] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1st ed. Cambridge University Press, Apr. 2015. [Online]. Available: https://www.cambridge.org/core/product/identifier/9781139025751/type/book

[39] "The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke," *The Lancet*, vol. 349, no. 9065, pp. 1569–1581, May 1997, publisher: Elsevier. [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(97)04011-7/fulltext

[40] E. Word and A. Others, "Student/Teacher Achievement Ratio (STAR) Tennessee's K-3 Class Size Study. Final Summary Report 1985-1990," Tech. Rep., 1990, eRIC Number: ED320692.

[41] J. L. Hill, "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, Jan. 2011, publisher: Taylor & Francis _eprint: https://doi.org/10.1198/jcgs.2010.08162. [Online]. Available: https://doi.org/10.1198/jcgs.2010.08162

[42] R. Lalonde, "Evaluating the Econometric Evaluations of Training Programs with Experiment Data," *American Economic Review*, vol. 76, pp. 604–20, Feb. 1986.

```
Algorithm 4: DP difference-in-means estimator at site
Data: Dataset: $D = \{(Y_i(1), Y_i(0))\}_{i=1}^N$, Number of treated/control individuals: $N_t$, $N_c$, Privacy parameter: $\epsilon$
Sample $W_1, \ldots, W_N$ according to Eq. (2)
Observe outcomes: $\{Y_i^{\text{obs}} = W_i Y_i(1) + (1 - W_i) Y_i(0)\}_{i=1}^N$
Compute average outcome of treated individuals with DP: $\bar{Y}_{t-\text{DP}}(D) = \frac{1}{N_t}\left(\sum_{i:W_i=1} Y_i^{\text{obs}} + \xi_t\right)$, where $\xi_t \sim \text{Lap}(B/\epsilon)$
Compute average outcome of control individuals with DP: $\bar{Y}_{c-\text{DP}}(D) = \frac{1}{N_c}\left(\sum_{i:W_i=0} Y_i^{\text{obs}} + \xi_c\right)$, where
  $\xi_c \sim \text{Lap}(B/\epsilon)$
return $\hat{\tau}_{\text{DP}} = \bar{Y}_{t-\text{DP}}(D) - \bar{Y}_{c-\text{DP}}(D)$
```

Appendix A
Randomized Trial Data & Difference-in-means Estimator

A. Privacy Guarantee

In randomized trials, the number of treated individuals N_t ($N_c = N - N_t$) is a fixed public information, and the treatment indicators W_i's are assigned randomly by the site so that $\sum_i W_i = N_t$. More specifically,

$$\Pr[W_1, \ldots, W_N] = \begin{cases} \frac{1}{\binom{N}{N_t}} & \text{if} \quad \sum_i W_i = N_t \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Note that this assignment does not depend on the individuals' potential outcomes, $Y_i(1)$'s and $Y_i(0)$'s. Recall that we define a dataset to be a set of individual potential outcomes, i.e., $D = \{(Y_i(1), Y_i(0))\}_{i=1}^N$ in randomized trials. This definition implies that neighboring datasets differ by one person's potential outcomes.

Under this definition of D, we show that our DP version of the difference-in-means estimator satisfies ϵ-DP. Recall that our DP estimator is as follows:

$$\begin{aligned} \hat{\tau}_{\text{DP}} &= \frac{1}{N_t}\left(\sum_{i:W_i=1} Y_i^{\text{obs}} + \xi_t\right) - \frac{1}{N_c}\left(\sum_{i:W_i=0} Y_i^{\text{obs}} + \xi_c\right) \\ &= \hat{\tau} + \frac{\xi_t}{N_t} - \frac{\xi_c}{N_c}, \end{aligned}$$

where $\xi_t, \xi_c \sim \text{Lap}(B/\epsilon)$.

Since the treatment indicators W_i's are assigned independently of personal information, one could see them as public information. In such a case, one can use the parallel composition since the change in one record of D only changes the value of either $\sum_{i:W_i=1} Y_i^{\text{obs}}$ or $\sum_{i:W_i=0} Y_i^{\text{obs}}$.

