A Simple and Practical Method for Reducing the Disparate Impact of Differential Privacy

Lucas Rosenblatt, Julia Stoyanovich, Christopher Musco

New York University
{lucas.rosenblatt, stoyanovich, cmusco}@nyu.edu

Abstract

Differentially private (DP) mechanisms have been deployed in a variety of high-impact social settings (perhaps most notably by the U.S. Census). Since all DP mechanisms involve adding noise to results of statistical queries, they are expected to impact our ability to accurately analyze and learn from data, in effect trading off privacy with utility. Alarmingly, the impact of DP on utility can vary significantly among different sub-populations. A simple way to reduce this disparity is with *stratification*. First compute an independent private estimate for each group in the data set (which may be the intersection of several protected classes), then, to compute estimates of global statistics, appropriately recombine these group estimates. Our main observation is that naive stratification often yields high-accuracy estimates of population-level statistics, without the need for additional privacy budget. We support this observation theoretically and empirically. Our theoretical results center on the private mean estimation problem, while our empirical results center on extensive experiments on private data synthesis to demonstrate the effectiveness of stratification on a variety of private mechanisms. Overall, we argue that this straightforward approach provides a strong baseline against which future work on reducing utility disparities of DP mechanisms should be compared.

Introduction

Two moral and legal imperatives, data privacy and algorithmic equity, have received significant recent research attention. For ensuring data privacy, differential privacy (DP) has emerged as a gold standard technique (Dwork, Roth et al. 2014). Academics and practitioners alike use DP algorithms to solve problems with sensitive data; notably, big tech companies like Google, Microsoft and Apple rely on differential privacy for protecting customer data (Erlingsson, Pihur, and Korolova 2014; Ding, Kulkarni, and Yekhanin 2017; Cormode et al. 2018). One recent high-profile use of DP was in the 2020 United States Census, which includes statistical disclosures with data for millions of Americans (Bureau 2021; Christ, Radway, and Bellovin 2022; Groshen and Goroff 2022; Hawes 2020). Additionally, a wide variety of popular algorithms have DP versions, from fundamental statistical methods (Dwork, Roth et al. 2014), to machine learning (Ji, Lipton, and Elkan 2014) and even deep learning (Abadi et al. 2016) techniques.

However, DP[1] can be in tension with algorithmic equity, where, unlike for privacy, there is no "gold standard" definition, and where potential harms are murky and plentiful (Corbett-Davies and Goel 2018; Mitchell et al. 2021). The 2020 Census, for example, was criticized over concerns that DP techniques would have adverse impacts on estimating demographic proportions (Ruggles et al. 2019) and on redistricting (Kenny et al. 2021). In fact, much attention has been paid lately to disparities in performance of DP mechanisms in a variety of settings, and to potential harms resulting from such disparities (Fioretto et al. 2022). Many of these works demonstrate disparities in performance for different demographic groups within the data (Bagdasaryan, Poursaeed, and Shmatikov 2019; Ganev, Oprisanu, and De Cristofaro 2022). In-line with recent literature, we broadly refer to such disparities as the *disparate impact* of DP.

As a concrete example, consider the problem of DP data synthesis. Suppose we have a data set X that can be split into k disjoint groups $G_1 \cup \ldots \cup G_k = X$, each of which might contain individuals in the intersection of several protected classes. A concern could be that a private data synthesis method run on X does not faithfully represent the data distribution for some of these groups, which can lead to technical bias (Friedman and Nissenbaum 1996) — in the sense of disparities in accuracy, appropriately measured — when machine learning (ML) models or statistics are fit to the synthetic data (Ganev, Oprisanu, and De Cristofaro 2022).

Contributions With the above example in mind, our paper analyzes a baseline approach to address the disparate impact of DP mechanisms: *stratification*. We note that general stratification-based methods are fundamental in statistics and may already be used by practitioners in privatizing data; however, we do not know of any work that *formally employs stratification to address the disparate impact of DP*. In particular, we could simply run a DP mechanism separately on $G_1, \ldots, G_k$ and report the results. If we have access to publicly available estimates of the size of each group, we could then take a weighted combination of the results to get a statistical estimate, or to generate DP synthetic data, for the global population. For example, we might fit data synthe-

[1] DP will mean "differential privacy" or "differentially private."

sizers $D_1, \ldots, D_k$ for each group. Then, to obtain synthetic data for the entirety of X, we would sample from synthesizer D_i with probability proportional to $|G_i| / \sum_{i=1}^{k} |G_i|$.

Intuitively, stratification minimizes disparate impact: for each group, we represent the data as well as we would have if the other groups did not exist.[2] So, the main question we ask is whether or not this simple approach can lead to high-quality estimates of *global* population statistics. Is something lost by treating individual groups separately? We make progress on answering this question, arguing that the cost of stratification is small or even negligible. Specifically, we make the following contributions:

(1) We validate the stratification strategy by first considering the problem of mean estimation. By making Dirichlet assumptions on the prior distribution of group sizes, we develop a theoretical understanding of the impact of stratification on the population mean estimate, and show that this impact is limited. Furthermore, for some state-of-the-art adaptive mean estimation techniques like COINPRESS (Biswas et al. 2020), stratification can even *reduce error in estimating the global population mean*, while also giving estimates for each stratum that are as accurate as one can hope for.

(2) Next, we validate our strategy on stratified DP data synthesizers, motivated by prior work highlighting the disparate impacts of these algorithms (Bagdasaryan, Poursaeed, and Shmatikov 2019; Ganev, Oprisanu, and De Cristofaro 2022). Our results indicate that stratification leads to *minimal overall utility loss for synthetic data* in practical privacy regimes, while also *reducing disparities* in utility across subgroups in the data.

Related Works

Disparate impact of DP Informally, a DP mechanism exhibits *disparate impact* when it leads to adverse outcomes for historically disadvantaged (i.e., protected) population groups, even if the mechanism appears neutral or unbiased on its face. Legally, a practice that adversely impacts protected groups can be considered discriminatory even without obvious categorization or intent to harm (Garrow 2014; Feldman et al. 2015; Barocas and Selbst 2016).

