Robin Hood and *Matthew* Effects: Differential Privacy Has Disparate Impact on Synthetic Data

Georgi Ganev [1 2] **Bristena Oprisanu** [1] **Emiliano De Cristofaro** [1]

Abstract

Generative models trained with Differential Privacy (DP) can be used to generate synthetic data while minimizing privacy risks. We analyze the impact of DP on these models vis-à-vis *underrepresented* classes/subgroups of data, specifically, studying: 1) the *size* of classes/subgroups in the synthetic data, and 2) the *accuracy* of classification tasks run on them. We also evaluate the effect of various levels of imbalance and privacy budgets. Our analysis uses three state-of-the-art DP models (PrivBayes, DP-WGAN, and PATE-GAN) and shows that DP yields opposite size distributions in the generated synthetic data. It affects the gap between the majority and minority classes/subgroups; in some cases by reducing it (a "Robin Hood" effect) and, in others, by increasing it (a "Matthew" effect). Either way, this leads to (similar) disparate impacts on the accuracy of classification tasks on the synthetic data, affecting disproportionately more the underrepresented subparts of the data. Consequently, when training models on synthetic data, one might incur the risk of treating different subpopulations unevenly, leading to unreliable or unfair conclusions.

1. Introduction

Releasing synthetic data is an increasingly advocated and adopted approach to reduce privacy risks while sharing data (Van Der Schaar & Maxfield, 2020). Synthetic data initiatives have been promoted by, e.g., the US Census Bureau (Benedetto et al., 2018), England's National Health Service (NHS England, 2021), and NIST (NIST, 2018a;b).

The idea is to train *generative* machine learning models to learn the probabilistic distribution of the (real) data and then sample from the model to generate new (synthetic) data records. However, real-world datasets often contain personal and sensitive information about individuals (Thompson & Warzel, 2019) that could leak into/through models that are trained on them. Generative models can overfit or memorize individual data points (Carlini et al., 2019; Webster et al., 2019), which facilitates privacy attacks such as membership or property inference attacks (Hayes et al., 2019; Chen et al., 2020a; Stadler et al., 2022).

The state-of-the-art method for training models that provably minimize inferences is to do while satisfying Differential Privacy (DP) (Dwork et al., 2014). DP provides a mathematical guarantee on the privacy of all records in the training dataset by bounding their individual contribution. This can be achieved by applying noise (e.g., using the Laplace Mechanism (Dwork et al., 2006b)), relying on techniques such as DP-Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016), or Private Aggregation of Teacher Ensembles (PATE) (Papernot et al., 2016; 2018).

Naturally, as they rely on perturbation, DP methods inherently reduce accuracy in the task the data is used for. Incidentally, this degradation is often disproportionate; for instance, the accuracy of *DP classifiers* often drops more for the underrepresented classes and subgroups of the dataset. Prior work (Bagdasaryan et al., 2019; Farrand et al., 2020; Uniyal et al., 2021) illustrates this effect when deep neural networks are trained with DP-SGD or PATE on imbalanced datasets. Moreover, *DP statistics* have also been shown to lead to disproportionate biases (Kuppam et al., 2019).

Problem Statement. So far, this "disparate effect" caused by DP and its applications have only been analyzed in the context of discriminative models. This paper focuses on DP generative models and tabular synthetic data. We look at the problem from two angles: 1) *counts comparisons* and 2) *downstream tasks* such as classification. We analyze three widely used DP generative models: PrivBayes (Zhang et al., 2017), DP-WGAN (Alzantot & Srivastava, 2019), and PATE-GAN (Jordon et al., 2018), which rely, respectively, on the Laplace Mechanism, DP-SGD, and PATE.

Our work aims to answer the following research questions:

– **RQ1:** Do DP generative models generate data in simi-

[1]University College London, London, UK [2]Hazy, London, UK. Correspondence to: Georgi Ganev <georgi.ganev.16@ucl.ac.uk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

lar classes and subgroups proportions to the real data?

- **RQ2:** Does training a classifier on DP synthetic data lead to the same disparate impact on accuracy as training a DP classifier on the real data?
- **RQ3:** Do different DP mechanisms for DP synthetic data behave similarly under different privacy and data imbalance levels?

Main Findings. Overall, our experiments show that:

1. There is a disparate effect on the classes and subgroups sizes in the synthetic data generated by all DP generative models. This effect is dependent on the generative model and DP mechanism; e.g., PrivBayes evens the data, while PATE-GAN increases the imbalance.
2. There also is a disparate effect on the accuracy of classifiers trained on synthetic data generated by all generative models; for instance, underrepresented classes and subgroups suffer bigger and/or more variable drops. Furthermore, majority classes with similar characteristics to minority classes could also suffer from a disproportionate drop in utility.
3. The magnitude of these effects on size and accuracy increases when stronger privacy guarantees are imposed. Higher data imbalance levels further intensify them. Also, some generative models are better suited for specific privacy budgets and imbalance levels.
4. While classifiers trained on data generated by PATE-GAN perform much better than, or on par with, DP-WGAN, we observe some undesirable behaviors: PATE-GAN completely fails to learn some subparts of the data with highly imbalanced multi-class data. With low privacy budgets, it also generates synthetic data with artificially enhanced correlation between the subgroup and the target columns.

2. Preliminaries

In this section, we present some background information, then, we introduce the datasets, the classifiers used for baselines, and the generative models used for producing synthetic data. (The source of the implementations we use, when applicable, is also reported.)

2.1. Background

Generative Models and Synthetic Data. During fitting, the generative model training algorithm $GM(D^n)$ takes in input D^n (a sample dataset consisting of n records drawn iid from the population $D^n \sim P(\mathbb{D})$), updates its internal parameters to learn $P_g(D^n)$, a (lower-dimensional) representation of the joint probability distribution of the sample dataset $P(D^n)$, and outputs a trained model $g(D^n)$. Then, one can sample from the trained model to generate a synthetic dataset of size m, $S^m \sim P(g(D^n))$. Both the fitting and generation steps are stochastic; in order to get confidence intervals, one can train the generative model l times and sample k synthetic datasets for each trained model.

While several different approaches exist to build generative models, in this paper, we focus on two of them, specifically: 1) Bayesian networks (Koller & Friedman, 2009; Barber, 2012), and 2) Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). The former is a graphical model that breaks down the joint distribution by explicit lower-dimensional conditional distributions. The latter approximates the dataset distribution implicitly by iteratively optimizing a min-max "game" between two neural networks: a generator, producing synthetic data, and a discriminator, trying to distinguish real from synthetic samples.

Differential Privacy (DP). Let ϵ be a positive and real number and $\mathcal{A}$ a randomized algorithm. $\mathcal{A}$ satisfies ϵ-DP if, for all neighboring datasets D_1 and D_2 (differing in a single data record), and all possible outputs S of $\mathcal{A}$, the following holds (Dwork et al., 2014):

$$P[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon) \cdot P[\mathcal{A}(D_2) \in S]$$

In other words, looking at the output of the algorithm, one cannot distinguish whether any individual's data was included in the input dataset or not. The level of that indistinguishability is measured by ϵ, also called a privacy budget.

In the context of machine learning, $\mathcal{A}$ is usually the training procedure. In this paper, we focus on three DP techniques: the Laplace mechanism (Dwork et al., 2006b), DP-SGD (Abadi et al., 2016), and PATE (Papernot et al., 2016; 2018) (for more details, see Sec. 2.3). The last two techniques use a relaxation of DP called (ϵ, δ)-DP (Dwork et al., 2014); here, δ, usually a small number, denotes a probability of failure. Finally, due to its robustness to post-processing, DP allows for DP-trained models to be re-used without further privacy leakage.

Disparate Impact Metrics. For the downstream task evaluation (see Sec. 3.1), we follow the disparate impact metrics proposed in (Bagdasaryan et al., 2019) and use *accuracy parity*, a weaker form of "equal odds" (Hardt et al., 2016). Specifically, we focus on model accuracy on imbalanced classes (and multi-classes) and imbalanced subgroups (with balanced classes) of the dataset. Similarly to (Bagdasaryan et al., 2019), we do not consider (other) fairness evaluations, leaving them as items for future work.

2.2. Datasets

We consider several tabular and one image datasets from different domains, which are widely used in the ML research community. All have an associated classification task or have slightly been modified for this purpose.

Adult. The Adult dataset (Dua & Graff, 2017) is extracted from the 1994 Census database, consisting of 32,561 training and 16,281 testing records. It has 15 attributes: 6 numerical, including age, and 9 categorical, including sex and race. The target column indicates whether the individual's income exceeds \$50K/year.

Texas. The Texas Hospital Inpatient Discharge dataset (DSHS, 2013) contains data on discharges from Texas hospitals. As done in previous work (Stadler et al., 2022), we sample 49,983 records from 2013 and select 12 attributes, 1 numerical and 11 categorical, including age, sex, and race. To create a classification task, we convert the numerical attribute, indicating the length of stay in the hospital, into a categorical one by specifying whether the person's hospitalization was a week or longer.

Purchases. The Purchases dataset is based on Kaggle's "Acquire Valued Shoppers Challenge" (Kaggle, 2013), aimed at predicting whether customers would become loyal to products based on incentives. As done in previous work (Shokri et al., 2017), we modify the main task to be predicting customers' purchase style. First, we filter products purchased at least 750,000 times and customers who made at least 500 purchases. Then, we summarize the data so that each row represents a customer with 108 binary features, corresponding to whether the customer has bought that product or not. Finally, to create the different purchasing styles, we cluster the customers into 20 clusters using a Mixture of Gaussian models. This yields a dataset with 152,369 customers and 109 attributes, including the style. Unlike the previous datasets, the classification task here is multi-class.

MNIST. The MNIST dataset (LeCun et al., 2010) consist of 60,000 training and 10,000 testing black and white handwritten 784-pixel digits. The goal is to classify the digit, making it a multi-class problem with 10 classes.