We can also show that the estimator satisfies ϵ-DP when W_i's are sensitive information and not public. In particular, we consider the randomized mechanism which explicitly involves the sampling of W_i's as in Algorithm 4 and show that it satisfies ϵ-DP. It suffices to show that for any neighboring dataset D, D' and any subset S of the image of $(\bar{Y}_{t-\text{DP}}, \bar{Y}_{c-\text{DP}})$,

$$\Pr[(\bar{Y}_{t-\text{DP}}(D), \bar{Y}_{c-\text{DP}}(D)) \in S] \leq e^{\epsilon} \Pr[(\bar{Y}_{t-\text{DP}}(D'), \bar{Y}_{c-\text{DP}}(D')) \in S].$$

We show it by marginalizing over $W_1, \ldots, W_N$ and observing the fact that $(\bar{Y}_{t-\text{DP}}, \bar{Y}_{c-\text{DP}})$ satisfies ϵ-DP for a fixed $W_1, \ldots, W_N$, which is equivalent to the case where W_i's are public.

Proof.

$$\begin{aligned} &\Pr[(\bar{Y}_{t-\text{DP}}(D), \bar{Y}_{c-\text{DP}}(D)) \in S] \\ =&\sum_{W_1,\ldots,W_N} \Pr[W_1, \ldots, W_N] \cdot \Pr[(\bar{Y}_{t-\text{DP}}(D), \bar{Y}_{c-\text{DP}}(D)) \in S | W_1, \ldots, W_N] \\ \leq&\sum_{W_1,\ldots,W_N} \Pr[W_1, \ldots, W_N] \cdot e^{\epsilon} \Pr[(\bar{Y}_{t-\text{DP}}(D'), \bar{Y}_{c-\text{DP}}(D')) \in S | W_1, \ldots, W_N] \\ =&e^{\epsilon} \Pr[(\bar{Y}_{t-\text{DP}}(D'), \bar{Y}_{c-\text{DP}}(D')) \in S]. \end{aligned}$$

□

Finally, by the post-processing theorem of DP, our DP difference-in-means estimator satisfies ϵ-DP for randomized trials. One can apply the same argument to the variance estimation as well.

APPENDIX B
OBSERVATIONAL STUDY DATA & MATCHING ESTIMATOR

Since the treatment indicator W_i depends on the covariates X_i, without loss of generality, we use $D = \{W_i, Y_i^{\mathrm{obs}}, X_i\}_{i=1}^N$ as the dataset definition in this section. The neighboring datasets can differ by the treatment assignment as well as the outcome and/or covariates.

A. Detail on Modification of Matching Estimator

We modify the exact single matching estimator so as to balance the number of individuals matched to a particular individual in the same covariate stratum. The reason for doing this is to obtain the tighter bound of the sensitivity. In particular, suppose for a stratum $X = x$, we have $|T_x|$ treated and $|C_x|$ control individuals. Without specifying any, the first control individual can be matched with all $|T_x|$ treated in the worst case (left side of Figure 2 in the main paper). Thus, the first one contributes $|T_x| + 1$ times to the final estimate which leads to larger sensitivity and more DP noise vaiance. As such, we avoid such cases by greedily balancing the number of matches for each individual (right side of Figure 2 in the main paper) within each covariate stratum. More specifically, i-th treated individual is matched with $i \bmod |C_x|$-th control, and j-th control individual is matched with $j \bmod |T_x|$-th treated. This way we guarantee that each treated (or control) individual contributes up to $\lceil |C_x|/|T_x| \rceil + 1$ (or $\lceil |T_x|/|C_x| \rceil + 1$) times to the final estiamte. As a result, we obtain the tighter sensitivity bound. Note that we only specify how to handle individuals with the same covariate in the exact single matching; thus, no bias is introduced by our modification.

B. Smooth Sensitivity of Matching Estimator

1) Proof of Theorem 1:

Proof. Let L_i be the number of individuals matched with i-th individual, namely, the number of individuals who use Y_i^{obs} as the imputed value. By our modification to the estimator in the main paper, we have $\lfloor \frac{|C_x|}{|T_x|} \rfloor \le L_i \le \lceil \frac{|C_x|}{|T_x|} \rceil$ when $X_i = x$ and $W_i = 1$. Then, the exact single matching estimator is written as below.

$$\hat{\tau}(D) = \frac{1}{N} \sum_x \sum_{i \in T_x} (1 + L_i) Y_i^{\mathrm{obs}} - \sum_{i \in C_x} (1 + L_i) Y_i^{\mathrm{obs}}$$

Let $f(D) = N \cdot \hat{\tau}(D)$. Since $\mathrm{LS}_{\hat{\tau}}(D) = \frac{1}{N}\mathrm{LS}_f(D)$, we instead consider the local sensitivity of f.