Prior work examining DP mechanisms found concerning disparate impact and other fairness trends. Some of this work has focused on bias introduced by private stochastic gradient descent (DP-SGD) (Abadi et al. 2016; Mironov 2017): Empirically, Bagdasaryan, Poursaeed, and Shmatikov (2019) discovered that DP-SGD amplifies noise in the data and adversely impacts certain subgroups, while theoretically, Tran, Dinh, and Fioretto (2021) showed that multiplicative effects on the Hessian loss in DP-SGD affect the proximity of group-specific data to the decision boundary. Others have also investigated DP-SGD's effects on image generation tasks (Cheng et al. 2021) and proposed adaptive clipping mechanisms to reduce negative subgroup impacts (Xu, Du, and Wu 2021), or suggested that different DP mechanisms for deep learning have reduced disparate impact (Uniyal et al. 2021). Other works assessing the fairness impacts of DP methods have primarily focused on tabular data and machine learning. Several methods have been proposed to balance the trade-off between privacy and fairness in classification, both theoretically and empirically (Cummings et al. 2019; Pujol et al. 2020; Xu, Yuan, and Wu 2019). Of particular relevance to our work, Ganev, Oprisanu, and De Cristofaro (2022) demonstrated the "Matthew Effects" (i.e., better performance for majority groups) of DP tabular synthetic data, which is the main focus of our experiments.

Private mean estimation The first part of our paper deals with finding a private empirical mean of a distribution, which is necessary because empirical means have been shown to reveal personally identifiable or otherwise sensitive information (Dinur and Nissim 2003; Dwork et al. 2017, 2015). Foundational work by Karwa and Vadhan (2017) studied algorithms for privately calculating statistical properties of finite-sample Gaussian distributions in various settings. Follow-up work introduced practical methods for incorporating distributional assumptions for multivariate settings (Biswas et al. 2020; Kamath, Singhal, and Ullman 2020) or for long-tailed distributions (Kamath et al. 2022). Complementary work discussed robustness guarantees to data ablations (Liu et al. 2021b), and eschewing distributional assumptions on Gaussians (Ashtiani and Liaw 2022). Of greatest relevance to our work is that of Biswas et al. (2020), who present an adaptive mean estimation algorithm, which we discuss in detail later. Additionally, contemporaneous work by Lin et al. (2023) studied DP stratification for confidence intervals, but they do not consider distributional assumptions on strata group sizes as we do.

Private synthetic data A substantial amount of work on DP data synthesis has been conducted in recent years (Aydore et al. 2021; Boedihardjo, Strohmer, and Vershynin 2022; Cai et al. 2021; McKenna, Sheldon, and Miklau 2019; Rosenblatt et al. 2020; Vietri et al. 2020; Zhang et al. 2021), with the best-performing methods following the "*Select, Measure, Project*" paradigm (discussed further in our results section) (Tao et al. 2021; Rosenblatt et al. 2022). We present our synthetic data results for three state-of-the-art algorithms: MST (McKenna, Miklau, and Sheldon 2021), GEM (Liu, Vietri, and Wu 2021) and AIM (McKenna et al. 2022), each of which offers a different flavor of this paradigm.

Preliminaries and Problem Statement

Notation We notate a standard Gaussian normal distribution as $\mathcal{N}(\mu, \sigma^2)$, using μ for mean and σ^2 for variance. A common distribution in data privacy is the centered *Laplace* distribution, notated $Lap(b)$, with PDF $f(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$. We use X to denote the data set in consideration, with $x_1, ..., x_n \in X$ items corresponding to individu-

[2] Some DP synthesizers based on the "*Select, Measure, Project*" approach, like MWEM (Hardt, Ligett, and McSherry 2012), use stratification in an *implicit way* because they take measurements of data based on marginal queries. However, these algorithms often rely on mutual information or other metrics to select measurements that are most informative about the joint distribution; in doing so, they may not measure marginals with respect to *all* attributes in a data set. This can leave groups vulnerable to disparate impact, and in fact, seems the likely culprit in the observed disparate impacts by (Ganev, Oprisanu, and De Cristofaro 2022) and others.

als. Each x_i is a single value or a vector of multiple values. We let $G_1 \cup \ldots \cup G_k = X$ denote disjoint subsets of X.

Differential privacy basics Differential privacy (DP) guarantees that the result of a data analysis or a query remains virtually unchanged even when one record in the dataset is modified or removed, thus preventing any deductions about the inclusion or exclusion of any specific individual. Modifying or removing a record from dataset X induces a *neighboring* dataset X'. We usually fix our definition of neighboring datasets, and define DP accordingly. For the purposes of our paper, X and X' are neighboring if one can be obtained by removing a single item x_i from the other.

The definitions for classical DP and zero-concentrated DP (ρ-zCDP) (a common alternative definition relevant to our work) are given in Definition 1 and Definition 2 respectively.

Definition 1 ((ϵ, δ)-Differential Privacy)**.** *A randomized mechanism $\mathcal{M}$ provides (ϵ, δ)-differential privacy if, for all pairs of neighboring datasets X and X', and all subsets R of possible outputs:*

$$\Pr[\mathcal{M}(X) \in R] \leq e^{\epsilon} \Pr[\mathcal{M}(X') \in R] + \delta$$

Definition 2 (Concentrated Differential Privacy (ρ-zCDP) (Bun and Steinke 2016))**.** *Here, $D_\alpha\left(\mathcal{M}(X)||\mathcal{M}(X')\right)$ denotes α-Rényi divergence. Then, a randomized algorithm $\mathcal{M}$ satisfies ρ-*zCDP *if for all pairs of neighboring datasets X and X',*

$$D_\alpha\left(\mathcal{M}(X)||\mathcal{M}(X')\right) \leq \rho\alpha, \forall \alpha \in (1, \infty)$$

These two closely related definitions scale with different relative privacy parameters. As Bun and Steinke (2016) showed, an ordering over the guarantees is as follows: An $(\epsilon, 0)$-DP mechanism gives $\frac{\epsilon^2}{2}$-zCDP, which gives $(\epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP for every $\delta > 0$.

