FFPDG: FAST, FAIR AND PRIVATE DATA GENERATION

Weijie Xu, Jinjin Zhao, Francis Iannacci, Bo Wang
Amazon

ABSTRACT

Generative modeling has been used frequently in synthetic data generation. Fairness and privacy are two big concerns for synthetic data. Although Recent GAN [Goodfellow et al. (2014)] based methods show good results in preserving privacy, the generated data may be more biased. At the same time, these methods require high computation resources. In this work, We design a fast, fair, flexible and private data generation method. We show the effectiveness of our method theoretically and empirically. We show that models trained on data generated by the proposed method can perform well (in inference stage) on real application scenarios.

1 INTRODUCTION

Synthetic data [Rubin] is data that is artificially created rather than being generated by actual events. The availability of large synthetic data can bring many collaboration opportunities between related industries and research community. Synthetic data has been valuable to business functions, such as health [Hwang et al. (2020)], robotics [Duczek et al. (2021)], and financial services [Samuel Assefa & Manuela Veloso1]. However, personnel related synthetic data is rarely available online. [Rich (2020)]

There are two reasons. First, personal information is very sensitive. [Navaz et al. (2013)] Any potential leakage of personal information is harmful. Recent work shows that machine learning models are highly susceptible to leak information from their training data [Reza Shokri & Shmatikov. (2017)]. If attackers have some real data, they are able to distinguish synthetic data from real data with high accuracy. [Choquette-Choo et al. (2021)] Second, models trained on biased data have different predictive power across protected groups, such as race or gender. Bringing equal opportunity to protected groups is essential for synthetic data, as noted in [Bolukbasi et al. (2016) Joy Buolamwini (2018)]. Many methods [Donini et al. (2020) Zafar et al. (2017) Zafar et al. (2019) Hardt et al. (2016) Woodworth et al. (2017)] have been provided to mitigate bias during model training or post model training. However, [Agarwal Rachel Cummings (2019)] prove that there is no algorithm that is private, fair and better than a constant classifier. The proof relies on that test data distribution can be very different. Thus, we attempt to make datasets fair and private before training any machine learning methods.

In this work, we propose a fast, fair, and private data generation method (FFPDG). Our algorithm can be used in supervised and unsupervised modeling applications. Our algorithm is fast since it has low run time complexity. We show that our algorithm has good fairness guarantee through experiments and mathematical proof. We show that our method preserves privacy under certain conditions.

2 RELATED WORK

2.1 PRIVATE DATA GENERATION

[Feldman et al. (2015)] shows that removing personal information does not protect privacy. A privacy preserving dataset does not change outcome very much by inclusion or exclusion of a particular example.

Definition 1(Differencial Privacy). A mechanism A on a query functions f is ϵ-differentially private if for all neighboring datasets $X, X^{'}$ which differ in a single record and for all possible measurement $S \subseteq R$, $\frac{Pr[A(f(X)) \in S]}{Pr[A(f(X')) \in S]} \leq exp(\epsilon)$ [Cynthia Dwork & Naor (2006)]

Differentially Private Stochastic Gradient Descent (DP-SGD) [Abadi et al. (2016)] is one of the first studies to make the Stochatic Gradient Descent computation differential private. It preserves differential privacy by clipping the gradient in the optimization's l_2 norm and adding noise.

Private Aggregation of Teach Ensembles (PATE)[Papernot et al. (2017) Papernot et al. (2018)]deploy multiple teacher models which trained by disjoint datasets. PATE then deploy the teach models to unseen data to make predictions. For unseen data, the teacher models vote to determine the label combining random noise generated by laplace distribution. PATE further trains student model by only accessing to the privatized labels generated from the teacher's vote. Student cannot relearn an individual teacher's model as teacher is trained on disjoint datasets with laplacian noise.

Generative Adversarial Network and Variational Autoencoders [Kingma & Welling (2014)] provide powerful methods to generate synthetic data using real data but they do not provide any privacy guarantee. These generation methods can combine with differentially private methods to achieve better privacy guarantee for generated data.

DPGAN [Xie et al. (2018)] proposes a framework for modifying the GAN [Goodfellow et al. (2014)] framework to be differentially private, which relies on the PostProcessing Theorem [Dwork & Roth (2014)] by learning a differentially private discriminator during training.