Additionally, let $d_{ar}(D, D')$ be the minimum number of row addition/removal from D to obtain D'. Furthermore, let $d_{ar}^{+}(D, D') \le 1 \Leftrightarrow d_{ar}(D, D') \le 1 \land |D'| = |D| + 1$ and define $\mathrm{LS}_f^{+}(D) = \max_{D' : d_{ar}^{+}(D,D') \le 1} |f(D) - f(D')|$. Similarly, let $d_{ar}^{-}(D, D') \le 1 \Leftrightarrow d_{ar}(D, D') \le 1 \land |D'| = |D| - 1$ and define $\mathrm{LS}_f^{-}(D) = \max_{D' : d_{ar}^{-}(D,D') \le 1} |f(D) - f(D')|$. Then, by the triangle inequality, it holds that

$$\begin{aligned}
\mathrm{LS}_f(D) &\le \max_{D'' : d_{ar}^{+}(D, D'') \le 1} \max_{D' : d_{ar}^{-}(D'', D') \le 1} |f(D) - f(D'')| + |f(D'') - f(D')| \\
&= \max_{D'' : d_{ar}^{+}(D, D'') \le 1} (|f(D) - f(D'')| + \max_{D' : d_{ar}^{-}(D'', D') \le 1} |f(D'') - f(D')|) \\
&\le \mathrm{LS}_f^{+}(D) + \max_{D'' : d_{ar}^{+}(D, D'') \le 1} \mathrm{LS}_f^{-}(D'')
\end{aligned}$$

We first consider $\mathrm{LS}_f^{+}(D)$. We write the neighboring dataset $D' = \{(W_i', Y_i^{\mathrm{obs}\prime}, X_i')\}_{i=1}^{N+1}$ and also define T_x', C_x', and L_i' accordingly. W.l.o.g., $D_{N+1} \in D'$ is the added individual data. Let x be $x = X_{N+1}'$.

Here we assume w.l.o.g. $W_{N+1} = 1$. When C_x, T_x are non-empty, it holds that

$$\begin{aligned}
|f(D) - f(D')| &= |(\sum_{i \in T_x} (1 + L_i) Y_i^{\mathrm{obs}} - \sum_{i \in C_x} (1 + L_i) Y_i^{\mathrm{obs}}) - (\sum_{i \in T_x'} (1 + L_i') Y_i^{\mathrm{obs}\prime} - \sum_{i \in C_x'} (1 + L_i') Y_i^{\mathrm{obs}\prime})| \\
&= | - (1 + L_{N+1}') Y_{N+1}^{\mathrm{obs}\prime} + \sum_{i \in T_x} (L_i - L_i') Y_i^{\mathrm{obs}} - \sum_{i \in C_x} (L_i - L_i') Y_i^{\mathrm{obs}}| \\
&\le |(1 + L_{N+1}') Y_{N+1}^{\mathrm{obs}\prime}| + |\sum_{i \in T_x} (L_i - L_i') Y_i^{\mathrm{obs}}| + |\sum_{i \in C_x} (L_i - L_i') Y_i^{\mathrm{obs}}| \\
&\le (1 + L_{N+1}') B + L_{N+1}' B + B \\
&= 2(1 + L_{N+1}') B \\
&\le 2(1 + \lceil \frac{|C_x'|}{|T_x'|} \rceil) B = 2(1 + \lceil \frac{|C_x|}{|T_x| + 1} \rceil) B.
\end{aligned}$$