Definition 3 (Sensitivity)**.** *Let $f : X \to \mathbb{R}$ be a real-valued function. The sensitivity Δf of f is defined as: $\Delta f = \max |f(X) - f(X')|$, where the maximum is taken over all pairs of possibly neighboring data sets X, X'.*

Definition 4 (Laplace Mechanism)**.** *Given a real-valued function f, the Laplace mechanism provides "pure" $(\epsilon, 0)$-differential privacy: $\mathcal{M}_{Lap}(f, X, \epsilon) = f(X) + Lap\left(\frac{\Delta f}{\epsilon}\right)$.*

To understand the impact of adding Laplace noise on the utility of a DP estimate, we will also require the following standard tail bound for our theoretical analysis:

Definition 5 (Laplace Tail Bound)**.** *Let Y be a random variable draw from a centered Laplace distribution with parameter b. Then $\Pr[|Y| \geq \alpha b] \leq e^{-\alpha}$.*

Definition 6 (Composition rules (Bun and Steinke 2016; Dwork, Roth et al. 2014))**.** *(ϵ, δ)-DP composes gracefully. For "sequential composition," if two randomized algorithms $\mathcal{M}$ and $\mathcal{M}'$ satisfy (ϵ_1, δ_1)-DP and (ϵ_2, δ_2)-DP, respectively, then the sequential $M^*(X) = (\mathcal{M}(X), \mathcal{M}'(X))$ satisfies $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP. For "parallel composition," if $\mathcal{M}$ satisfies (ϵ, δ)-DP, and $\{X_1, ..., X_k\}$ are disjoint, then the parallel mechanism $\mathcal{M}^*(X) = (\mathcal{M}(X_1), ..., \mathcal{M}(X_k))$ also satisfies (ϵ, δ)-DP. Note that ρ-zCDP composes analogously.*

Problem statement We consider a setting where the dataset X can be divided into groups of individuals. Specifically, we assume k *disjoint* subsets of X: $G_1, \ldots, G_k$, each of size $\{g_1, g_2, ..., g_k\} = \mathbf{g}$, that partition X s.t. $G_1 \cup G_2 \cup ... \cup G_k = X$. Groups are typically defined by *sensitive attributes*. For example, suppose we have two sensitive attributes, race with N_r possible values, and gender identity with N_g values. Then we would create a group G_i for each of the $N_r \cdot N_g$ possible combinations of race and gender identity. Our goal will be to release DP statistics about each group G_i so as to prioritize the highest possible accuracy for each group while also achieving acceptable accuracy for the full population when those statistics are aggregated.

Access to public weights In this paper, we assume limited access to public data: namely, available estimates of group sizes. This data is often already available. For example, in the case of intersectional groups (e.g., between race, gender identity, and income), the proportion of these groups in many populations is known (e.g., from Census data) across data contexts. Studying the implications of public data access for privacy is common, and assuming information about group sizes is quite mild in comparison to assumptions made in most prior work in this area (Bie, Kamath, and Singhal 2022; Ji and Elkan 2013; Liu et al. 2021a). Without accurate estimates of group sizes, we expect that the performance of stratification methods in approximating global statistics would degrade, although this topic is beyond the scope of our paper.

Stratified Private Mean Estimation

We begin by studying the effect of stratification on the problem of DP mean estimation for (single-variate) Gaussian data. We consider the standard simplified setting where the data variance is fixed and known, so data can be scaled to have variance 1. I.e., $X = \{x_1, x_2, ..., x_n\}$ consists of n i.i.d. draws from $\mathcal{N}(\mu, 1)$ where $x_i \in \mathbb{R}$. Additionally, we assume a known upper bound R on the absolute value of the mean, μ. In this setting, the standard DP estimator combines the Laplace mechanism with a data clipping step (Dwork and Lei 2009; Karwa and Vadhan 2017). For a scalar value x and a chosen constant $\gamma \in (0, 1)$, we let:

$$\text{clip}(x) = \begin{cases} x & \text{if } x \in [R - \sqrt{\log n/\gamma}, R + \sqrt{\log n/\gamma}] \\ \text{sign}(x) \cdot (R + \sqrt{\log n/\gamma}) & \text{otherwise.} \end{cases}$$

We then return the empirical mean with clipping and Laplace noise as a DP estimate $\hat{p}$. I.e.,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \text{clip}(x_i) + Lap\left(\frac{2R + 2\sqrt{\log n/\gamma}}{n\epsilon}\right).$$

It can be checked that $\hat{p}$ is $(\epsilon, 0)$-differentially private. To bound estimate error, we can use a triangle inequality involving the (non-private) empirical mean $p = \frac{1}{n}\sum_{i=1}^{n} x_i$:

$$|\hat{p} - \mu| \leq |p - \mu| + |\hat{p} - p|.$$

By standard Gaussian concentration, we have that the first term, which represents inherent statistical error, is bounded

by $O(\frac{1}{\sqrt{n}})$ with high probability. We'd like to bound the second term, which represents additional error incurred by privatization. Applying Definition 5, this second term is bounded with probability $1 - O(\gamma)$ by:

$$|\hat{p} - p| \leq O\left(\log(1/\gamma) \cdot \left(\frac{R + \sqrt{\log n/\gamma}}{n\epsilon}\right)\right). \quad (1)$$

In the setting where X can be split into disjoint groups, $G_1, \ldots, G_k$, we assume that each group i contains normally distributed data with mean μ_i and standard deviation 1. I.e., X follows a mixture of unit-variance Gaussian distributions.[3] Since the global mean can easily wash out information about individual groups, the natural stratification approach to minimizing disparate impact would be to compute DP estimates $\hat{p}_1, \ldots, \hat{p}_k$ for each $\mu_1, \mu_2, ...\mu_k$. I.e.,

$$\hat{p}_i = \frac{1}{|G_i|} \sum_{x \in G_i} \text{clip}(x) + Lap\left(\frac{2R + 2\sqrt{\log |G_i|/\gamma}}{|G_i|\epsilon}\right). \quad (2)$$

Since each of these means is based on a disjoint subset of X, by parallel composition (Definition 6), we can compute group-specific means with privacy parameters $(\epsilon, 0)$, and still obtain an overall $(\epsilon, 0)$-private method.