2.3. Generative Models

Our evaluation includes three of the most popular DP generative models: a statistical one based on Bayesian networks and two GANs incorporating DP mechanisms. Unless stated otherwise, we use the default hyperparameters, as provided by the authors.

PrivBayes. PrivBayes (Zhang et al., 2017; Ping et al., 2017) first constructs an optimal Bayesian network that approximates the joint data distribution by low-dimensional conditional distributions and then estimates them. Both of these steps are done with ϵ-DP guarantees, respectively, using the Exponential Mechanism (McSherry & Talwar, 2007) to choose the parents for each child node and the Laplace Mechanism (Dwork et al., 2006b) to construct noisy contingency tables before converting them to distributions. Looking at the step involving Laplace Mechanism in more detail, any negative noisy counts are clipped at 0 before being normalized to a distribution, potentially leading to a biased estimator. We discretize numerical columns to 50 bins, as opposed to 20, and set the degree of the network to 3 for all datasets except for Purchases, where it is 2. Furthermore, we identified an industry-wide bug in the open-source package violating the DP guarantees and fixed it.[1]

DP-WGAN. DP-WGAN (Alzantot & Srivastava, 2019) is one of the top 5 solutions to the 2018 NIST Contest (NIST, 2018a). It relies on the WGAN architecture (Arjovsky et al., 2017), which improves training stability and performance by using the Wasserstein distance instead of the Jensen-Shannon divergence as in GANs. Furthermore, (ϵ, δ)-DP of the output is achieved using DP-SGD (Abadi et al., 2016), which sanitizes the gradients (clips the ℓ_2 norm of the individual gradients and applies Gaussian Mechanism (Dwork et al., 2006a) to the sum) of the discriminator during training. The privacy budget is tracked using the moments accountant method (Abadi et al., 2016). To be consistent with PATE-GAN (see below), we set $\delta = 10^{-5}$ for all experiments.

PATE-GAN. PATE-GAN (Jordon et al., 2018) adapts the PATE framework (Papernot et al., 2016; 2018) for training GANs. Instead of a single discriminator, there are k teacher-discriminators and a student-discriminator. The teacher-discriminators only see a disjoint partition of the real data. They are trained to minimize the classification loss when classifying samples as real or fake. In contrast, the student-discriminator is trained using noisy labels (using the Laplace Mechanism) predicted by the teachers. As before, the privacy budget of the algorithm is calculated using the moments accountant (Abadi et al., 2016) and the output is (ϵ, δ)-DP, with $\delta = 10^{-5}$.

2.4. Discriminative Models

As mentioned, our experiments include a downstream task (classification) run on the synthetic data. We use Logistic Regression (LR) to avoid another layer of stochasticity; specifically two versions of LR: the standard one in Scikit-Learn (Pedregosa et al., 2011) and one with DP guarantees (Chaudhuri et al., 2011; Holohan et al., 2019). The latter achieves ϵ-DP by perturbing the objective function before optimization.

3. Experimental Evaluation

3.1. Evaluation Methodology

Broadly, our goal is to empirically measure the impact of generative models with different DP mechanisms, ϵ levels, and data imbalance ratios have on class/subgroups distributions in the generated synthetic data and downstream task

[1] https://github.com/DataResponsibly/DataSynthesizer/issues/34

performance. We consider four settings:

S1) *Binary class size, precision, and recall.* We focus on the effect on binary classes, reporting class recall and precision because the target columns in all datasets are imbalanced.

S2) *Multi-class size, precision, and recall.* We study the effect on multi-classes. As in the previous setting, we report class recall and precision.

S3) *Single-attribute subgroup size and accuracy.* We analyze the effect on a single-attribute subgroup. Here, we treat a single feature (e.g., sex) as a subgroup. We imbalance the dataset, so the minority subgroup comprises the desired ratio of the population while keeping the class per subgroup balanced.

S4) *Multi-attribute subgroup size and accuracy.* We focus on the effect on multi-attribute subgroups. We treat an intersection of features (e.g., age, sex, and race) as small fine-grained subgroups. As in the previous setting, we balance the data only according to a single-attribute subgroup; otherwise, we risk throwing too much data out. Thus, we discard subgroups with fewer than 25 members.

All evaluation settings follow three steps: dataset preparation, synthetic data generation, and prediction – see below.

Dataset Preparation. First, we split the dataset into training and testing if the latter is not explicitly provided. If the subgroup imbalance level is provided (S3 and S4), for both training and testing datasets, we balance the subgroup by class, so there are 50% of each class per subgroup (we only consider binary classes in these settings).

Then, we imbalance the datasets to the desired subgroup imbalance level (while maintaining class parity), where the level represents the ratio of minority subgroup to the total size of the dataset.

Synthetic Data Generation. For a given generative model and privacy budget ϵ, we train l (we set $l = 10$) generators and generate k (we set $k = 10$) synthetic datasets with size equal to the input dataset. This results in $l \cdot k$ synthetic datasets. We measure the class/subgroups distributions. If single/multi-attribute subgroup is provided (S3 and S4), we also measure correlation between the subgroup and target columns by calculating the mutual information (reported in Appendix A.5).

Classifiers Prediction. We capture the performance of three types of classifiers: 1) *real classifier* – we train a single LR on the real dataset and predict on the test dataset to serve as an overall baseline; 2) *DP classifiers* – we train $l \cdot k$ (equal to the number of synthetic datasets) DP LRs on the real dataset and predict on the test dataset; 3) *synth classifiers* – we train a single LR per synthetic dataset (in total $l \cdot k$) and predict on the test dataset.

3.2. S1: Binary Class Size and Precision

We consider privacy budgets (ϵ) of 0.01, 0.1, 1, 10, 100, and infinity ("no-DP") for the binary classification datasets (Adult and Texas). We do not imbalance the data because all datasets already have imbalanced classes – specifically, the proportion of the minority class to the total number of records is 0.24 in Adult and 0.195 in Texas.

Size. In the first row of Fig. 1 (and 7 in Appendix A.1), we plot the class size distribution in the synthetic data for Adult and Texas, respectively. For both datasets, for PrivBayes, we observe that decreasing ϵ results in synthetic data with reduced class imbalance; for PATE-GAN, the opposite is true – decreasing ϵ leads to increased class imbalance (except for $\epsilon = 0.01$ for Texas). These results are consistent with the disparate effects from applying Laplace to DP statistics (Kuppam et al., 2019) as well as PATE to DP neural networks classifiers (Uniyal et al., 2021). Interestingly, DP-WGAN preserves the imbalance for $\epsilon > 0.1$. As expected, there is an increased standard deviation for smaller values of ϵ for all synthetic datasets.

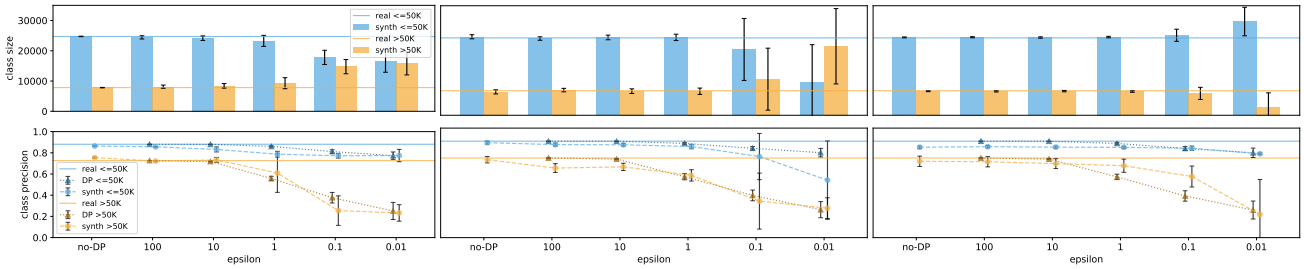
Precision. In the bottom rows of Fig. 1 (and 7 in Appendix A.1) we plot the precision of the real, DP, and synth classifiers on the two datasets (recall plots are also in Appendix A.1). For the DP classifiers, we find that precision drops disproportionately more for the underrepresented class when decreasing ϵ for all datasets.

Synth classifiers follow very similar behavior even with small privacy budgets ($\epsilon < 1$), regardless of the direction of class distortion in the synthetic datasets.

3.3. S2: Multi-Class Size and Recall

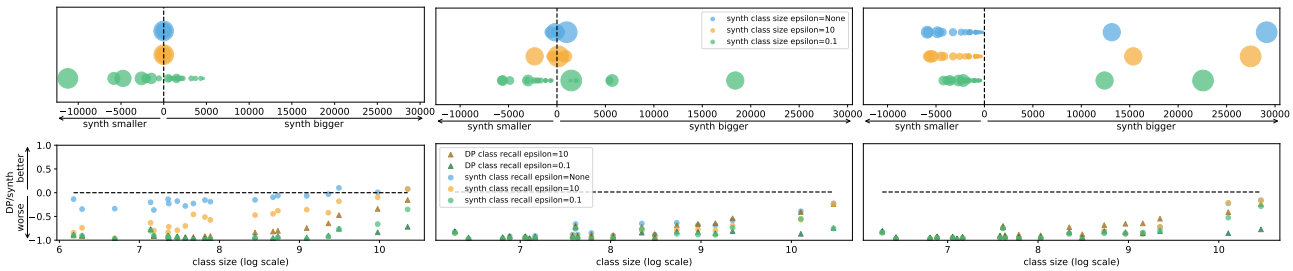
We experiment with privacy budgets (ϵ) of 0.1, 10, and infinity ("no-DP") for Purchases and 0.5, 5, 15, and infinity ("no-DP") for MNIST (to maintain consistency with previous work (Uniyal et al., 2021)). While for Purchases we do not imbalance the dataset, since it is already imbalanced (the proportion of the smallest to the largest class is 0.015), we imbalance the class "8" in MNIST to be 0.25 and 0.1 of the largest class. Furthermore, for MNIST, we only compare DP-SGD and PATE-GAN because the data has too many dimensions for PrivBayes.