The second to last inequality holds due to properties achieved by the (greedy) exact single matching estimator: (1) for $i \in T_x$, $L_i \geq L'_i$, (2)$\sum_{i \in T_x} L_i - L'_i = L'_{N+1}$ since $\sum_{i \in T_x} L_i = \sum_{i \in T'_x} L'_i = |C_x|$, and (3) for $i \in C_x$, there exists only one $i' \in C_x$ s.t. $L_{i'} - L'_{i'} = -1$ and for $i \neq i'$, $L_i = L'_i$. In particular, the second term is bounded as follows.

$$|\sum_{i \in T_x} (L_i - L'_i) Y_i^{\text{obs}}| \leq \sum_{i \in T_x} |L_i - L'_i||Y_i^{\text{obs}}| \leq B \sum_{i \in T_x} |L_i - L'_i| = B \sum_{i \in T_x} (L_i - L'_i) = B L'_{N+1}$$

The last inequality holds the exact single matching estimator balances the number of matches.

When C_x is empty, $|f(D) - f(D')| = 0$. When T_x is empty and C_x is not empty, $|f(D) - f(D')| = |Y_{N+1}^{\text{obs}} - \hat{Y}_{N+1}(0) - \sum_{i \in C_x} Y_{N+1}^{\text{obs}} - Y_i^{\text{obs}}| \leq 2(|C_x| + 1)B$.

Therefore, by the symmetry, the following holds.

$$\text{LS}_f^+(D) = \max_x \begin{cases} 0 & |T_x| = |C_x| = 0 \\ 2(1 + \max(\lceil \frac{|C_x|}{|T_x|+1} \rceil, \lceil \frac{|T_x|}{|C_x|+1} \rceil))B & o.w. \end{cases}$$

With similar arguments, we have the following.

$$\text{LS}_f^-(D) = \max_x \begin{cases} 0 & |T_x| = 0 \vee |C_x| = 0 \\ 2(1 + \max(\lceil \frac{|C_x|}{|T_x|} \rceil, \lceil \frac{|T_x|}{|C_x|} \rceil))B & o.w. \end{cases}$$

Therefore,

$$\max_{D'': d_{ar}^+(D, D'') \leq 1} \text{LS}_f^-(D'') = \max_x \begin{cases} 0 & |T_x| = |C_x| = 0 \\ 2(1 + |C_x|)B & |T_x| = 0 \wedge |C_x| > 0 \\ 2(1 + |T_x|)B & |T_x| > 0 \wedge |C_x| = 0 \\ 2(1 + \max(\lceil \frac{1+|C_x|}{|T_x|} \rceil, \lceil \frac{1+|T_x|}{|C_x|} \rceil))B & o.w. \end{cases}$$

Finally, by combining above, we have the upper bound on the local sensitivity as follows.

$$\begin{aligned} \text{LS}_{\hat{\tau}}(D) &\leq \frac{1}{N} \max_x \begin{cases} 0 & |T_x| = |C_x| = 0 \\ 4(1 + |C_x|)B & |T_x| = 0 \wedge |C_x| > 0 \\ 4(1 + |T_x|)B & |T_x| > 0 \wedge |C_x| = 0 \\ 2(2 + \max(\lceil \frac{|C_x|}{|T_x|+1} \rceil, \lceil \frac{|T_x|}{|C_x|+1} \rceil) + \max(\lceil \frac{1+|C_x|}{|T_x|} \rceil, \lceil \frac{1+|T_x|}{|C_x|} \rceil))B & o.w. \end{cases} \\ &= \frac{1}{N} \max_x \begin{cases} 0 & |T_x| = |C_x| = 0 \\ 4(1 + |C_x|)B & |T_x| = 0 \wedge |C_x| > 0 \\ 4(1 + |T_x|)B & |T_x| > 0 \wedge |C_x| = 0 \\ 2(2 + \max(\lceil \frac{|C_x|}{|T_x|+1} \rceil + \lceil \frac{1+|C_x|}{|T_x|} \rceil, \lceil \frac{|T_x|}{|C_x|+1} \rceil + \lceil \frac{1+|T_x|}{|C_x|} \rceil))B & o.w. \end{cases} \\ &\leq \frac{1}{N} \max_x \begin{cases} 0 & |T_x| = |C_x| = 0 \\ 4(1 + |C_x|)B & |T_x| = 0 \wedge |C_x| > 0 \\ 4(1 + |T_x|)B & |T_x| > 0 \wedge |C_x| = 0 \\ 4(1 + \max(\lceil \frac{1+|C_x|}{|T_x|} \rceil, \lceil \frac{1+|T_x|}{|C_x|} \rceil))B & o.w. \end{cases} \end{aligned}$$