Given individual group estimates, how do we then compute an estimate of the global mean, μ? One natural approach is to do so from scratch, using a different DP estimator. Doing so, however, has a few drawbacks: (1) Since X is not disjoint from $G_1, \ldots, G_k$, we would only obtain a $(2\epsilon, 0)$-private method via serial composition if we report all individual means as well as the group mean. That is, for the same level of privacy, our error will be greater by a factor of two. (2) If we separately estimate the global mean, then this estimate may be "inconsistent," in that it could differ from a weighted average of the per-group means. An alternative is to simply return the private estimate:

$$\hat{p}_{strat} = \frac{1}{n} \sum_{i=1}^{k} |G_i| \cdot \hat{p}_i. \quad (3)$$

Proposition 7 (Privacy of $\hat{p}_{strat}$)**.** *The stratified mean with the Laplace mechanism is $(\epsilon, 0)$-DP by parallel composition rules (see Definition 6).*

But how accurate will the estimate be in comparison to a "fresh" DP estimate of the global mean? Again, letting p equal the empirical mean $\frac{1}{n}\sum_{i=1}^{n} x_i$, we find the following (complete proof deferred to supplementary materials):

Proposition 8 (Worst Case Bound for Stratified Mean Estimation)**.** *Let $\hat{p}_{strat}$ be the estimator defined in Eq.* (8)*. With probability $1 - O(\gamma)$,*

$$|p - \hat{p}_{strat}| \leq O\left(\log(1/\gamma) \cdot \sqrt{k}\left(\frac{R + \sqrt{\log(n/k\gamma)}}{n\epsilon}\right)\right) \quad (4)$$

When k is small (e.g., $k = n^c$ for constant $c < 1$), then $O(\log(n/k\gamma) = O(\log(n/\gamma))$. So, the above bound appears *worse* than what we obtain from the standard DP estimate $\hat{p}$ in Eq. (1) by a multiplicative factor of $\sqrt{k}$. Nevertheless, when we compute $\hat{p}_{strat}$ in practice, **we find its accuracy is competitive with $\hat{p}$.** It is natural to ask why this is the case.

[3]DP methods for estimating mixtures of Gaussians were studied extensively by (Kamath et al. 2019), differing in that the identities of the sub-populations are *unknown*.

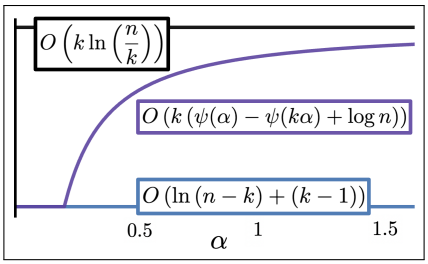

Figure 1: Visualizing the affect of the α parameter on this expectation term, with fixed k and n. α essentially controls the "sparsity" of the distribution; a small α implies expected error closer to the theoretical lower bound.

Distributional assumptions To better understand this question, first note that the error term involving R is typically a lower-order term, since an accurate estimate for R can be found using adaptive DP mean estimation techniques (Biswas et al. 2020). Removing all shared multiplicative terms and assuming γ is a small constant, we then see that the difference between the error of a fresh DP estimate, as in Eq. (1), and $\hat{p}_{strat}$, as in Eq. (9) is a matter of $\log(n)$ vs. $k\log(n/k)$. In the proof of Proposition 8, the $k\log(n/k)$ term arises from the following sum involving the group sizes: $\sum_{i=1}^{k} \log(|G_i|)$.

This sum is maximized when all sizes are equal, so $|G_i| = n/k$. However, when the sum is smaller, we obtain a tighter bound than Proposition 8. In the most extreme case, when all group sizes equal one, except for a single (majority) group, the bound can be as good as $(k-1) + \log(n-k)$, which nearly matches the $\log n$ dependence of $\hat{p}$. And, in fact, we *rarely* observe uniform group sizes in practice, particularly when considering intersectional groups. Often, a small number of majority groups dominate. To better explain the error, take a standard Bayesian assumption that our group size vector is drawn from a *Dirichlet* distribution, $\mathcal{D}(\alpha, k)$. (Our supplementary materials contains notes and details on Dirichlet parameters and behavior for completeness.) In this case, we prove the following bound (see supplementary materials):

Theorem 9 (Error Upper Bound with Dirichlet Assumption)**.** *Consider the Dirichlet distribution $\mathcal{D}(\alpha, k)$ with parameters k, α and let $\mathbf{g} \sim \mathcal{D}(\alpha, k)$ be a vector drawn from this distribution. If for each group i, $|G_i| = g_i/n$, then, letting ψ denote the digamma function, we have:*

$$\mathbb{E}\left[\sum_{i=1}^{k} \log |G_i|\right] \leq k\left(\psi(\alpha) - \psi(k\alpha) + \log n\right) \quad (5)$$

To better understand the bound of Theorem 9, please refer to Figure 1, which plots the bound in comparison to the weaker upper bound of $k\log(n/k)$ from Propostion 8, and in comparison to our informal lower bound $O(k + \log(n))$. As we can see, as the parameter α of the Dirichlet-distributed group size vectors varies from $[0.2, 1.0)$, we interpolate between the lower and upper bound. Small $\alpha < 1.0$ implies "sparsity", i.e. one or two dominant entries in the vector $\mathbf{g}$. Overall, we frame this result as follows: In cases when minority groups in the data are relatively small compared to the

majority groups, additional noise from privacy is expected to be small when aggregated.

Practical Implications of Theory Our theory for stratified private mean estimation (PME) bounds worst case costs (Proposition 8) and expected costs (Theorem 9) for stratification. Much can be learned from these bounds practically. For a salient example, consider a sample of American Community Survey data for New York from 2018, where $n = 196967$ and the RACE variable has $k = 9$. The group size vector here is $v_g = [0.70, 0.12, 0.086, 0.056, 0.029, ...]$. Using an iterative Monte Carlo approach to find a likely Dirichlet prior over v_g, we find $\alpha \approx 0.13$. Applying Proposition 8 and Theorem 9 then tells us that, for any setting of R or ϵ, our *expected* absolute error from stratification in estimating the mean for a given variable is approximately $\sim 80\%$ better than the worst case error, relative to the best case stratification error. This at least partially explains the strong performance of stratification empirically, seen in Figure 2. Put another way, our theory shows that, when the group sizes follow a Dirichlet distribution, then, as $\alpha \to 0$, the error of stratified PME scales with $O(k + \log(n) + R)$. This is close to the error of non-stratified PME, which scales as $O(\log(n) + R)$.

Adaptive Private Mean Estimation

In the previous section, we focused on bounding the additional error introduced by differential privacy, i.e., on $|\hat{p} - p|$, where $\hat{p}$ is a DP mean estimate and p is the empirical mean. However, as we highlighted, this error is always in addition to the inherent statistic error $|\mu - p|$ between the true population mean μ and our empirical estimate p. It is well known that stratification can help reduce this statistic error (Cochran 1977; Botev and Ridder 2017). This helps explain the value of stratification in practice: extra error introduced by DP noise can be offset by reduced statistical error.