PATE-GAN [Yoon et al. (2019)] modifies PATE framework to apply to GANs by applying PATE mechanism to the discriminator. The dataset is first partitioned into k subsets to train k teachers. Each teacher is trained by discrimination between generated data by generator and a disjoint subset of original data. The student discriminator is trained by voting from teachers result plus laplacian noise. At the end, the generator is trained to fool the student discriminator.

Random OrthoNormal projection with GAUSSian generative model(RON-Gauss) [Chanyaswad (2019)] combines dimension reduction via random orthonormal projection and the Gaussian generative model for synthesizing differential private data. They are inspired by Diaconis-Freedman-Meckes (DFM) effect [Meckes (2012)] which shows that under suitable conditions, most projections of high dimensional data are nearly Gaussian. The data was generated by a process of normalization, random projection and gaussian modeling. Noise is added in the data normalization and gaussian model stages.

2.2 Fair Processing

We first give definition of Disparate Impact [Barocas & Narayanan (2018)], Equal Opportunity[Heidari et al. (2018)] and Statistical Disparity.

Definition 2(Disparate Impact). Given data set $D = (X, Y, C)$ with binary protected attribute C (e.g. race, sex, religion, etc, 0 is unprivileged group and 1 is privileged group), remaining attributes X and binary class to be predicted Y, we will say that C has disparate impact if $\frac{Pr(Y=1|C=0)}{Pr(Y=1|C=1)} \leq 0.8$.

Definition 3(Equal Opportunity/Equality of Odds) requires equal True Positive Rate(TPR) across subgroups: $P(Y^{'} = 1|Y = 1, C = 0) = P(Y^{'} = 1|Y = 1, C = 1)$ where $Y^{'}$ is the model output.

Definition 4(Statistical Parity) requires positive predictions to be unaffected by the value of the protected attribute, regardless of true label $P(Y^{'} = 1|C = 0) = P(Y^{'} = 1|C = 1)$

Another common definition is individual fairness, which means that people who are similar with respect to the task should be given similar predictions or decisions.

Definition 5(Individual Fairness). IF $f : R^n \rightarrow R$ is a decision model, given appropriate distance functions - $d(.,.)$ on R^n (the domain of f) and $D(.,.)$ on R (the co-domain of f) - as well as threshold $\epsilon \geq 0$ and $\delta \geq 0$, the model is individually fair if, for any pair of inputs $x, x^{'}$ such that $d(x, x^{'}) \leq \epsilon$, we have $D(f(x), f(x')) \leq \delta$ [Dwork et al. (2011b)]

We can remove bias during data generation or during the downstream task. To address it from downstream tasks, there are two categories: 1. inprocessing methods [Donini et al. (2020) Zafar et al.

(2017) Zafar et al. (2019) Perrone et al. (2021)] that change the objective function optimized during training to include fairness constraints and 2. post-processing [Hardt et al. (2016) Woodworth et al. (2017)] methods that modify the outcome of the existing models by changing decision boundary. Some of these method such as [Jagielski et al. (2018)] preserve privacy and fairness. Our work focuses on removing bias during data generation process and stronger privacy assumption.

Disparate impact remover (DIR) [Feldman et al. (2015)] creates a distribution for each protected attribute. DIR then creates a median distribution over all distribution. The outcome variable is then projected by median distribution. This method is rank preserving for samples from each protected group.

Fair Max Entropy Distributions (FairMaxEnt) [Celis et al. (2020)] is a maximum entropy based approach to eliminate bias from data. The method uses priors and marginal vector to constrain on statistical rate [Feldman et al. (2015)] and representation rate [Hardt et al. (2016)] of the generated data. Then, it solves dual of the max-entrophy framework to optimize data distribution. This method is faster than methods such as [Calmon et al. (2017)]. However, this method does not consider individual fairness and only works for binary data.

3 Methodology

In our proposed method, we first use FairMaxEnt to create unbiased data and then use RON-Gauss to generate private data.(See Algorithm 1)

For step 2, we use discretization methods suggested in the original paper from [Celis et al. (2020)]. For step 1 and 3, we use a dictionary to store the mappings from binary data to original data. For step 4,we use the same preprocessing methods described in [Xu & Veeramachaneni (2018)] for categorical features. After data generation, we map binary features to original datasets by uniformly sampling from mappings. Since FairMaxEnt is a distribution over the domain rather than reweighting, some generated data does not have mapping to original datasets. We then sample data from the closest mappings. For step 7, we randomly sample matrix and use QR factorization [Trefethen & Bau (1997)] to get W. For step 8, we use normal distribution as generative model for unsupervised and continuous outcome variable. We can use Gaussian Mixture Model [Amendola et al. (2016)] to generate data for categorical outcome variable. Thus, our method is flexible since it can be used in regression, classification and unsupervised settings. For step 9, our study uses training data percentile as threshold to map continuous data back to binary/categorical features and we also restrict continuous data to range of the minimum and maximum from the original data.