The local sensitivity depends only on $|T_x|$ and $|C_x|$, and thus, the smooth sensitivity is obtained as follows. Let $R_x(D)$ satisfy $\text{LS}(D) = \frac{4B}{N}(1 + \max_l R_x(D))$. Also, let $R_x^{(k)}(D) = \max_{D': d(D, D') \leq k} R_x(D')$. Then,

$$\begin{aligned} A^{(k)}(D) &= \max_{D': d(D, D') \leq k} \text{LS}(D') \\ &= \frac{4B}{N}(1 + \max_x R_x^{(k)}(D)). \end{aligned}$$

Here, $R_x^{(k)}(D)$ is as follows.

$$R_x(k) = \begin{cases} |T_x| + k & |T_x| \geq |C_x| \wedge k \geq |C_x| \\ \lceil \frac{|T_x| + k + 1}{|C_x| - k} \rceil & |T_x| \geq |C_x| \wedge k < |C_x| \\ |C_x| + k & |C_x| \geq |T_x| \wedge k \geq |T_x| \\ \lceil \frac{|C_x| + k + 1}{|T_x| - k} \rceil & |C_x| \geq |T_x| \wedge k < |T_x| \end{cases}$$

Thus, the β-smooth sensitivity is

$$S^*_{\hat{\tau},\beta}(D) = \max_{k=0,\dots,N} e^{-k\beta} \frac{4B}{N}(1 + \max_x R_x^{(k)}(D)).$$

This can be computed by storing $|T_x|$ and $|C_x|$ for each $x \in \mathcal{X}$, which is present in the dataset, and enumerating over k; thus, we need $\mathcal{O}(\min(|\mathcal{X}|, N))$ space and $\mathcal{O}(N \cdot \min(|\mathcal{X}|, N))$ time. Note that for $x \in \mathcal{X}$ which is not present in th dataset $R_x^{(k)}(D)$ is constant for fixed k, i.e., $R_x^{(k)}(D) = k$. □

C. Differentially Private Variance Estimation

Let $\hat{\tau}_{\mathrm{DP}}$ be our DP matching estimator, SmoothDPMatching. Recall that the variance of $\hat{\tau}_{\mathrm{DP}}$ is $\mathbb{V}[\hat{\tau}_{\mathrm{DP}}] = \mathbb{V}[\hat{\tau}] + \mathbb{V}[(2S^*_{\hat{\tau},\beta}(D)/\epsilon) \cdot \eta] = \mathbb{V}[\hat{\tau}] + 8(S^*_{\hat{\tau},\beta}(D))^2/\epsilon^2$. Since both terms are data-dependant, we need to estimate them from data privately.

We guarantee (ϵ_2, δ_2)-DP and (ϵ_3, δ_3)-DP separately for each term. Suppose the private ATE estimation is done with (ϵ_1, δ_1)-DP. Then, publishing the private ATE estimate and the variance estimate satisfies $(\epsilon_1 + \epsilon_2 + \epsilon_3, \delta_1 + \delta_2 + \delta_3)$-DP in total.

1) Smooth Sensitivity of Variance of Matching Estimator: By Section 19 of [38], we have the following variance estimate of the exact single matching estimator, $\hat{\mathbb{V}}[\hat{\tau}]$:

$$\hat{\mathbb{V}}[\hat{\tau}] = \frac{1}{2N^2} \sum_x \left\{ \sum_{i \in T_x} (1 + L_i)^2 (\hat{Y}_i(1) - \hat{Y}_i(0))^2 + \sum_{i \in C_x} (1 + L_i)^2 (\hat{Y}_i(1) - \hat{Y}_i(0))^2 \right\}.$$

We produce DP version of the variance estimate by adding noise calibrated to the smooth sensitivity of this quantity. In the rest of this section, we present its smooth sensitivity.