Size. The size of the synthetic data is shown in the top rows of Fig. 2, 3, (and 10 in Appendix A.2). For the Purchases dataset, we see similar trends with PrivBayes and PATE-GAN as in S1; the former evens the classes, while the latter increases the gap by "transferring" counts from the minority to the majority classes. PATE-GAN exhibits the



(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Figure 1: Synthetic data class size (top) and real, DP, and synthetic classifiers precision (bottom) for different levels of ϵ, ***Adult***, (***S1***).



(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Figure 2: Synthetic data class (multi-class) size relative to real (top) (each bubble denotes a distinct class while the size its relative count in the real data) and DP and synthetic classifiers recall relative to real (bottom) for different levels of ϵ, ***Purchases***, (***S2***).

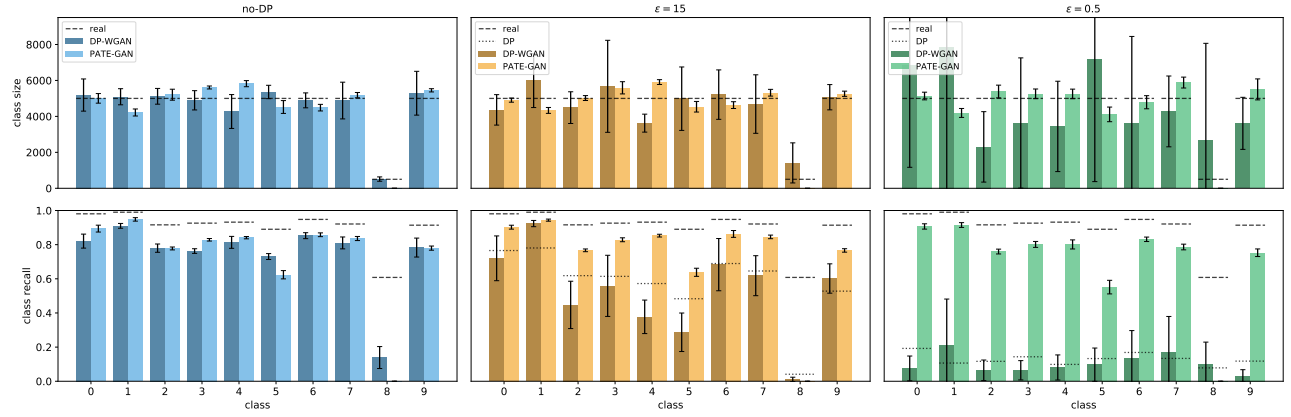


Figure 3: (S2) Synthetic data class (multi-class) size (top) and real, DP, and synthetic classifiers recall (bottom) for different digits and levels of ϵ, ***MNIST*** with class "8" downsampled to 0.1 its count, (***S2***).

strongest disparity, even with "no-DP," which contradicts findings from (Uniyal et al., 2021).However, the data imbalance here is of different nature as we have two classes with much higher counts than the others rather than a single underrepresented class. In turn, this could potentially bias the generator towards these classes as the teacher-discriminators are exposed predominantly to them and thus, learn to distinguish them from fake examples better. DP-WGAN does not preserve the class sizes as successfully as in S1.

For MNIST, PATE-GAN exhibits far better performance for both imbalances and preserves the counts even for lower ϵ budgets. Looking at the minority class "8," however, PATE-GAN fails to generate any digits for imbalance 0.1 (even for "no-DP" as well). This could be because the teachers fail to pass samples "8" labeled as real to the student even though when applied to classification, PATE is more robust under similar imbalance levels (Uniyal et al., 2021).

Recall. In the bottom two rows of Fig. 2, 3, (and 10 in Appendix A.2), we report the recall of the real, DP, and synth classifiers on the two datasets. For Purchases, the synth classifiers trained on data from PrivBayes far outperform both the DP classifiers and the other synth classifiers. Even with "no-DP," the synth classifiers trained on DP-WGAN and PATE-GAN incur a severe recall drop on smaller subgroups.

For MNIST, again PATE-GAN performs much better than

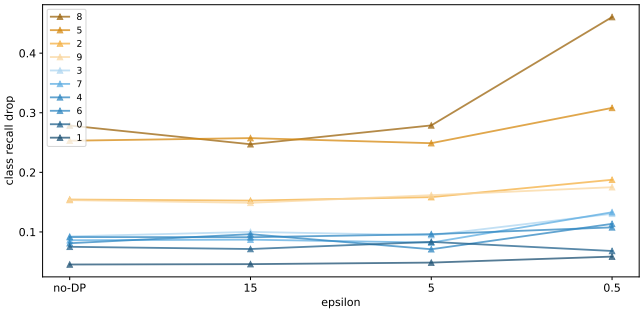


Figure 4: (S2) PATE-GAN synthetic classifier class recall drop relative to real for different digits and levels of ϵ, ***MNIST*** with class "8" downsampled to 0.25 its count, (***S2***).

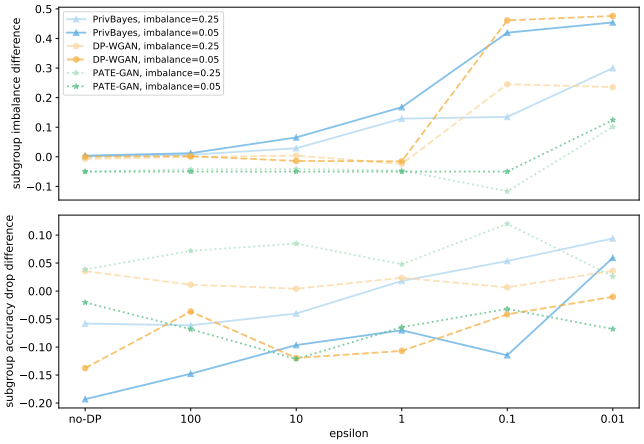


Figure 5: Minority single-attribute (sex) subgroup imbalance level difference (top) and minority subgroup accuracy drop difference (bottom) relative to majority for different subgroup imbalance and ϵ levels, ***Texas***, (***S3***).

DP-WGAN and DP classifiers – the recall drops are not so acute, and their standard deviations are much lower. DP-WGAN follows closely DP classifiers but is slightly worse for all levels of ϵ and imbalances. DP-WGAN's performance looks random for $\epsilon = 0.5$, which means that the classifiers failed to learn anything, most likely due to bad quality of the synthetic data (Fig. 11 in Appendix A.2). While expectedly DP-WGAN monotonically drops in terms of recall with decreasing ϵ, PATE-GAN's performance actually increases when DP is applied, e.g., $\epsilon = 15$ and 5 yield marginally better results than "no-DP" for both imbalances. This is most likely due to the fact that the teacher-discriminators are exposed to different subsets of the real data, and as result, do not learn exactly the same distributions as well as the noise added to their votes, which further enables generalization. Interestingly, the performance on some digits suffers a lot more than others (e.g., "2," "5," "9"), which could be explained because they are visually close to "8." This phenomenon is displayed in Fig. 4. The observation allows us to speculate that applying DP could not only lead to worse performance for the underrepresented subparts of the data but also for those with similar characteristics as well.

3.4. S3: Single-Attribute Subgroup Size and Accuracy

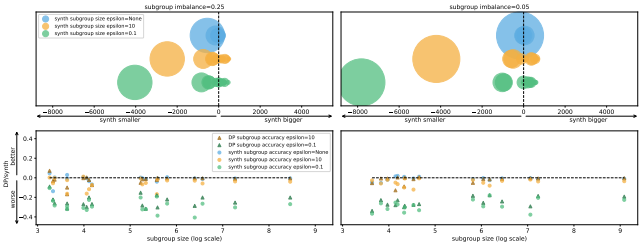
We treat a single feature – namely, sex – as a subgroup in the Adult and Texas datasets. We consider privacy budgets (ϵ) of 0.01, 0.1, 1, 10, 100, and infinity ("no-DP") as well as imbalance ratios of 0.01, 0.05, 0.1, 0.25, and 0.5.

Size. In the top rows of the plots in Fig. 12 in Appendix A.3, we report the full experiments for ϵ and imbalance effects on subgroup size on the two datasets while in Fig. 5 we summarize the trends for Texas. Once again, we find that, with PrivBayes, decreasing ϵ results in synthetic data with reduced subgroup imbalance for all datasets – the higher the initial imbalance, the more PrivBayes balances the subgroups (could be seen in the slope of the blue lines). As before, this effect could be attributed to the truncation of negative noisy counts after the Laplace mechanism is applied. PATE-GAN synthetic datasets follow the opposite trend. The gap becomes so large that, for imbalances lower than 0.1 for Adult and 0.25 for Texas, PATE-GAN barely generates the underrepresented subgroup for all ϵ values except 0.01. DP-WGAN is again the most successful at preserving the imbalance for $\epsilon > 0.1$. For $\epsilon = 0.1$ and 0.01, the subgroup size appears random, as the DP-WGAN models are trained only for a few iterations before the full privacy budget is spent.

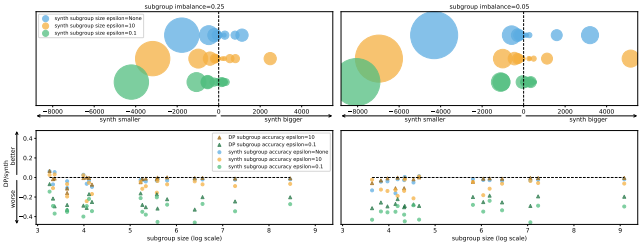
Accuracy. The bottom rows of Fig. 12 in Appendix A.3 report the accuracy of the various classifiers, while Fig. 5 displays a summary for Texas. Interestingly, we find that the real classifier, with Adult, achieves higher accuracy on the underrepresented subgroup "Female" than the overrepresented "Male" for all imbalances.