In particular, consider the case when each group has fixed variance 1, but the means μ_i for the groups can differ substantially i.e. the overall data variance σ is much larger than 1. Then we expect statistical error on the order of $O(\sigma/\sqrt{n})$. On the other hand, if we assume we know exact group sizes and stratify, statistical error should scale as $O(1/\sqrt{n})$, which can be substantially better (Botev and Ridder 2017).

The naive mean estimation method analyzed in the previous section will likely not benefit from this scenario, since we will need to choose a large range R (thus scaling our noise) if our group means differ substantially. However, adaptive mean estimation methods like CoinPress (Biswas et al. 2020) have been introduced that only depend logarithmically on R. In our experiments, we observe that the aggregate population level means of these methods can actually *benefit* from stratification.

Measuring disparate impact of DP We consider the following formalism for our setting to measure the impact of private mechanisms on subgroups in data. Intuitively, we want the error for any protected group in a population to be comparable to the error for other groups. Consider dataset X, which consists of k strata $G_1 \cup ... \cup G_k$. Consider a non-private function that maps X to a real-valued vector with values for each of the k strata as well as a single global value $f(X) : X \to \mathbb{R}^{k+1}$. A privatized version of $f(x)$ is denoted $\mathcal{M}_f(X) : X \to \mathbb{R}^{k+1}$. Results for both on stratum i are denoted $f(G_i)$ and $\mathcal{M}_f(G_i)$ respectively. Population level results are then $f(X)_{k+1}$ and $\mathcal{M}_f(X)_{k+1}$.

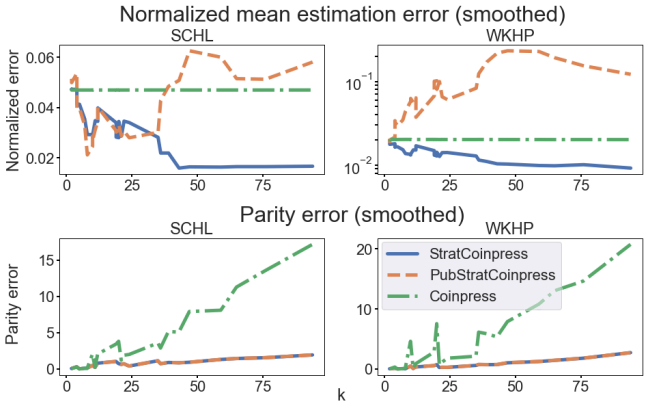

Figure 2: Comparisons of CoinPress and StratCoinPress on U.S. Census data from Folktables; SCHL is years spent in school (ordinal, 0-24), WKHP is hours worked per week (ordinal, 0-168). Top row shows normalized mean estimation error, bottom row shows parity error, both as the number of groups k increases.

Definition 10 (Parity error)**.** *Parity error ϕ is the average normalized absolute error in approximating $f(G_i)$ for each stratum i, plus the normalized absolute error in approximating $f(X)_{k+1}$, weighted by a positive parameter ω that determines the emphasis on population-level accuracy:*

$$\omega \left| \frac{f(X)_{k+1} - \mathcal{M}_f(X)_{k+1}}{f(X)_{k+1}} \right| + \sum_{i=1}^{k} \left| \frac{f(G_i) - \mathcal{M}_f(G_i)}{f(G_i)} \right| \quad (6)$$

In Definition 10, ω weights the faithfulness to the overall population estimate in the sum. We emphasize faithfulness for the protected class strata, and so we set $\omega = \frac{1}{k}$ (weighting population-level estimates the same as any single-stratum estimate).

Data and Setup We provide results of two experiments to validate the stratified adaptive private mean estimation approach. The first experiment (Figures 3, 7, 6) uses a synthetic mixture of k Gaussians, over a dataset of size n, with Dirichlet parameter α, to compare CoinPress and StratCoinPress. (The supplementary materials includes parameters for our synthetic Gaussians, which are chosen illustratively and are independent of our theory). Our second experiment (Figure 2) is on demographic census data for New York State from *Folktables*, a standard dataset in fair-ML literature (Ding et al. 2021). For consistency and comparability with prior work, we maintain Census variable codes in all of our plots and results; for example, SCHL is a discrete variable denoting years of education, and ranges from 0-24. For a complete list of Census variables, counts, marginals, and their corresponding meanings and domains, see the supplementary materials. We also use common, legally protected classes, like SEX, AGE and RAC1P, to create groups

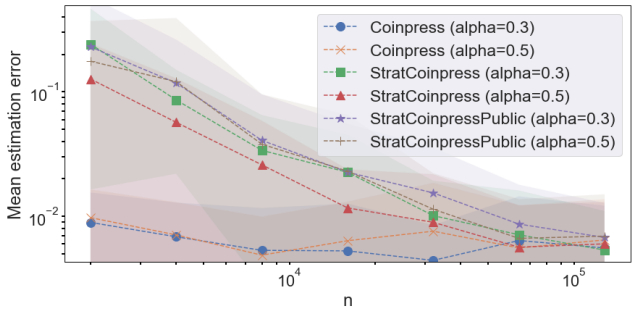

Figure 3: Mean estimation error, varying data size n in synth Gaussian mixture (50 runs). Stratified variants converge to performance of non-stratified COINPRESS as n grows (for $n \geq 10000$, error $\leq 1\%$). In other words, with large samples, we don't have to "pay" in *global* estimation error for a reduction in *group-specific* error.

for stratification. Note that we present two stratified variants: STRATCOINPRESS assumes direct access to the demographic weights necessary for stratification, while PUB-STRATCOINPRESS calculates those weights on a smaller, non-overlapping public holdout set.

In Figure 3, we depict the convergence of STRATCOINPRESS to the performance of COINPRESS as n grows linearly. We do not report on disparate impact experiments with the synthetic data, as we can arbitrarily *increase* harm (measured by parity error) by increasing variance of the mixed Gaussian means. Figures 6 and 7 on this experiment, which show the effects of varying k and α, are available in the supplementary materials.