Algorithm 1: Fast, Fair and Private Data Generation

```
Data: dataset $X \in R^{d \times n}$, dimensions $p < d$, and $\epsilon_{\mu}, \epsilon_{\sum} > 0$
Result: $x_1^{DP}$, .... $x_{n'}^{DP}$
Step 1: Create a dictionary D to map data X to binary data B;
Step 2: Obtain the fair processed data $B^{'}$ from FairMaxEnt with inputs B;
Step 3: Map $B^{'}$ back to X using nearest neighbor in D and uniform sampling;
Step 4: Pre-normalize: change categorical feature to one hot encoding plus white noise and
  $x_i := \frac{x_i}{\|x_i\|}$ for all $x_i \in X$;
Step 5: Center the data: $\bar{X} = X - \mu^{DP} 1^T$ where $\mu^{DP} = (\frac{1}{n} \sum_{i=1}^{n} x_i) + Z$ and $z_j(i)$ is drawn
  i.i.d from $Lap(2\sqrt{d}/n\epsilon\mu)$ ;
Step 6: Re-normalize: $\bar{x_i} = \bar{x_i}/\|\bar{x_i}\|$ for all $\bar{x_i} \in \bar{X}$.;
Step 7: Construct a RON projection matrix $W = [q_1, ...q_p] \in R^{d \times p}$ and project the data
  $X' = W^T \bar{X} \in R^{p \times n}$.;
Step 8: Draw synthetic data $x_i^{DP} \in R^p$ from $N(0, \sum^{DP})$ where $\sum^{DP} = (\frac{1}{n} X' X'^T) + Z$ and
  $z_j(i)$ is drawn i.i.d from $lap(2\sqrt{p}/n\epsilon_{\sum})$.;
Step 9: Project it back to original space using W and transform it to original data format.
```

The proposed method is fast. Suppose d is the dimension of data and n is the number of samples in the data and the number of data FFPDG generates. Step 1 runs in time nd. Step 2 runs in time polynomial in d, n and the bit complexity of the input parameters. (See Appendix D.3 in [Celis et al.

(2020)]). Step 3 runs in time n^2 as identifying nearest neighbor runs in time N at most. Step 4 to 6 runs in time nd. Steps 7 runs in time nd^2 as projected dimension is smaller than real dimension and QR decomposition runs in d^3. Steps 8 runs in time n^2d and step 9 runs in time n. Thus, the whole algorithm runs in time polynomial in d^3, n^2 and the bit complexity of the input parameters. If we want to generate data from extreme large data set, this method is faster and more computationally efficient than GAN based methods.

The proposed method is differentially private under some conditions. For step 1-4, data is generated by using KL divergence [Shlens (2014)] framework. The resulting posterior is robust in terms of KL-divergence to small changes in the data.(See Theorem 1 in Appendix). The generated samples are differential private under lipschitz continuity assumption.(See Theorem 2 in Appendix). Thus, till this step, our method is 2L-differentially private where L is mentioned in Appendix B Assumption 1. For step 5-8, it is approved by [Chanyaswad (2019)] that it is $(\epsilon_{\mu} + \epsilon_{\sum})$-differentially private. According to serial composition theorem from [Agrawal M. (2008)], if we can find L, our method is $(2L + \epsilon_{\mu} + \epsilon_{\sum})$-differentially private. We inject noise in different part of process such as sampling, data normalizing and data generation. Even if attackers know the whole generation process, it is almost impossible for them to estimate most parameters accurately.

Our method preserves group fairness. FairMaxEnt preserve both fairness constraints. While a few studies [Dwork & Mulligan (2013) Michael D Ekstrand & Mehrpouyan. (2018) Satya Kuppam & Miklau (2019)] argue that differentially private algorithms make unfair decisions, the datasets used in their study are not fair. We do not find evidence that shows that differentially private algorithm make unfair decision on fair dataset.