As for DP classifiers, over a certain ϵ, decreasing it further reduces the accuracy of the minority subgroup more than that of the majority. Additionally, this reduction in accuracy is more accentuated with increasing subgroup imbalance. For example, looking at the Adult plots, the accuracy on "Female" drops more than "Male" for $\epsilon \leq 0.1$ and imbalance 0.5. For increasing imbalances, the drop on the minority subgroup overtakes the majority for larger privacy budgets, i.e., $\epsilon \leq 1$ for imbalances 0.25 and 0.1, $\epsilon \leq 10$ for imbalance 0.05, and finally $\epsilon \leq 100$ for imbalance 0.01.

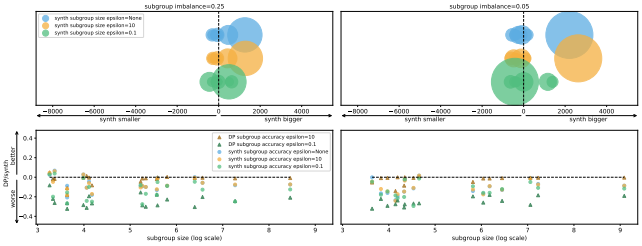
The synth classifiers incur a bigger accuracy drop in the underrepresented subgroup—regardless of the subgroup sizes in the synthetic data (as observed in the overall positive slopes of all lines in the bottom row of Fig. 5). Classifiers trained on PrivBayes synthetic data follow the behavior of DP classifiers the closest. Overall, DP-WGAN synth classifiers perform worse than the others as they are more unstable,



(a) PrivBayes



(b) DP-WGAN



(c) PATE-GAN

Figure 6: Synthetic data multi-attribute (intersection of age, sex, and race) subgroup size relative to real (top) (each bubble denotes a distinct subgroup while the size its relative count in the real data) and DP and synthetic classifiers accuracy relative to real (bottom) for different single-attribute (sex) subgroup imbalance and ϵ levels, ***Adult***, (***S4***).

with some noticeable accuracy drops (e.g., $\epsilon = 100$ for imbalance 0.5 in Adult, $\epsilon = 1$ for imbalances 0.1 and 0.05 in Texas). PATE-GAN synth classifiers have better accuracy than DP classifiers for both subgroups for $\epsilon < 10$; this is perhaps surprising, especially for the underrepresented subgroup, as the synth classifiers are trained on synthetic data containing only a small number of the minority subgroup.

3.5. S4: Multi-Attribute Subgroup Size and Accuracy

In our last set of experiments, we treat the intersection between three features – age, sex, and race – as complex subgroups in the Adult and Texas datasets. We consider privacy budgets (ϵ) of 0.1, 10, and infinity ("no-DP") as well as imbalance ratios of 0.05 and 0.25. This results in 16 subgroups for Adult, 21 for Texas for imbalance 0.05 and 19, and 27 for 0.25, respectively.

Size. In the top rows of the plots in Fig. 6 (also see Fig. 14 in Appendix A.4), we report the sizes of the subgroups, with the three different models and the two datasets. Once again, PrivBayes reduces the gap between majority and minority subgroups in the synthetic data, whereas PATE-GAN increases it. DP-WGAN behaves similarly to PrivBayes, which is inconsistent with the multi-class case discussed in Sec. 3.4. For Adult, DP-WGAN does not manage to keep the subgroups distribution, even for "no-DP." Finally, the effect of increased subgroup imbalance is evident in the higher disparity for all models and datasets.

Accuracy. The bottom rows of the plots in Fig. 6 (and 14) report the performance of the classifiers. In contrast to previous work (Bagdasaryan et al., 2019), we do not observe "the rich get richer, the poor get poorer" effect for either DP or synth classifiers; instead, "everybody gets poorer," with very few exceptions. This should not come as a surprise, as it could not be expected from a generative model to synthesize data with better utility than the real data. The mentioned exceptions are only for subgroups with small sizes; however, the accuracy on these subgroups has a much higher standard deviation compared to larger subgroups.

There is also a clear distinction between classifiers with $\epsilon = 10$ and 0.1, both DP and synth. Synth classifiers trained on PATE-GAN and PrivBayes synthetic data incur a smaller drop than DP-WGAN synthetic data. Similar to S3, PATE-GAN trained synth classifiers have better accuracy than DP classifiers for $\epsilon = 0.1$.

3.6. Main Take-Aways

Disparate Effects. Overall, our experiments provide an empirical demonstration that DP generative models *do* have different disparate effects on synthetic data. Analyzing the size of the classes and subgroups in the generated data, we consistently observe that PrivBayes reduces the gap between the majority and minority classes/subgroups (thus exhibiting a "Robin Hood" effect), PATE-GAN increases it (exhibiting a "Matthew" effect), and DP-WGAN has a mixed one.

Downstream Classification. When performing classification tasks on data produced by generative models, one faces an even bigger (or more variable) accuracy drop on minority classes and subgroups. We also see that higher privacy guarantees and more imbalanced datasets result in more substantial disparate effects. For example, even though PATE-GAN displays better overall behavior than DP-WGAN, an imbalance of 0.1 prevents it from learning an entire class of the data even for low privacy settings. High privacy guarantees also result in PATE-GAN generating undesirable artifacts in the synthetic data in the form of a much stronger correlation between low correlated columns.

Revisiting Our Research Questions. Recall from Section 1 that our work aimed to answer three main research questions; we now summarize some concise answers.

RQ1: Do DP generative models generate data in similar classes and subgroups proportions to the real data?
Not really. DP distorts the proportions, yielding Robin Hood vs Matthew effects depending on the DP generative model.

RQ2: Does training a classifier on DP synthetic data lead to the same disparate impact on accuracy as training a DP classifier on the real data?
Overall, yes. Smaller classes/subgroups suffer more similarly to DP classifiers. However, we do not see the rich get richer, the poor get poorer; everybody gets poorer. Incidentally, sometimes synthetic classifiers are better than DP classifiers; studying this in detail is left to future work.

RQ3: Do different DP mechanisms for DP synthetic data behave similarly under different privacy and data imbalance levels?
No, different DP generative models behave differently. For example, PATE-GAN performs better than DP-WGAN, with some very specific exceptions, while PrivBayes is the only one that manages to maintain the data utility for the multi-class tabular data Purchases.

4. Related Work

(Kuppam et al., 2019) show that, if resource allocation is decided based on DP statistics, smaller districts could get more funding and larger ones less. Prior work has also studied deep neural network classifiers trained using DP-SGD (Bagdasaryan et al., 2019; Farrand et al., 2020; Suriyakumar et al., 2021) on imbalanced datasets (mainly images). Essentially, they show that underrepresented groups in a dataset that already incur lower accuracy end up losing even more accuracy when DP is applied. In particular, (Farrand et al., 2020) show that even small imbalances and loose privacy guarantees can cause disparate impacts.

(Uniyal et al., 2021) find that classifiers trained with PATE exhibit disparate drops in performance but less severely than with DP-SGD. Furthermore, (Feldman, 2020) formalizes the need for accurate discriminative models to memorize training data and studies the disparate effects of privacy and model compression on subgroups. (Chen et al., 2020b) provide a theoretical analysis that quantifies the clipping bias on convergence with a disparity measure between the gradient and a geometrically symmetric distribution. There are also papers focusing on learning DP classifiers with fairness constraints (Jagielski et al., 2019; Tran et al., 2021b) and on analyzing the PATE framework from a fairness point of view (Tran et al., 2021a). Unlike our work, these efforts focus on discriminative (rather than generative) models.

(Cheng et al., 2021) show that training classifiers on DP synthetic images can result in significant utility degradation and increased majority subgroup influence, but not worse group unfairness measures. However, they only use a single generative model and only look at the utility of balanced DP synthetic datasets.

Finally, recent work has focused on tabular DP synthetic data. (Ghalebikesabi et al., 2021) introduce novel bias mitigation techniques, which, unfortunately, lead to reduced usefulness of the synthetic data. Perhaps closer to our work is that by (Pereira et al., 2021), who mainly look at single-attribute subgroup fairness and overall classification performance. Their work, however, does not investigate the utility disparity on different single and multi-attribute subgroups of the data, nor the effect of data imbalance. Furthermore, we consider the size disparities in the generated synthetic data as we do not use conditional generative models. We also experiment with a far wider range of epsilon budgets. Overall, we are the first to highlight and analyze the disparate effects of several factors: generative model type, DP mechanism, privacy budget, class and single/multi-attribute subgroup imbalance on the resulting synthetic data in terms of both statistical analysis and downstream classification. We do so through experiments geared to isolate the effect of these factors.

There is a rich literature with DP generative models for tabular data (Acs et al., 2018; Xie et al., 2018; Tantipongpipat et al., 2021; Frigerio et al., 2019; Zhang et al., 2021; McKenna et al., 2021). Our goal is, however, not to benchmark all possible models but to focus on the best known and accessible state-of-the-art models relying on different/well-studied DP mechanisms (Laplace, DP-SGD, PATE).

5. Conclusion

This work analyzed the effects of privacy-preserving generative models, using different DP methods, on 1) class/subgroups distributions in the generated synthetic data and 2) the performance of downstream tasks. We found that applying DP to synthetic data generation disparately affects the minority subpopulations. As for the class/subgroup distribution in the synthetic data, DP can have opposing effects depending on the underlying DP method; e.g., PrivBayes reduces the imbalance, PATE-GAN increases it. However, when training a classifier on the synthetic data, minority subpopulations suffer stronger and/or more varying decreases in accuracy. We also showed that the privacy budget and data imbalance are important factors and further intensify these effects.

Overall, our work motivates the need for practitioners and companies to take the disparate effects into consideration and adopt more extensive testing before deploying synthetic data, given that the studied technologies are already in production in the real world (Brown, 2020; US Census Bureau, 2021) and there are concerns and scepticism from the public (Wezerek & Van Riper, 2020; Hong, 2020).