Effectiveness In Figure 2, we show the effectiveness of a stratified adaptive mean estimation in practice. We vary k by creating a different number of intersectional groups by combining $sex \in [0, 1]$, $race \in [0, 8]$, $age \in [0, 4]$ (bucketed), and $physical\ disability \in [0, 1]$. The top row of Figure 2 shows that the error of the stratified variants of COINPRESS is controlled in settings where the range of possible values is small relative to the total population size. In fact, when the subgroups are particularly meaningful (such as with `SCHL`, `WKHP` and `JWMNP`), stratification can sometimes *improve* on the overall mean estimates. Challenging settings for stratification are large continuous spaces such as with `PINCP`, where the strata-specific estimates (and, thus, the accuracy of aggregation) suffer from well-known private mean estimation challenges for wide range long-tailed distributions.

The bottom row of Figure 2 tells the story of harm reduction on subgroups in the data. Because private estimates are done over disjoint strata under composition (Def. 6), parity error can be tightly controlled by STRATCOINPRESS. On the other hand, non-stratified COINPRESS incurs substantial utility loss for most intersectional groups, because their mean differs significantly from the population-level mean.

Stratified Data Synthesis

The disparate impacts of DP have been most commonly framed as occurring through private synthetic data (Bagdasaryan, Poursaeed, and Shmatikov 2019; Ganev, Oprisanu, and De Cristofaro 2022). Luckily, the principles of stratification (subset, estimate and aggregate) are not limited to mean estimation. Stratifying private synthesizers follows the same straightforward process: We learn separate parametric private distributions for each group-specific stratum. Then, to compose population-level data, we sample elements from strata-specific models *proportionally to their representation in the full population*. We do not offer theoretical bounds for private synthesizers as we did for mean estimation, as their guarantees are limited to specific query workloads selected during model fitting. This makes theoretical analysis challenging, and we leave it to future work.

Synthesizers and Privacy Settings We conducted an extensive empirical evaluation on the viability of stratified synthetic data. We selected three state-of-the-art DP synthesizers: MST (McKenna, Miklau, and Sheldon 2021), GEM (Liu, Vietri, and Wu 2021), and AIM (McKenna et al. 2022). MST is data-aware, GEM is both data- and workload-aware, and AIM is data-, workload-, and privacy budget-aware (McKenna et al. 2022). Benchmarking DP synthesizers is non-trivial and computationally expensive (Rosenblatt et al. 2022). We ran our experiments on an NVIDIA T4 GPU cluster and on a high-performance CPU cluster. We ran our models on the same privacy settings as (McKenna et al. 2022), $\epsilon \in \{0.01, 0.05, 0.1, 0.5, 1.0, 5.0\}$, representing a low to medium privacy budget regime. We trained 5 differently seeded models for each synthesizer at each privacy setting. Our models took over 200 hours of compute time to fit. We employ two naming conventions in labeling our figures: the first is that `_vanilla` denotes the non-stratified version of the algorithm, and the second is that `_VARIABLE` means that the algorithm stratifies explicitly along a set of dimensions (for example, `SEX_RAC1P` implies separate models maintained for all subgroups induced by sex and race.)

Classification Setting First, we found that stratification only seemed to harm the overall utility of private synthetic data in low privacy regimes ($\epsilon < 0.1$). We demonstrate this on the Folktables employment prediction task (Ding et al. 2021) by training a classifier on DP synthetic versions of the data. Figure 4 shows that the strength of these classifiers increased as ϵ increased, implying that relationships in the data are maintained for nearly all variants, although we acknowledge the limitations of using classification as a proxy task higher-dimensional fidelity.

Parity Error Reduction Second, we find that stratified variants of all synthesizers reduce parity error (Definition 10). Figure 5 shows how additional privacy budget can generally improve the performance of the stratified variants (where the parity error function f is the aggregate normalized difference of means across all variables in the Folktables employment task data). The best-performing synthesizer in our tests (GEM) also stratifies most gracefully, and provides the best parity error in nearly all ϵ parameter settings. We defer some experimental results to supplementary materials, which further demonstrate that, for all synthesizers, as privacy budget increases: (1) the maximum disparity

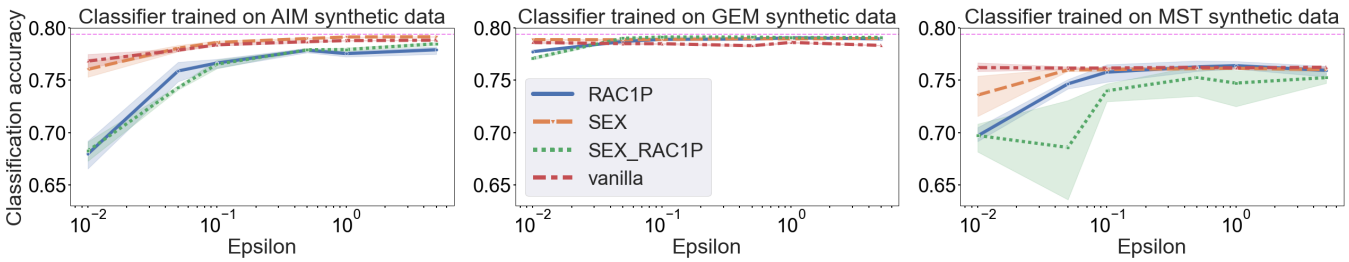

Figure 4: Performance of different stratified synthesizers, Folktables employment prediction. Note that variants of GEM and AIM approach a classifier trained on real data (accuracy ≈ 0.8) as privacy budget increases. From the perspective of maintaining predictive utility, stratification appears to incur only minor costs, if any (for $\epsilon \geq 1$, difference $\leq 2\%$).

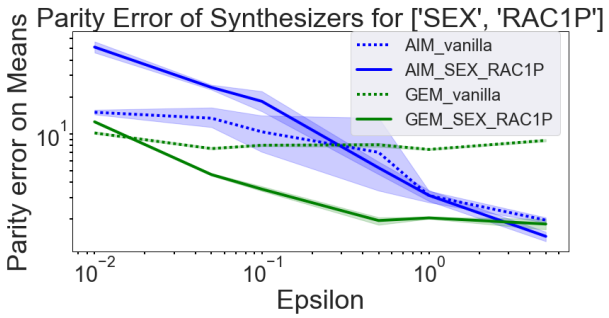

Figure 5: Parity error, where f aggregates the error of means calculated on all variables in the Folktables employment prediction dataset. Here we contrast the two higher performing synthesizers: AIM (adaptive, lower predictive performance) and GEM (non-adaptive, higher predictive performance). Vanilla GEM struggles without explicit stratification, but by stratifying we can reduce parity error significantly (a reduction of $\geq 200\%$). Vanilla AIM's budget and workload adaptivity has parity error benefits at low privacy budgets, though explicit stratification improves at higher budgets.

of all *means* gradually decreases (Figure 7a); (2) the *demographic parity* (Hardt, Price, and Srebro 2016) on the Folktables employment task increases towards 1 (Figure 7b); and (3) the maxmimum false negative rate difference between groups decreases (Figure 7c).