Our method preserves individual fairness with enough samples in regression settings . For step 1-4, we only sample data from previous distribution which preserves individual fairness. For step 5-7, after normalization and RON projection, we prove that the distance between different data point is getting closer. (See Theorem 3 in Appendix for proof) For step 8, we show that when outcome variable is generated by linear combination of feature space plus noise, sample generated from this model converge to data generation space as sample size goes to infinity. (See Theorem 4 in Appendix for proof). To summarize, our generated method only generate data that is closed to original data space.

4 METRIC AND EXPERIMENT

4.1 METRIC

AUCROC[Yoon et al. (2019)]: We train few models on generated data. Then, we test those models on real data and use area under the receiver operating characteristics curve(AUCROC) as our metric. If we saw high AUCROC on the real data for models that were trained on synthetic data, we can infer that synthetic data has captured the relationship between features and labels well. These synthetic data can be used to train models without ever seeing the real data. To have a stable result, we choose best AUCROC performance among models that include logistic regression, Random Forest[Breiman], Neural Network, Gaussian Naive Bayes[Rish (2001)], Gradient Boosting Classifier[Friedman], Bernoulli Naive Bayes, Decision Tree and Linear Discriminant Analysis.

DEO/DSP: $Y^{'}$ is the predicted value of outcome variable. Difference in equal opportunity DEO ($|P(Y^{'} = 1|Y = 1, C = 0) - P(Y^{'} = 1|Y = 1, C = 1)|$) and Difference in statistical parity DSP ($|P(Y^{'} = 1|C = 0) - P(Y^{'} = 1|C = 1)|$) (Note that DSP also called Demographics Parity Difference (DPD) [Zafar et al. (2017)]) [Perrone et al. (2021)] are two measures we use to evaluate group fairness of the generated result. We use logistic regression , Random Forest, Neural Network, Gaussian Naive Bayes, Gradient Boosting Classifier, Bernoulli Naive Bayes, Decision TreeQuinlan and Linear Discriminant Analysis and choose the average of them as default models to predict Y^{bar}. (For other metric, you can look at [Dwork et al. (2011b)])

LRD: We build a logistic regression classifier that learns to tell the synthetic data apart from the real data (LRD), which later on is evaluated using cross validation. The output of the metric is one minus the average AUCROC score obtained. The higher the score, the harder it is for the model to distinguish real data from synthetic data. We use this metric to understand if the generated data is robust against adversarial attacks such as membership inference attack.

4.2 HYPERPARAMETER TUNING

We set epsilon equals to 1 for differential privacy. We choose repair level for DIR equal to 1. We choose smooth factor equals to 0.1 and error parameter equal to 0 for FairMaxEnt. We choose σ equals to 2, clip coefficient equals to 0.1, micro batch size equals to 8, number of epochs equals to 500 and batch size equals to 64 for DPGAN. Number of teacher equals to 10, teacher and student iterations equal 5, number of moments equal to 100 for PATEGAN. For RON-Gauss, we set ϵ_μ to be 30 percent of ϵ and $\epsilon_\sum$ to be 70 percent of ϵ. All experiments are run on AWS with ml.p2.16xlarge.

4.3 EXPERIMENT

The Adult [Dua & Graff (2017)] dataset contains demographic information of individuals along with a binary label of whether their annual income is greater than 50k. In our analysis we include attributes race (white VS non white), sex, age and education years. We use gender as the protected attribute.

The COMPAS [Barenstein (2019)] dataset contains information on criminal defendants at the time of trial, along with post-trail instances of recidivism. We include attributes such as sex, race, priors count and charge degree as features. We use gender as the protected attribute.

We first study whether we should put fair processing first or private data generation first. Our finding is that if we put fair processing after private data generation, it becomes harder to decrease DSP and increase AUCROC. Although more experiments may be needed to see if this result can be generalized, we recommend that fair processing occurs before private data generation. Private data generation relies on adding noise in labeling or training process. If data has many outliers, this private generated data may exaggerate these outliers and generate more noisy data. Fair processing method such as FairMaxEnt can mitigate the influence of outliers by projecting feature space to binary and fitting a distribution.