We are confident that our results will motivate further research (including theoretical contributions) at the intersection of generative models, DP, and fairness. Hopefully, this will include novel generative models with modified/new DP learning algorithms that could reproduce the original data in a privacy-preserving manner and without disparate loss in utility. Another interesting direction would be to examine the conditions under which classifiers trained on DP synthetic data achieve better utility than DP classifiers trained on real data as observed in this paper.

To facilitate further research in this space, including reproducibility of our results and analysis a wider set of hyperparameters, additional generative models, datasets, and/or tasks beyond classification, we are also open-sourcing our code.[2] As part of future work, we plan to explore the relationship between disparate effects and fairness, support additional experiments, as well as release a re-usable, modular framework that integrates with other security and privacy evaluations of both discriminative and generative models, e.g., (Liu et al., 2021; Stadler et al., 2022).

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *ACM CCS*, 2016.

Acs, G., Melis, L., Castelluccia, C., and De Cristofaro, E. Differentially private mixture of generative neural networks. *IEEE TKDE*, 2018.

Alzantot, M. and Srivastava, M. Differential Privacy Synthetic Data Generation using WGANs. https://github.com/nesl/nist_differential_privacy_synthetic_data_challenge/, 2019.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *ICML*, 2017.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential privacy has disparate impact on model accuracy. In *NeurIPS*, 2019.

Barber, D. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

Benedetto, G., Stanley, J. C., Totty, E., et al. The creation and use of the SIPP synthetic Beta v7. 0. *US Census Bureau*, 2018.

Brown, A. Synthetic Data Promises Fair AI And Privacy Compliance, But How Exactly Does It Work? https://tinyurl.com/yc5vtrhb, 2020.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*, 2019.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *JMLR*, 2011.

Chen, D., Yu, N., Zhang, Y., and Fritz, M. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *ACM CCS*, 2020a.

Chen, X., Wu, S. Z., and Hong, M. Understanding gradient clipping in private SGD: A geometric perspective. *NeurIPS*, 2020b.

Cheng, V., Suriyakumar, V. M., Dullerud, N., Joshi, S., and Ghassemi, M. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *ACM FAccT*, 2021.

DSHS. Texas Hospital Inpatient Discharge Public Use Data File Q1-Q4, 2013. https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm, 2013.

Dua, D. and Graff, C. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/adult, 2017.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *EuroCrypt*, 2006a.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006b.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.

Farrand, T., Mireshghallah, F., Singh, S., and Trask, A. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Workshop on Privacy-Preserving Machine Learning in Practice*, 2020.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *STOC*, 2020.

Frigerio, L., de Oliveira, A. S., Gomez, L., and Duverger, P. Differentially private generative adversarial networks for time series, continuous, and discrete open data. In *IFIP SEC*, 2019.

Ghalebikesabi, S., Wilde, H., Jewson, J., Doucet, A., Vollmer, S., and Holmes, C. Bias Mitigated Learning from Differentially Private Synthetic Data: A Cautionary Tale. *arXiv:2108.10934*, 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *NeurIPS*, 2014.

[2] https://github.com/ganevgv/dp-gen-disparate

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *NeurIPS*, 2016.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: Membership inference attacks against generative models. In *PoPETs*, 2019.

Holohan, N., Braghin, S., Mac Aonghusa, P., and Levacher, K. Diffprivlib: the IBM differential privacy library. *arXiv:1907.02444*, 2019.

Hong, J. Census 2020 +/- 2: Census, Differential Privacy, and the Future of Data. https://www.hawaiidata.org/news/2020/9/24/census2020-census-differential-privacy-future-of-data, 2020.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. Differentially private fair learning. In *ICML*, 2019.

Jordon, J., Yoon, J., and Van Der Schaar, M. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *ICLR*, 2018.

Kaggle. Acquire Valued Shoppers Challenge. https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data, 2013.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

Kuppam, S., McKenna, R., Pujol, D., Hay, M., Machanavajjhala, A., and Miklau, G. Fair decision making using privacy-protected data. *arXiv:1905.12744*, 2019.

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database. *ATT Labs*, 2010.

Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., De Cristofaro, E., Fritz, M., and Zhang, Y. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. *arXiv:2102.02551*, 2021.

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

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *FOCS*, 2007.

NHS England. A&E Synthetic Data. https://data.england.nhs.uk/dataset/a-e-synthetic-data, 2021.

NIST. 2018 Differential Privacy Synthetic Data Challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic, 2018a.

NIST. 2018 The Unlinkable Data Challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-unlinkable-data-challenge, 2018b.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv:1610.05755*, 2016.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, Ú. Scalable private learning with pate. *arXiv:1802.08908*, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

Pereira, M., Kshirsagar, M., Mukherjee, S., Dodhia, R., and Ferres, J. L. An Analysis of the Deployment of Models Trained on Private Tabular Synthetic Data: Unexpected Surprises. *arXiv:2106.10241*, 2021.

Ping, H., Stoyanovich, J., and Howe, B. DataSynthesizer. https://github.com/DataResponsibly/DataSynthesizer, 2017.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *IEEE S&P*, 2017.

Stadler, T., Oprisanu, B., and Troncoso, C. Synthetic Data – Anonymization Groundhog Day. In *Usenix Security*, 2022.

Suriyakumar, V. M., Papernot, N., Goldenberg, A., and Ghassemi, M. Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. In *ACM FAccT*, 2021.

Tantipongpipat, U., Waites, C., Boob, D., Siva, A., and Cummings, R. Differentially private mixed-type data generation for unsupervised learning. 2021.

Thompson, S. A. and Warzel, C. The Privacy Project: Twelve Million Phones, One Dataset, Zero Privacy. https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html, 2019.

Tran, C., Dinh, M. H., Beiter, K., and Fioretto, F. A Fairness Analysis on Private Aggregation of Teacher Ensembles. *arXiv:2109.08630*, 2021a.

Tran, C., Fioretto, F., and Van Hentenryck, P. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. *AAAI*, 2021b.

Uniyal, A., Naidu, R., Kotti, S., Singh, S., Kenfack, P. J., Mireshghallah, F., and Trask, A. DP-SGD vs PATE:

Which Has Less Disparate Impact on Model Accuracy? *arXiv:2106.12576*, 2021.

US Census Bureau. Differential Privacy and the 2020 Census. https://www.census.gov/library/fact-sheets/2021/differential-privacy-and-the-2020-census.html, 2021.

Van Der Schaar, M. and Maxfield, N. Synthetic data: Breaking the data logjam in machine learning for healthcare. https://tinyurl.com/2hr4atnn, 2020.

Webster, R., Rabin, J., Simon, L., and Jurie, F. Detecting overfitting of deep generative networks via latent recovery. In *IEEE CVPR*, 2019.

Wezerek, G. and Van Riper, D. Changes to the Census Could Make Small Towns Disappear. https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html, 2020.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially private generative adversarial network. *arXiv:1802.06739*, 2018.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems*, 2017.

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

A. Additional Results and Plots

A.1. S1: Texas and Recall Plots

In Fig. 7 we show the size of the binary class in the real and synthetic data as well as the precision of the real, DP, and synth classifiers for the Texas dataset, they are discussed in Section 3.2.

In Fig. 8 and 9 we plot the recall of the real, DP, and synth classifiers on the Adult and Texas datasets. For the DP classifiers, recall follows similar patterns as precision for Adult, while, for Texas, there is close to no drop for the underrepresented class and a small drop for the overrepresented class for $\epsilon < 0.1$.

For all synth classifiers in the two datasets recall looks more noisy than precision; for most cases (except for PATE-GAN with $\epsilon = 0.01$ in Adult), after initially declining with decreasing ϵ values, the recall of underrepresented class actually starts increasing for $\epsilon < 0.1$. This is most likely because the generated synthetic data is more random; indeed, this is also evident from the large standard deviations in the class size and recall values. It is interesting to observe that for PATE-GAN in Fig. 9(c), the underrepresented class recall is actually larger than the real baseline for $\epsilon > 1$.

A.2. S2: Further Purchases and MNIST Plots

For the Purchases dataset, Fig. 16 is duplicate of Fig. 2 but with included error bars. Looking at the top row, where we display the size of the classes, we observe that PrivBayes has lowest standard deviation (approximately none for "no-DP" and $\epsilon = 10$), while DP-WGAN the highest. For PATE-GAN, unlike the other two models, bigger classes exhibit larger standard deviation. Finally, observing the recall in the bottom row, PATE-GAN classifiers have the lowest variation.

For MNIST, Fig. 10 displays the size and recall on all digits for MNIST with class "8" undersampled to 0.25 its original size. In Fig. 11 we can see random synthetic samples produced by DP-WGAN and PATE-GAN for various ϵ budgets. The results from both plots are analyzed in Section 3.3. We do not plot or analyze the precision for S2, as the trends are almost identical to recall.

A.3. S3: Full Plots

Due to space limitation, in Fig. 12 we plot the full set of experiments that are discussed in detail in Section 3.4.