Finally, in Figure 9 (also deferred to supplementary materials), we use the ALL 3-WAY MARGINAL workload error from (McKenna et al. 2022) as our function for parity error. This is a standard method of measuring performance of DP synthesizers: a k-way marginal for attribute set S (where $|S| = k$) is a histogram over $x_S \in X$; marginal workload error is the average difference between these marginals run on real data and on private synthetic data (see supplementary materials or (McKenna et al. 2022) for more details). Figure 9 demonstrates that, in the case of non-budget adaptive algorithms like GEM, one can stratify and achieve low parity error and great overall performance. In the case of the budget-adaptive algorithm AIM, we see that it eventually has ample budget to adapt to all groups, and that the stratification might hurt performance. However, *we rarely know what an acceptable budget setting is a priori*; insufficient privacy budget for AIM (here, $\epsilon < 1.0$) greatly increases the parity error of the `vanilla` adaptive algorithm relative to its stratified variant. Performance of the the stratified variant improves stably for all subgroups, and is thus much safer to use in regimes with a limited privacy budget, or when it is unclear how to set the budget.

Overall, we found that the stratified GEM variants convincingly outperformed MST and AIM with small privacy budget ($\epsilon \leq 1.0$). AIM did outperform GEM in the highest privacy regime of $\epsilon = 5.0$, likely due to AIM's ability to adapt and utilize "excess" privacy budget.

Limitations and future work Our experimental results suggest that stratified synthetic data is safe given sufficient privacy budget, and that it often helps improve utility of subgroup data while preserving population-level utility, even in the case of adaptive algorithms. However, this fundamentally relies on good aggregation proportions, which may not always be available (say, in a medical context for a *specific* hospital). Though we believe that stratification provides a strong baseline for future work on adaptive DP algorithms with disparate impact protections, a stronger theoretical underpinning for stratified private synthesizers would allow for greater confidence in deployment. Additionally, our approach essentially targets parity error, and perhaps unsurprisingly trades off good performance there with worse performance by other metrics. Future work could formally characterize this trade-off by identifying the Pareto frontier of this dual optimization problem.

Conclusion

Reducing disparate impact in DP data release is a laudable goal that has received much attention in recent years. We showed that, when access to public estimates of group proportions in the data can be assumed, a *stratified* approach to disparate impact reduction is surprisingly effective, and that it does not significantly reduce — and sometimes even improves — the accuracy of population-level private statistics when the stratified private data is aggregated. With this work, we hope to encourage interest in principled methodologies for harm reduction when privatizing social data. We also hope that that practitioners will find the simple strategy we outlined here immediately applicable for private data release on sensitive data with protected classes.

Acknowledgments

This research was supported in part by NSF Award Nos. 1916505, 2312930, 1922658, by NSF Award No. 2045590 and by the NSF Graduate Research Fellowship under Award No. DGE-2234660.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proc. of the 2016 ACM SIGSAC Conf. on CCS*, 308–318.

Ashtiani, H.; and Liaw, C. 2022. Private and polynomial time algorithms for learning gaussians and beyond. In *Conf on Learning Theory*, 1075–1076. PMLR.

Aydore, S.; Brown, W.; Kearns, M.; Kenthapadi, K.; Melis, L.; Roth, A.; and Siva, A. A. 2021. Differentially private query release through adaptive projection. In *International Conf on Machine Learning*, 457–467. PMLR.

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California law review*, 671–732.

Bie, A.; Kamath, G.; and Singhal, V. 2022. Private Estimation with Public Data. *arXiv:2208.07984*.

Biswas, S.; Dong, Y.; Kamath, G.; and Ullman, J. 2020. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, 33: 14475–14485.

Boedihardjo, M.; Strohmer, T.; and Vershynin, R. 2022. Private sampling: a noiseless approach for generating differentially private synthetic data. *SIAM Journal on Mathematics of Data Science*, 4(3): 1082–1115.

Botev, Z.; and Ridder, A. 2017. *Variance Reduction*, 1–6. John Wiley Sons, Ltd. ISBN 9781118445112.

Bun, M.; and Steinke, T. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography: 14th International Conf, TCC 2016-B, Beijing, China, October 31-November 3, 2016, Proceedings, Part I*, 635–658. Springer.

Bureau, U. 2021. Disclosure avoidance for the 2020 census: An introduction.

Cai, K.; Lei, X.; Wei, J.; and Xiao, X. 2021. Data synthesis via differentially private markov random fields. *Proc. of the VLDB Endowment*, 14(11): 2190–2202.

Cheng, V.; Suriyakumar, V. M.; Dullerud, N.; Joshi, S.; and Ghassemi, M. 2021. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *Proc. of the 2021 ACM Conf. on FAccT*, 149–160.

Christ, M.; Radway, S.; and Bellovin, S. M. 2022. Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the US Census. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1564–1564. IEEE Computer Society.

Cochran, W. G. 1977. *Sampling techniques*. John Wiley & Sons.

Corbett-Davies, S.; and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023*.

Cormode, G.; Jha, S.; Kulkarni, T.; Li, N.; Srivastava, D.; and Wang, T. 2018. Privacy at scale: Local differential privacy in practice. In *Proc. of the 2018 International Conf on Management of Data*, 1655–1658.

Cummings, R.; Gupta, V.; Kimpara, D.; and Morgenstern, J. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conf on User Modeling, Adaptation and Personalization*, 309–315.

Ding, B.; Kulkarni, J.; and Yekhanin, S. 2017. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34: 6478–6490.

Dinur, I.; and Nissim, K. 2003. Revealing information while preserving privacy. In *Proc. of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 202–210.

Duchi, J. 2021. Lecture Notes for Statistics 311/Electrical Engineering 377.

Dwork, C.; and Lei, J. 2009. Differential privacy and robust statistics. In *Proc. of the forty-first annual ACM symposium on Theory of computing*, 371–380.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4): 211–407.

Dwork, C.; Smith, A.; Steinke, T.; and Ullman, J. 2017. Exposed! a survey of attacks on private data. *Annu. Rev. Stat. Appl*, 4(1): 61–84.