As in Section B-B1, we consider the local sensitivity of the unnormalized variance estimate, denoting by g, i.e., $g(D) = 2N^2 \cdot \hat{\mathbb{V}}[\hat{\tau}]$. Thus, we have $\mathrm{LS}_{\hat{\mathbb{V}}[\hat{\tau}]}(D) = \frac{1}{2N^2} \mathrm{LS}_g(D)$. We also upper bound $\mathrm{LS}_g(D) \le \mathrm{LS}_g^+(D) + \max_{D'' : d^+_{ar}(D, D'') \le 1} \mathrm{LS}_g^-(D'')$.

We first consider $\mathrm{LS}_g^+(D)$. We write the neighboring dataset $D' = \{(W_i', Y_i^{\mathrm{obs}\prime}, X_i')\}_{i=1}^{N+1}$ and also define T_x', C_x', and L_i' accordingly. W.l.o.g., $D_{N+1} \in D'$ is the added individual data. Let x be $x = X'_{N+1}$.

Here we assume w.l.o.g. $W_{N+1} = 1$. When C_x, T_x are non-empty, it holds that

$$\begin{aligned}
|g(D) - g(D')| &= |\sum_{i \in T_x \cup C_x} (1 + L_i)^2 (\hat{Y}_i(1) - \hat{Y}_i(0))^2 - \sum_{i \in T_x' \cup C_x'} (1 + L_i')^2 (\hat{Y}_i'(1) - \hat{Y}_i'(0))^2| \\
&= |\sum_{i \in T_x \cup C_x} \left\{ (1 + L_i)^2 (\hat{Y}_i(1) - \hat{Y}_i(0))^2 - (1 + L_i')^2 (\hat{Y}_i'(1) - \hat{Y}_i'(0))^2 \right\} \\
&\quad - (1 + L'_{N+1})^2 (\hat{Y}'_{N+1}(1) - \hat{Y}'_{N+1}(0))^2| \\
&\le B^2 (\sum_{i \in T_x \cup C_x} \{(1 + L_i)^2 + (1 + L_i')^2\} + (1 + L'_{N+1})^2) \\
&\le B^2 (|T_x|((1 + \lceil \frac{|C_x|}{|T_x|} \rceil)^2 + (1 + \lceil \frac{|C_x'|}{|T_x'|} \rceil)^2) + |C_x|((1 + \lceil \frac{|T_x|}{|C_x|} \rceil)^2 + (1 + \lceil \frac{|T_x'|}{|C_x'|} \rceil)^2) \\
&\quad + (1 + \lceil \frac{|C_x'|}{|T_x'|} \rceil)^2) \\
&\le B^2 (2|T_x|((1 + \lceil \frac{|C_x|}{|T_x|} \rceil)^2 + 2|C_x|((1 + \lceil \frac{|T_x| + 1}{|C_x|} \rceil)^2 + (1 + \lceil \frac{|C_x|}{|T_x| + 1} \rceil)^2)
\end{aligned}$$

When C_x is empty, $|g(D) - g(D')| = 0$. When T_x is empty and C_x is not empty, $|g(D) - g(D')| \le B^2((1 + |C_x|)^2 + |C_x|(1 + 1)^2) = B^2((1 + |C_x|)^2 + 4|C_x|)$.

Thus, by symmetry, we have the following upper bound on $\mathrm{LS}_g^+(D)$.

$$\mathrm{LS}_g^+(D) \le \max_x \begin{cases} 0 & |T_x| = |C_x| = 0 \\ B^2((1 + |C_x|)^2 + 4|C_x|) & |T_x| = 0 \land |C_x| > 0 \\ B^2((1 + |T_x|)^2 + 4|T_x|) & |T_x| > 0 \land |C_x| = 0 \\ B^2 \max(2|T_x|((1 + \lceil \frac{|C_x|}{|T_x|} \rceil)^2 + 2|C_x|((1 + \lceil \frac{|T_x|+1}{|C_x|} \rceil)^2 + (1 + \lceil \frac{|C_x|}{|T_x|+1} \rceil)^2, & \\ 2|C_x|((1 + \lceil \frac{|T_x|}{|C_x|} \rceil)^2 + 2|T_x|((1 + \lceil \frac{|C_x|+1}{|T_x|} \rceil)^2 + (1 + \lceil \frac{|T_x|}{|C_x|+1} \rceil)^2) & o.w. \end{cases}$$