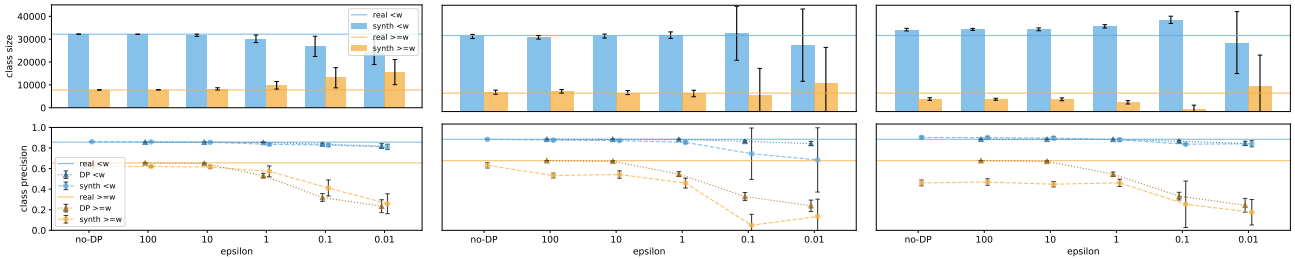
Additionally, in Fig. 13 we plot a summary of the overall trends for the Adult dataset. From the top row, it could be seen that if there is imbalance in the subgroups PrivBayes balances the data set (dark blue lines have positive slope), DP-WGAN maintains well the imbalance for $\epsilon < 1$ (orange dashed lines stay around 0), while PATE-GAN increases it (dotted green lines are negative). Looking at the bottom row, it could be seen that almost all lines have a positive slope, meaning that with decreased ϵ the accuracy of the minority subgroup drop quicker than the majority. Furthermore, darker lines (i.e., more imbalanced datasets) tend to be on top of lighter (i.e., more balanced datasets) which means that increasing the imbalance results in even larger/quicker minority subgroup accuracy drop relative to the majority.

A.4. S4: Texas Plots

Due to space limitation, in Fig. 14 we plot the multi-attribute subgroup experiments for the Texas dataset, they are discussed in Section 3.5.

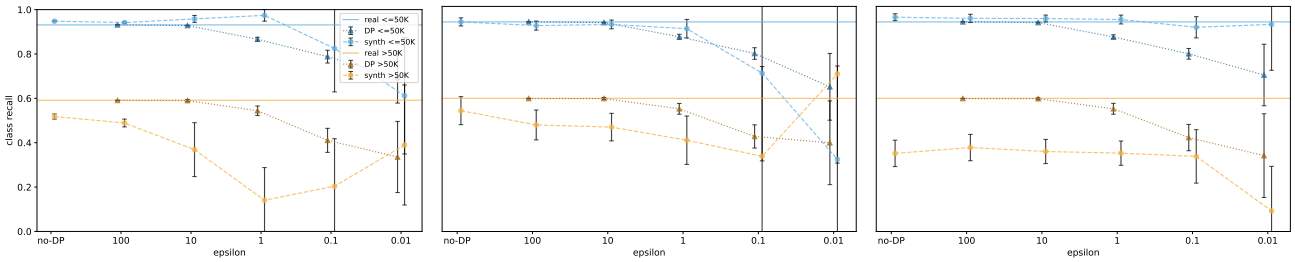
A.5. S3 and S4: Correlations

Correlation in S3. The mutual information between the subgroup and the target columns are displayed in Fig. 17. Since in the preparation step the datasets the subgroups were balanced by class, the expected (and the real) mutual information is 0. We observe that, DP-WGAN manages best to maintain this relationship, closely followed by PrivBayes which, however, has a few noisy exceptions for small privacy budgets $\epsilon <= 0.1$. In contrast, PATE-GAN introduces undesirable artifact in the synthetic data for $\epsilon <= 0.1$ for



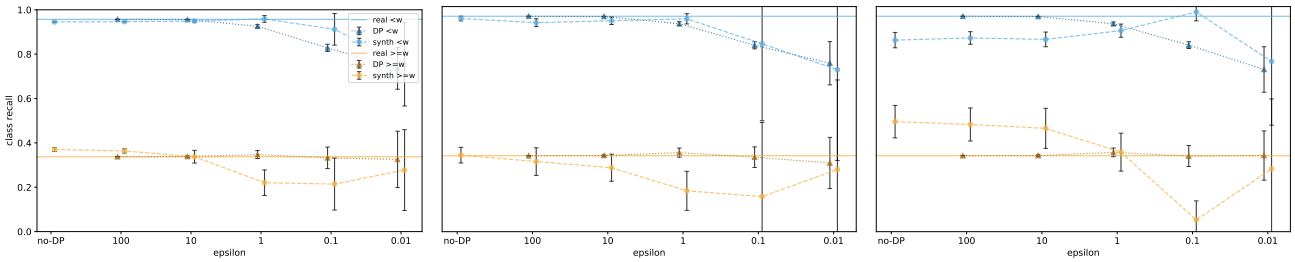
(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Figure 7: Synthetic data class size (top) and real, DP, and synthetic classifiers precision (bottom) for different levels of ϵ, ***Texas***, (***S1***).



(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Figure 8: Real, DP, and synthetic classifiers recall for different levels of ϵ, ***Adult***, (***S1***).



(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Figure 9: Real, DP, and synthetic classifiers recall for different levels of ϵ, ***Texas***, (***S1***).

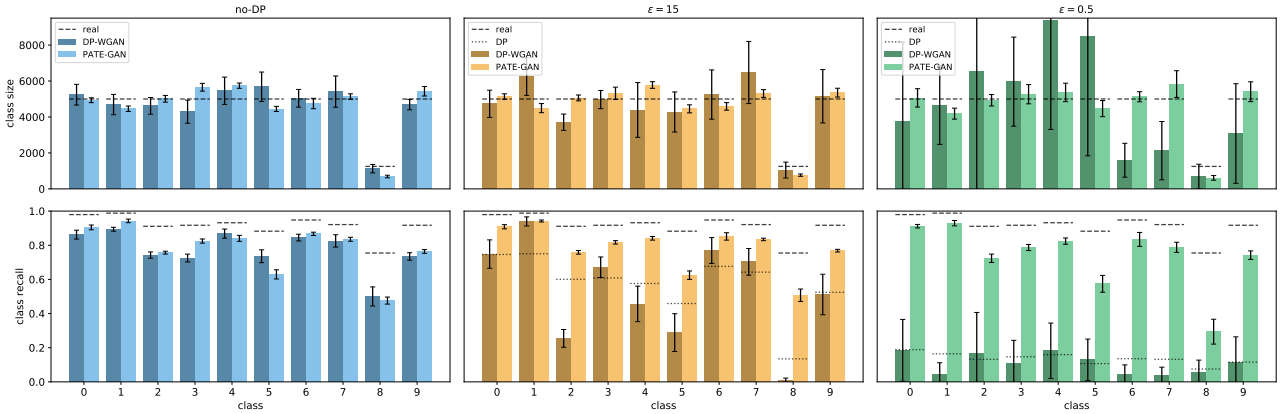
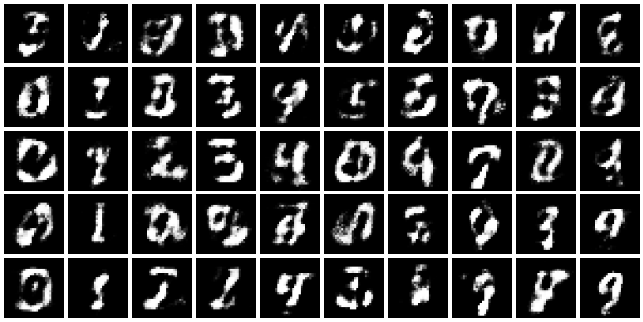
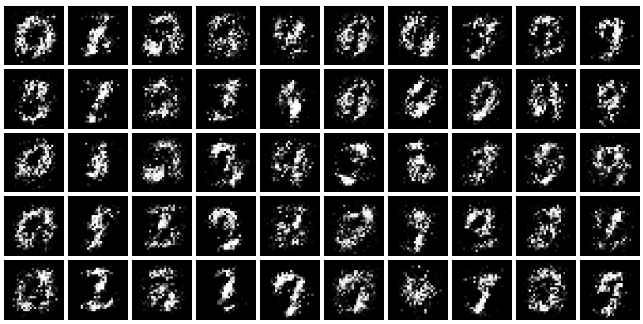


Figure 10: Synthetic data class (multi-class) size (top) and real, DP, and synthetic classifiers recall (bottom) for different digits and levels of ϵ, ***MNIST*** with class "8" downsampled to 0.25 its count, (***S2***).