Dwork, C.; Smith, A.; Steinke, T.; Ullman, J.; and Vadhan, S. 2015. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, 650–669. IEEE.

Erlingsson, Ú.; Pihur, V.; and Korolova, A. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proc. of the 2014 ACM SIGSAC*, 1054–1067.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proc. of the 21th ACM SIGKDD*, 259–268.

Fioretto, F.; Tran, C.; Van Hentenryck, P.; and Zhu, K. 2022. Differential privacy and fairness in decisions and learning tasks: A survey. *arXiv:2202.08187*.

Friedman, B.; and Nissenbaum, H. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.*, 14(3): 330–347.

Ganev, G.; Oprisanu, B.; and De Cristofaro, E. 2022. Robin Hood and Matthew Effects: Differential Privacy Has Disparate Impact on Synthetic Data. In *International Conf. on Machine Learning*, 6944–6959. PMLR.

Garrow, D. J. 2014. Toward a Definitive History of Griggs v. Duke Power Co. *Vand. L. Rev.*, 67: 197.

Groshen, E. L.; and Goroff, D. 2022. Disclosure avoidance and the 2020 Census: What do researchers need to know. *Harvard Data Science Review*.

Hardt, M.; Ligett, K.; and McSherry, F. 2012. A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*, 25.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hawes, M. B. 2020. Implementing differential privacy: Seven lessons from the 2020 United States Census. *Harvard Data Science Review*, 2(2).

Ji, Z.; and Elkan, C. 2013. Differential privacy based on importance weighting. *Machine learning*, 93: 163–183.

Ji, Z.; Lipton, Z. C.; and Elkan, C. 2014. Differential privacy and machine learning: a survey and review. *arXiv:1412.7584*.

Kamath, G.; Mouzakis, A.; Singhal, V.; Steinke, T.; and Ullman, J. 2022. A private and computationally-efficient estimator for unbounded gaussians. In *Conf on Learning Theory*, 544–572. PMLR.

Kamath, G.; Sheffet, O.; Singhal, V.; and Ullman, J. 2019. Differentially private algorithms for learning mixtures of separated gaussians. *Advances in Neural Information Processing Systems*, 32.

Kamath, G.; Singhal, V.; and Ullman, J. 2020. Private mean estimation of heavy-tailed distributions. In *Conf on Learning Theory*, 2204–2235. PMLR.

Karwa, V.; and Vadhan, S. 2017. Finite sample differentially private Confidence intervals. *arXiv preprint arXiv:1711.03908*.

Kenny, C. T.; Kuriwaki, S.; McCartan, C.; Rosenman, E. T.; Simko, T.; and Imai, K. 2021. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 US Census. *Science advances*, 7(41): eabk3283.

Lin, S.; Bun, M.; Gaboardi, M.; Kolaczyk, E. D.; and Smith, A. 2023. Differentially Private Confidence Intervals for Proportions under Stratified Random Sampling. *arXiv preprint arXiv:2301.08324*.

Liu, T.; Vietri, G.; Steinke, T.; Ullman, J.; and Wu, S. 2021a. Leveraging public data for practical private query release. In *International Conf. on Machine Learning*, 6968–6977. PMLR.

Liu, T.; Vietri, G.; and Wu, S. Z. 2021. Iterative methods for private synthetic data: Unifying framework and new methods. *Advances in Neural Information Processing Systems*, 34: 690–702.

Liu, X.; Kong, W.; Kakade, S.; and Oh, S. 2021b. Robust and differentially private mean estimation. *Advances in neural information processing systems*, 34: 3887–3901.

McKenna, R.; Miklau, G.; and Sheldon, D. 2021. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *arXiv:2108.04978*.

McKenna, R.; Mullins, B.; Sheldon, D.; and Miklau, G. 2022. Aim: An adaptive and iterative mechanism for differentially private synthetic data. *arXiv:2201.12677*.

McKenna, R.; Sheldon, D.; and Miklau, G. 2019. Graphical-model based estimation and inference for differential privacy. In *International Conf on Machine Learning*, 4435–4444. PMLR.

Mironov, I. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, 263–275. IEEE.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8: 141–163.

Pujol, D.; McKenna, R.; Kuppam, S.; Hay, M.; Machanavajjhala, A.; and Miklau, G. 2020. Fair decision making using privacy-protected data. In *Proceedings of FAccT 2020*, 189–199.

Rosenblatt, L.; Holovenko, A.; Rumezhak, T.; Stadnik, A.; Herman, B.; Stoyanovich, J.; and Howe, B. 2022. Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy. *arXiv:2208.12700*.

Rosenblatt, L.; Liu, X.; Pouyanfar, S.; de Leon, E.; Desai, A.; and Allen, J. 2020. Differentially private synthetic data: Applied evaluations and enhancements. *arXiv:2011.05537*.

Ruggles, S.; Fitch, C.; Magnuson, D.; and Schroeder, J. 2019. Differential privacy and census data: Implications for social and economic research. In *AEA papers and proceedings*, volume 109, 403–408. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Tao, Y.; McKenna, R.; Hay, M.; Machanavajjhala, A.; and Miklau, G. 2021. Benchmarking differentially private synthetic data generation algorithms. *arXiv:2112.09238*.

Tran, C.; Dinh, M.; and Fioretto, F. 2021. Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 34: 27555–27565.

Uniyal, A.; Naidu, R.; Kotti, S.; Singh, S.; Kenfack, P. J.; Mireshghallah, F.; and Trask, A. 2021. Dp-sgd vs pate: Which has less disparate impact on model accuracy? *arXiv:2106.12576*.

Vietri, G.; Tian, G.; Bun, M.; Steinke, T.; and Wu, S. 2020. New oracle-efficient algorithms for private synthetic data release. In *International Conf on Machine Learning*, 9765–9774. PMLR.

Wainwright, M. J. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Xu, D.; Du, W.; and Wu, X. 2021. Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In *Proc. of the 27th ACM SIGKDD Conf.*, 1924–1932.

Xu, D.; Yuan, S.; and Wu, X. 2019. Achieving differential privacy and fairness in logistic regression. In *Companion Proc. of the 2019 WWW Conf.*, 594–599.

Zhang, Z.; Wang, T.; Honorio, J.; Li, N.; Backes, M.; He, S.; Chen, J.; and Zhang, Y. 2021. Privsyn: Differentially private data synthesis.