AUCROC	DP WGAN	PATE GAN	RON GAUSS
FairMaxEnt First	0.62	0.75	0.76
FairMaxEnt Last	0.53	0.72	0.74

Table 1: Compare pre/post fair processing influence AUCROC

DSP	DP WGAN	PATE GAN	RON GAUSS
FairMaxEnt First	0.07	0.11	0.07
FairMaxEnt Last	0.22	0.16	0.40

Table 2: Compare pre/post fair processing influence DSP

Then, we compare our result with combinations of reweight[Kamiran & Calders (2012)]/DIR/FairMaxEnt as fair processing and DPWGAN/PATE-GAN/RON-Gauss as private data generation methods. Code we used: [bor (2019) Bellamy et al. (2018) vij (2020) Patki et al. (2016)]

We also compare the speed of our method with others. We choose the fastest DPWGAN and PATE-GAN based method. We then compare them with FFPDG on COMPAS and Adult dataset. We can see that FFPDG is much faster than GAN based methods.

Dataset	FairMaxEnt + DPWGAN	reweight + PATEGAN	FFPDG
COMPAS	36 seconds	29 minutes	1 seconds
Adult	47 seconds	29 minutes	1 seconds

Table 3: Compare pre/post fair processing influence DSP

Based on our experiment, our proposed method achieves good AUCROC while maintains the highest LRD and lowest DSP. (See table 3 and 4 in appendix for full results) We run the same experiment on COMPAS dataset[Barenstein (2019)] and see similar result but with relatively lower AUCROC. High AUCROC means data generated by our method has good predictability power on real data.

High LRD means that our generated data is hard to distinguish from real data. Low DSP shows that our method preserves group fairness. In general, FairMaxEnt performs better in DSP compare to all other fair processing methods. RON-Gauss performs better in LRD compare to DPWGAN and PATEGAN.

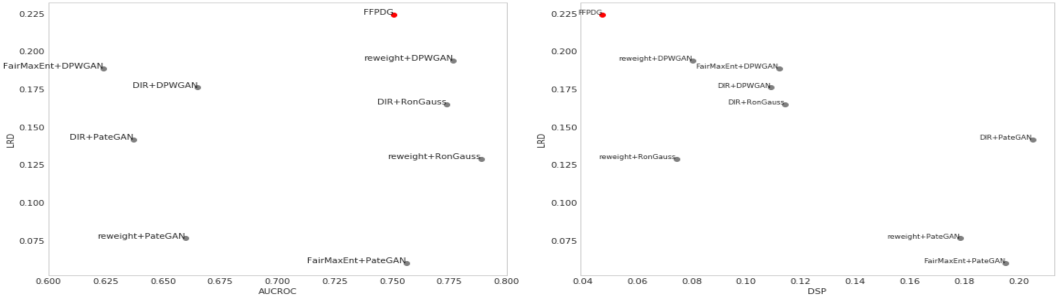

Figure 1: Comparison of different methods in Adult data set

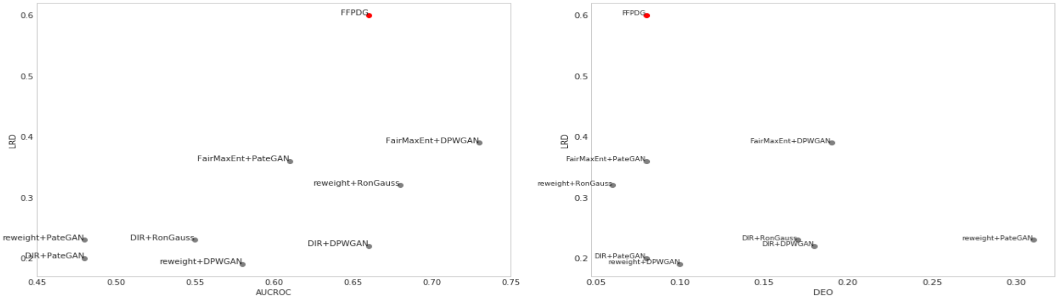

Figure 2: Comparison of different methods in COMPAS data set

5 Limitations and Conclusions

5.1 Limitations

Although, DFM shows that data can be projected to low dimensional Gaussian distribution. It becomes harder to find good projection with categorical data and increase of dimensions. We see the predictability on real data decreases as dimensions or categorical features increase. We may need to change step 8 in our proposed method for better data generation on high dimensional data.

LRD is very low for almost all private preserved generated data. We further study methods such as CTGAN[Xu et al. (2019)] and Copula GAN[Kamthe et al. (2021)] using Adult dataset to see if this result can be generalized for other data generation technique. The result shows that for the same procedure, the LRD for them are 0.6 and 0.79. However, if we use the same post processing such as capping by max and min and changing continous data back to binary/categorical features. The LRD drops to 0.19 and 0.15. To make our method scale, better postprocessing method for categorical feature is needed.