Similarly, we have the following.

$$\mathrm{LS}_g^-(D) \le \max_x \begin{cases} 0 & |T_x| = 0 \lor |C_x| = 0 \\ B^2 \max(2(1 + |T_x|)((1 + \lceil \frac{|C_x|}{|T_x|} \rceil)^2 + 2|C_x|((1 + \lceil \frac{|T_x|+1}{|C_x|} \rceil)^2 + (1 + \lceil \frac{|C_x|}{|T_x|} \rceil)^2, & \\ 2(1 + |C_x|)((1 + \lceil \frac{|T_x|}{|C_x|} \rceil)^2 + 2|T_x|((1 + \lceil \frac{|C_x|+1}{|T_x|} \rceil)^2 + (1 + \lceil \frac{|T_x|}{|C_x|} \rceil)^2) & o.w. \end{cases}$$

Therefore,

$$\begin{aligned}
&\max_{D'':d^+_{ar}(D,D'')\leq 1} \mathrm{LS}^-_g(D'') \\
&= \max_x \begin{cases}
0 & |T_x| = |C_x| = 0 \\
B^2 \max(8(1+|C_x|) + 2(2+|C_x|)^2 + 4, & \\
4(1+|C_x|)^2 + 2|C_x|((1+\lceil\frac{2}{|C_x|}\rceil)^2 + (1+|C_x|)^2) & |T_x| = 0 \wedge |C_x| > 0 \\
B^2 \max(8(1+|T_x|) + 2(2+|T_x|)^2 + 4, & \\
4(1+|T_x|)^2 + 2|T_x|((1+\lceil\frac{2}{|T_x|}\rceil)^2 + (1+|T_x|)^2) & |T_x| > 0 \wedge |C_x| = 0 \\
B^2 \max(2(1+|C_x|)(1+\lceil\frac{1+|T_x|}{|C_x|}\rceil)^2 + 2(1+|T_x|)(1+\lceil\frac{1+|C_x|}{1+|T_x|}\rceil)^2 + (1+\lceil\frac{1+|T_x|}{|C_x|}\rceil)^2, & \\
2(2+|T_x|)(1+\lceil\frac{|C_x|}{1+|T_x|}\rceil)^2 + 2|C_x|(1+\lceil\frac{2+|T_x|}{|C_x|}\rceil)^2 + (1+\lceil\frac{|C_x|}{1+|T_x|}\rceil)^2, & \\
2(1+|T_x|)(1+\lceil\frac{1+|C_x|}{|T_x|}\rceil)^2 + 2(1+|C_x|)(1+\lceil\frac{1+|T_x|}{1+|C_x|}\rceil)^2 + (1+\lceil\frac{1+|C_x|}{|T_x|}\rceil)^2, & \\
2(2+|C_x|)(1+\lceil\frac{|T_x|}{1+|C_x|}\rceil)^2 + 2|T_x|(1+\lceil\frac{2+|C_x|}{|T_x|}\rceil)^2 + (1+\lceil\frac{|T_x|}{1+|C_x|}\rceil)^2) & o.w.
\end{cases}
\end{aligned}$$

Combining above, we have the upper bound on $\mathrm{LS}_{\hat{\mathbb{V}}[\hat{\tau}]}(D)$ as below:

$$\mathrm{LS}_{\hat{\mathbb{V}}[\hat{\tau}]}(D) \leq \frac{1}{2N^2}\left(\mathrm{LS}^+_g(D) + \max_{D'':d^+_{ar}(D,D'')\leq 1} \mathrm{LS}^-_g(D'')\right).$$

The β-smooth sensitivity is by definition:

$$S^*_{\hat{\mathbb{V}}[\hat{\tau}],\beta}(D) = \max_{k=0,\ldots,N} e^{-k\beta} \max_{D':d(D,D')\leq k} \mathrm{LS}_{\hat{\mathbb{V}}[\hat{\tau}]}(D').$$

Since the local sensitivity depends only on $|T_x|$ and $|C_x|$, it remains to consider all possible $|T'_x|$'s and $|C'_x|$'s such that $d(D,D') \leq k$ and compute the local sensitivity for all k and take the maximum.