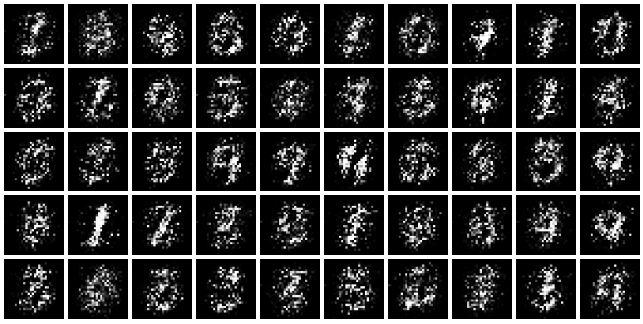
(a) DP-WGAN, no-DP

(b) PATE-GAN, no-DP



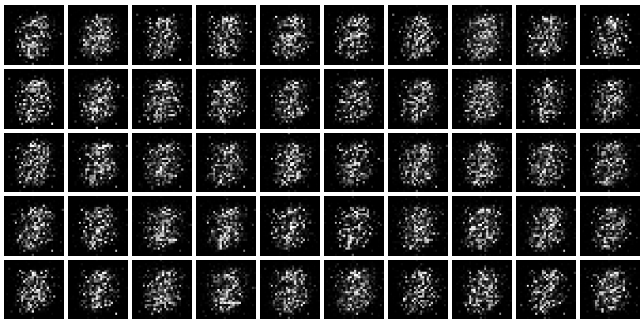
(c) DP-WGAN, $\epsilon = 15$

(d) PATE-GAN, $\epsilon = 15$

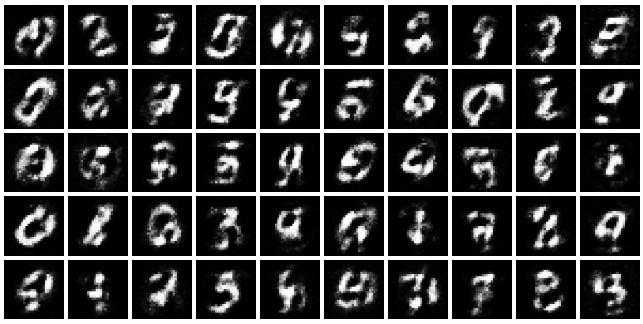


(e) DP-WGAN, $\epsilon = 5$

(f) PATE-GAN, $\epsilon = 5$

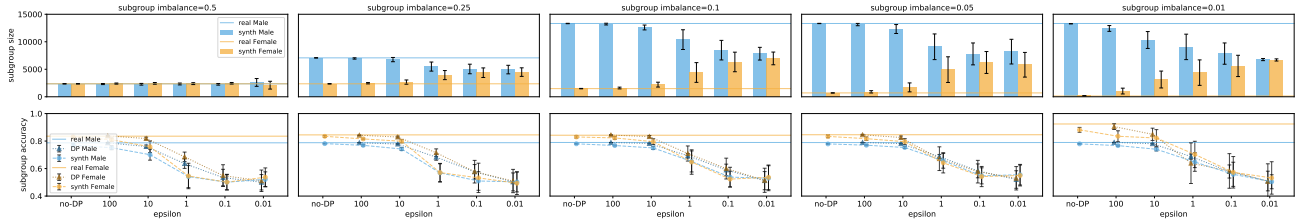


(g) DP-WGAN, $\epsilon = 0.5$

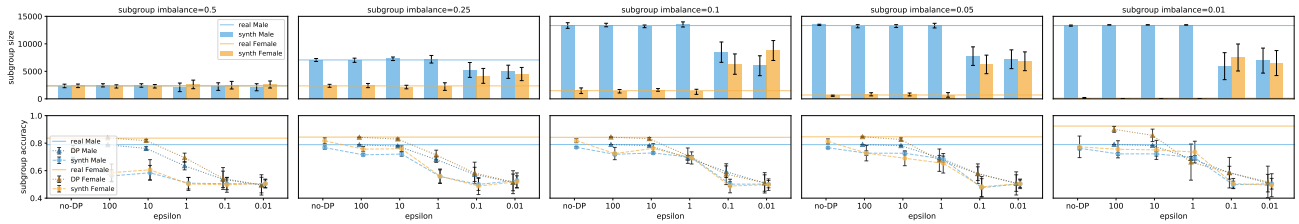


(h) PATE-GAN, $\epsilon = 0.5$

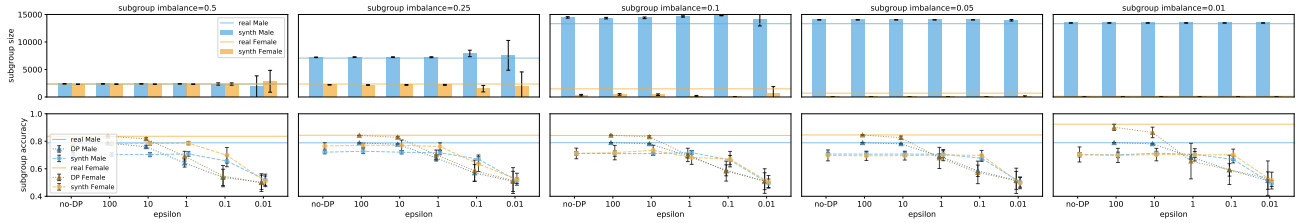
Figure 11: Synthetic samples (ordered from "0" to "9" in each subplot) generated by DP-WGAN (left) and PATE-GAN (right) for different ϵ levels, ***MNIST*** with class "8" downsampled to 0.25 its count, (***S2***).



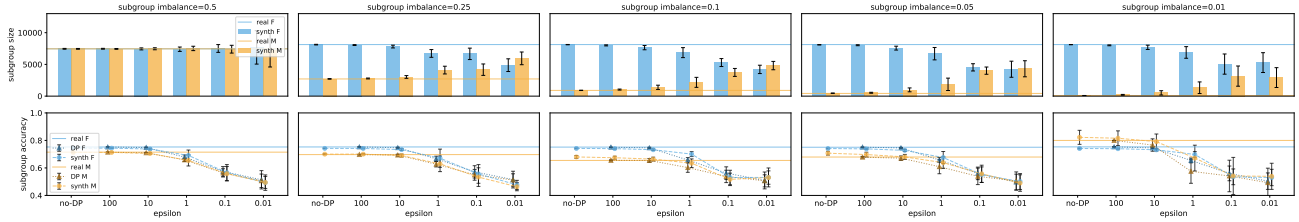
(a) PrivBayes, Adult



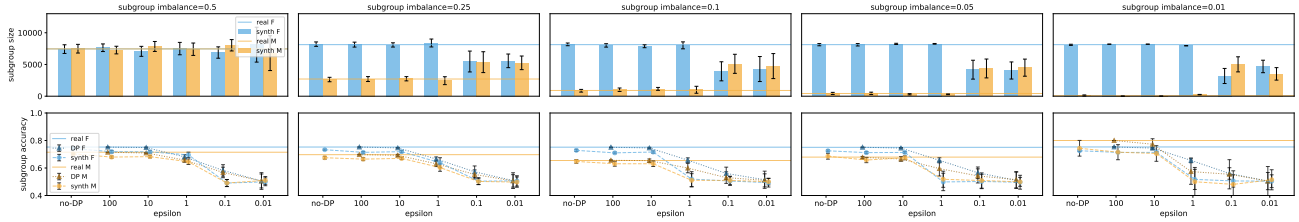
(b) DP-WGAN, Adult



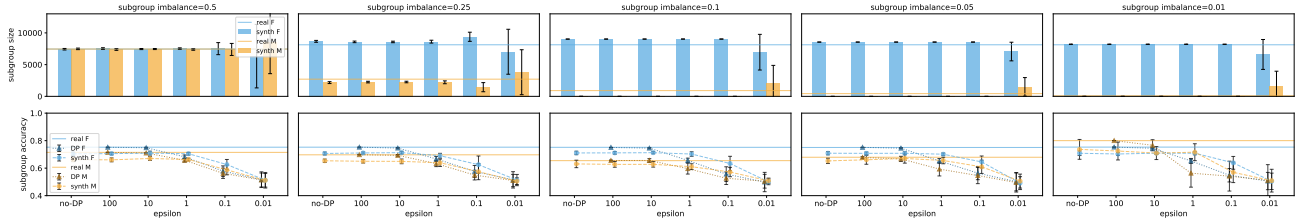
(c) PATE-GAN, Adult



(d) PrivBayes, Texas



(e) DP-WGAN, Texas



(f) PATE-GAN, Texas

Figure 12: Synthetic data single-attribute (sex) subgroup size (top) and real, DP, and synthetic classifiers recall accuracy (bottom) for different single-attribute subgroup imbalance and ϵ levels, ***Adult*** (top 3) and ***Texas*** (bottom 3), (***S3***).

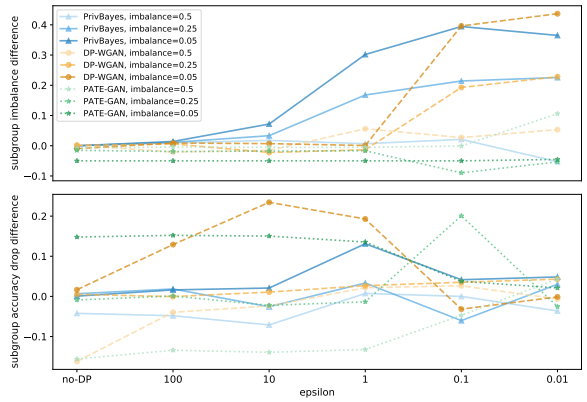
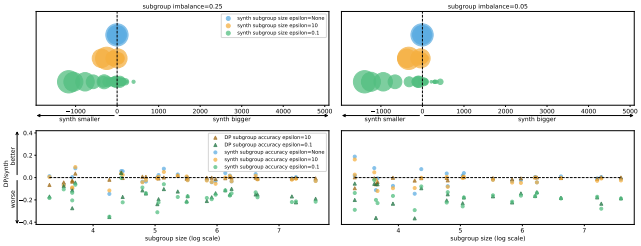
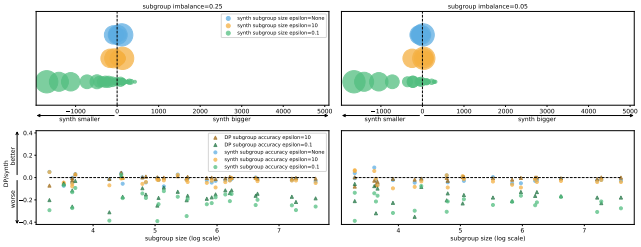


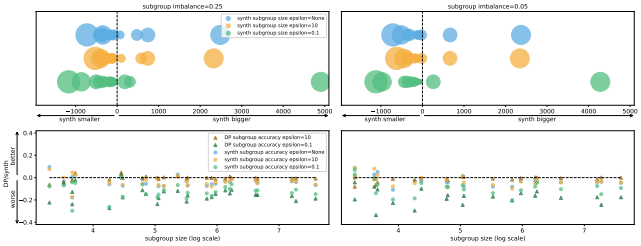
Figure 13: Minority single-attribute (sex) subgroup imbalance level difference (top) and minority subgroup accuracy drop difference (bottom) relative to majority for different subgroup imbalance and ϵ levels, ***Adult***, (***S3***).