Our method does not inference well on dataset that is biased. Our method make data unbiased. If test data is hugely biased, distribution shift may occur between generated data and test data. Model trained on these generated data may not inference well on biased test data.

5.2 Conclusion

To conclude, we design a method to generate fair and unbiased data. The method of data generation has the following benefits: It is flexible and fast, the data created can produce models that perform well in real datasets, robust to membership inference attack and preserve individual/group fairness.

There are few directions of future study: (1) Make our algorithm scale to high dimensional datasets. (2) Find a better method to process categorical features. (3) Merge fair processing and private data generation into one step.

References

private-data-generation. `https://https://github.com/BorealisAI/private-data-generation`, 2019.

Fair-max-entropy-distributions. `https://https://github.com/vijaykeswani/Fair-Max-Entropy-Distributions`, 2020.

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016. doi: 10.1145/2976749.2978318. URL `http://dx.doi.org/10.1145/2976749.2978318`.

Sushant Agarwal. On the tradeoffs between privacy and fairness.

Duan Z. Li A Dwork C Agrawal M., Du D. Differential privacy: A survey of results., 2008.

Carlos Amendola, Jean-Charles Faugere, and Bernd Sturmfels. Moment varieties of gaussian mixtures. *Journal of Algebraic Statistics*, 7(1), Jul 2016. ISSN 1309-3452. doi: 10.18409/jas.v7i1.42. URL `http://dx.doi.org/10.18409/jas.v7i1.42`.

Matias Barenstein. Propublica's compas data revisited, 2019.

S Hardtm N Barocas and A Narayanan. *Fairness and Machine Learning*. 2018.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016.

L Breiman. Random forests. *Machine Learning 45*.

Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized data pre-processing for discrimination prevention, 2017.

L. Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1349–1359. PMLR, 13–18 Jul 2020. URL `http://proceedings.mlr.press/v119/celis20a.html`.

Liu C. Mittal P. Chanyaswad, T. Ron-gauss: Enhancing utility in non-interactive private data release. *Proceedings on Privacy Enhancing Technologies*, pp. 26–46, 2019.

Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks, 2021.

Frank McSherry llya Mironov Cynthia Dwork, Krishnaram Kenthapadi and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, 2006.

Christos Dimitrakakis, Blaine Nelson, Zuhe Zhang, Aikaterini Mitrokotsa, and Benjamin I. P. Rubinstein. Differential privacy for bayesian inference through posterior sampling. *Journal of Machine Learning Research*, 18(11):1–39, 2017. URL `http://jmlr.org/papers/v18/15-257.html`.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints, 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Nicolas Duczek, Matthias Kerzel, and Stefan Wermter. Continual learning from synthetic data for a humanoid exercise robot, 2021.

Cynthia Dwork and Christina Ilvento. Fairness under composition. *CoRR*, abs/1806.06122, 2018. URL http://arxiv.org/abs/1806.06122.

Cynthia Dwork and Deirdre K Mulligan. It's not privacy, and it's not fair. *Stan. L. Rev. Online*, 2013.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. 2014.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011a.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011b. URL http://arxiv.org/abs/1104.3913.

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. 2015.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. A moral framework for understanding of fair ml through economic models of equality of opportunity, 2018.

Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications, 2020.

Matthew Jagielski, Michael J. Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan R. Ullman. Differentially private fair learning. *CoRR*, abs/1812.02696, 2018. URL http://arxiv.org/abs/1812.02696.

Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models, 2020.

Timnit Gebru Joy Buolamwini. Gender shades: Intersectional accuracy disparities in commercial gender classification. *The 1st Conference on Fairness, Accountability and Transparency,*, pp. 81:77–91, 2018.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination, 2012.

Sanket Kamthe, Samuel Assefa, and Marc Deisenroth. Copula flows for synthetic data generation, 2021.

Michael Kearns, Aaron Roth, and Saeed Sharifi-Malvajerdi. Average individual fairness: Algorithms, generalization and experiments, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores, 2016.

Elizabeth Meckes. Projections of probability distributions: A measure-theoretic dvoretzky theorem. *Proceedings on Privacy Enhancing Technologies*, pp. 317–326, 2012.

Rezvan Joshaghani Michael D Ekstrand and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. *In Conference on Fairness, Accountability and Transparency*, pp. 35–47, 2018.

A. S. Syed Navaz, A. S. Syed Fiaz, C. Prabhadevi, V. Sangeetha, and S. Gopalakrishnan. Human resource management system, 2013.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data, 2017.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate, 2018.

N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, Oct 2016. doi: 10.1109/DSAA.2016.49.

Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. Fair bayesian optimization, 2021.

J.Ross Quinlan. Induction of decision trees. *Mach Learn*.

Dhamma Kimpara Jamie Morgenstern Rachel Cummings, Varun Gupta. On the compatibility of privacy and fairness. *UMAP*, pp. 309–315, 2019. URL https://doi.org/10.1145/3314183.3323847.

Congzheng Song Reza Shokri, Marco Stronati and Vitaly Shmatikov. Membership inference attacks against machine learning models. *In Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 3–18, 2017.

Dr. Rich, 2020. URL https://www.kaggle.com/rhuebner/human-resources-data-set.

I. Rish. An empirical study of the naive bayes classifier. 2001.

Donald Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*.

Mahmoud Mahfouz Tucker Balch Prashant Reddy Samuel Assefa, Danial Dervovic and journal = Neural Information Processing Systems year = 2019 Manuela Veloso1, title = Generating synthetic data in finance: opportunities, challenges and pitfalls.

David Pujol Michael Hay Ashwin Machanavajjhala Satya Kuppam, Ryan McKenna and Gerome Miklau. Fair decision making using privacy-protected data. *CoRR*, 2019.

Jonathon Shlens. Notes on kullback-leibler divergence and likelihood, 2014.

Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997. ISBN 0898713617.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors, 2017.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network, 2018.

Lei Xu and Kalyan Veeramachaneni. Synthesizing tabular data using generative adversarial networks, 2018.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. *CoRR*, abs/1907.00503, 2019. URL http://arxiv.org/abs/1907.00503.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. URL http://jmlr.org/papers/v20/18-262.html.

A Tables

A.1 Results from COMPAS

	Metric	Original Data	DPWGAN	PATEGAN	RON GAUSS
No Preprocessing	AUCROC	0.84	0.52	0.75	0.71
No Preprocessing	DEO	0.39	0.24	0.16	0.51
No Preprocessing	DSP	0.13	0.12	0.08	0.21
No Preprocessing	LRD	0.99	0.16	0.12	0.18
REWEIGHT	AUCROC	0.83	**0.78**	0.65	**0.79**
REWEIGHT	DEO	0.41	0.20	**0.01**	0.16
REWEIGHT	DSP	0.13	0.08	0.18	0.07
REWEIGHT	LRD	0.99	0.19	0.08	0.13
DIR	AUCROC	0.83	0.67	0.64	**0.77**
DIR	DEO	0.41	0.02	**0.01**	0.32
DIR	DSP	0.13	0.11	0.20	0.11
DIR	LRD	0.99	0.18	0.14	0.16
FairMaxEnt	AUCROC	0.81	0.62	0.75	**0.76**
FairMaxEnt	DEO	0.06	0.07	0.12	0.07
FairMaxEnt	DSP	0.12	0.07	0.20	**0.05**
FairMaxEnt	LRD	0.82	0.19	0.06	**0.22**

Table 4: Model performance with different metric for combination of fair preprocessing and private data generation for Adult data

	Metric	Original Data	DPWGAN	PATEGAN	RON GAUSS
No Preprocessing	AUCROC	0.74	0.63	0.73	0.66
No Preprocessing	DEO	0.38	0.09	0.06	0.08
No Preprocessing	DSP	0.01	0.04	0.04	0.05
No Preprocessing	LRD	0.99	0.33	0.33	0.24
REWEIGHT	AUCROC	0.73	0.58	0.48	0.68
REWEIGHT	DEO	0.38	0.10	0.31	0.06
REWEIGHT	DSP	0.01	0.07	0.05	0.02
REWEIGHT	LRD	0.98	0.19	0.23	0.32
DIR	AUCROC	0.73	0.66	0.48	0.55
DIR	DEO	0.36	0.18	0.08	0.17
DIR	DSP	0.01	0.06	0.04	0.04
DIR	LRD	0.89	0.22	0.20	0.23
FairMaxEnt	AUCROC	0.73	0.73	0.61	0.66
FairMaxEnt	DEO	0.06	0.19	0.08	0.08
FairMaxEnt	DSP	0.02	0.02	0.05	0.04
FairMaxEnt	LRD	0.68	0.39	0.36	**0.60**