(a) PrivBayes



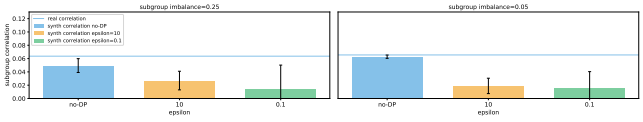
(b) DP-WGAN



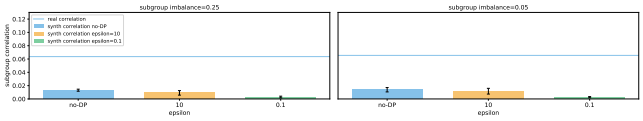
(c) PATE-GAN

Figure 14: Synthetic data multi-attribute (intersection of age, sex, and race) subgroup size relative to real (top) (each bubble denotes a distinct subgroup while the size its relative count in the real data) and DP and synthetic classifiers accuracy relative to real (bottom) for different single-attribute (sex) subgroup imbalance and ϵ levels, ***Texas***, (***S4***).

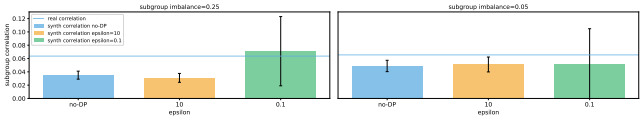
both datasets. In other words, the model creates data with stronger relationship between the subgroup column and the target column.



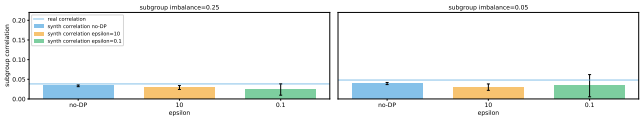
(a) PrivBayes, Adult



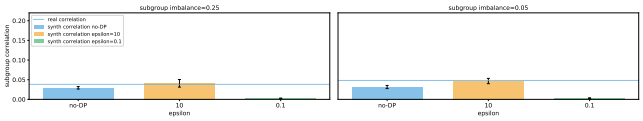
(b) DP-WGAN, Adult



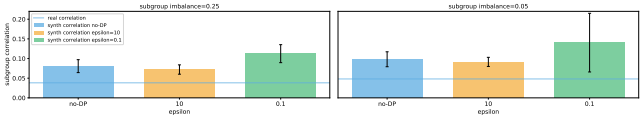
(c) PATE-GAN, Adult



(d) PrivBayes, Texas



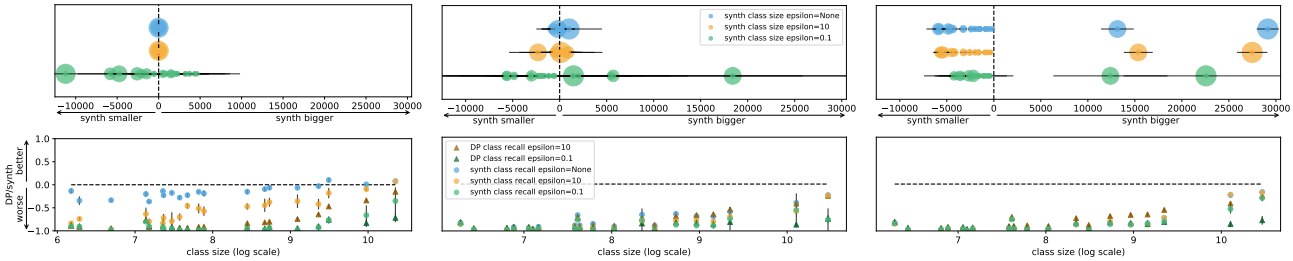
(e) DP-WGAN, Texas



(f) PATE-GAN, Texas

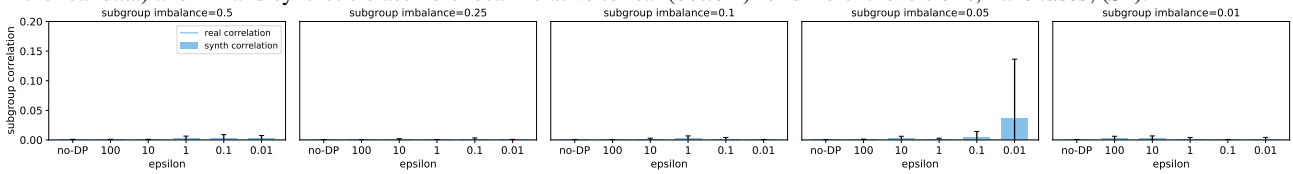
Figure 15: Mutual information between the multi-attribute (intersection of age, sex, and race) subgroup and the target (income/length of stay) columns for different single-attribute subgroup imbalance (sex) and ϵ levels, ***Adult*** (top 3) and ***Texas*** (bottom 3), (***S4***).

Correlation in S4. In Fig. 15, we display the mutual information between the multi-attribute subgroup and the target. Unlike the single-attribute scenario, the baseline mutual information here is not 0 because only one of the attributes (sex) was balanced by class. For both datasets, PrivBayes exhibits the most expected behavior – increasing the privacy budget ϵ results in more distorted synthetic data, thus reducing the mutual information between the subgroup and target columns. On the other hand, PATE-GAN displays similar to u-shaped behavior: incorporating some privacy ($\epsilon = 10$) initially reduces the mutual information but adding more privacy ($\epsilon = 0.1$) increases it to level even higher than when "no-DP" is applied. In particular, for the Texas dataset, PATE-GAN enforces the dependency between the subgroups and target columns for all privacy budgets. Finally, DP-WGAN performs the worst in the Adult dataset.

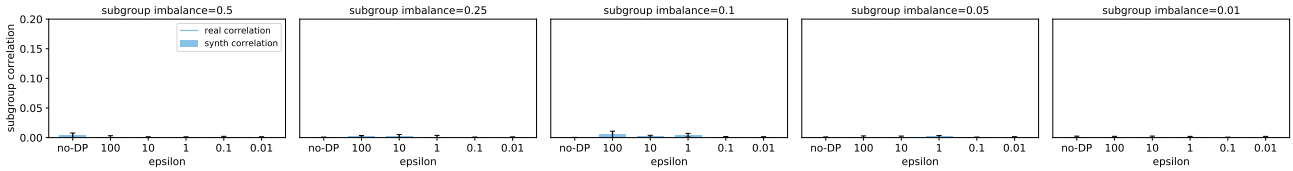


(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

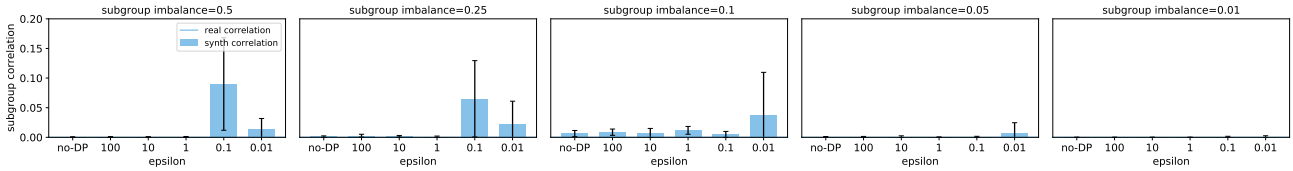
Figure 16: Synthetic data class (multi-class) size relative to real (top) (each bubble denotes a distinct class while the size its relative count in the real data) and DP and synthetic classifiers recall relative to real (bottom) for different levels of ϵ, ***Purchases***, (***S2***).



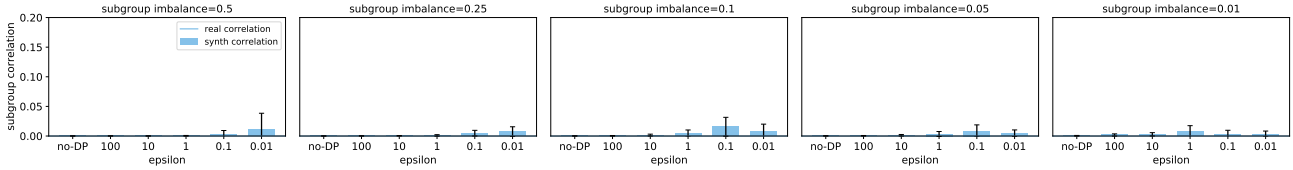
(a) PrivBayes, Adult



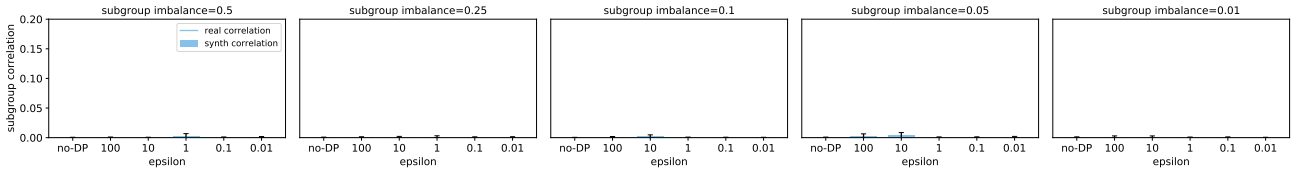
(b) DP-WGAN, Adult



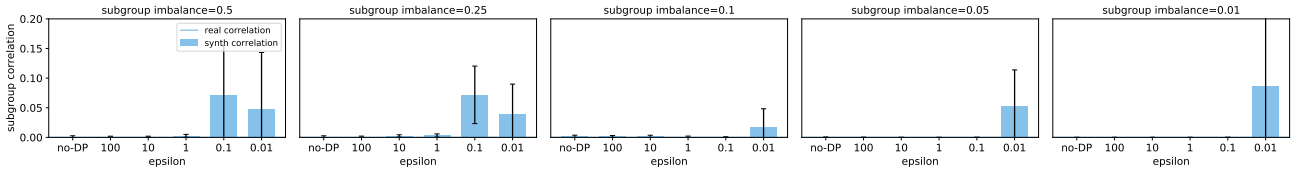
(c) PATE-GAN, Adult



(d) PrivBayes, Texas



(e) DP-WGAN, Texas



(f) PATE-GAN, Texas

Figure 17: Mutual information between the single-attribute subgroup (sex) and the target (income/length of stay) columns for different single-attribute subgroup imbalance and ϵ levels, ***Adult*** (top 3) and ***Texas*** (bottom 3), (***S3***).