Table 5: Model performance with different metric for combination of fair preprocessing and private data generation for COMPAS data

B THEOREM

Assumption 1 (Individual Bias) $f : R^n \rightarrow R$ said to be individually biased if there exists a pair of valid inputs x and $x^{'}$, with $|f(x) - f(x^{'})| > \delta$, such that $|x_i - x_i^{'}| \leq \epsilon_j$for all $i \in S_j$, and for all $j = 1, ..., t$. Such a pair $(x, x^{'})$ is called an individual bias instance of the model f. [John et al. (2020) Kleinberg et al. (2016)]

Assumption 1 (Lipschitz continuity) Let $f(x, \theta) = \ln p_\theta(x)$ be the log probability of x under θ. $\rho : S \times S \rightarrow R_+$ is the pesudo distance metric. The Lipschitz constant for a parameter value θ is $l(\theta) = inf\{\mu : |f(x, \theta) - f(y, \theta)| \leq \mu\rho(x, y) \forall x, y \in S\}$. We assume there exist some $L < \infty$ such that:

$$l(\theta) \leq L, \theta \in \Theta$$

[Dimitrakakis et al. (2017)]

Theorem 1 When ξ is a prior distribution on Θ and $\xi(\cdot|x)$ and $\xi(\cdot|y)$ are the respective posterior distribution for data sets $x, y \in S$, under a peseudi-metric ρ and $L > 0$ satisfying Assumption 1,

$$D(\xi(\cdot|x)\|\xi(\cdot|y)) \leq 2L\rho(x, y)$$

[Dimitrakakis et al. (2017)]

Theorem 2 Under a pseudo-metric ρ and $L > 0$ satisfying Assumption 1, for all $x, y \in S$, $B \in \sigma_\theta$:

$$\xi(B|x) \leq exp\{2L\rho(x, y)\}\xi(B|y)$$

(This means posterior ξ is $(2L, 0)$-differentially private under pseudo-metric ρ) [Dimitrakakis et al. (2017)]

Theorem 3 for $\forall x_1, x_2 \in D$, if $z_i = z_j$, then $\|f(x_i) - f(x_j)\|_F < \|x_i - x_j\|_F$ where z_i, z_j is noise drawn from laplacian noise and f represent step 5-7. W is ron projection and μ is the average from normalization.

Proof: Since we assume $z_i = z_j$, $u_i = \frac{1}{n}\sum_{i=1}^{n}(x_i) + z_i = \frac{1}{n}\sum_{i=1}^{n}(x_i) + z_j = u_j$

$$\|f(x_i) - f(x_j)\|_F = \|W^T(x_i - \mu_i) - W^T(x_j - \mu_j)\|_F$$

$$= \|W^T(x_i - x_j)\|_F$$

$$= \sqrt{tr((x_i - x_j)^T W W^T (x_i - x_j))}$$

$$= \sqrt{tr((x_i - x_j)^T W W^T W W^T (x_i - x_j))}$$

[This is because $W^T W = I$]

$$= \|W W^T (x_i - x_j)\|_F$$

[Let $P = W W^T$]

$$= \|P(x_i - x_j)\|_F$$

$$=< P(x_i - x_j), P(x_i - x_j) >_F$$

$$=< P(x_i - x_j), (x_i - x_j) >_F$$

[Consider $x' = x - Px$, $< Px, x >_F =< Px, x' + Px >=< Px, Px >_F + < Px, x' >_F =< Px, Px >_F$ since $Px \perp x'$, $< Px, x' >_F = 0$]

On the other hand, we have: $\|P(x_i - x_j)\|_F^2 =< P(x_i - x_j), (x_i - x_j) >_F^2 \leq \|P(x_i - x_j)\|_F \|(x_i - x_j)\|_F$ because of Cauchy-Schwarz inequality. This means $\|P(x_i - x_j)\|_F < \|(x_i - x_j)\|_F$ Thus we have $\|f(x_i) - f(x_j)\|_F = \|P(x_i - x_j)\|_F < \|(x_i - x_j)\|_F$

Theorem 4 Suppose $Y = X\beta + \epsilon$, in regression settings, $\begin{bmatrix} x_i \\ y_i \end{bmatrix} \in R^{p+1}$ by drawing samples from $N(0, \begin{bmatrix} \frac{1}{n}(XX^T + Z_x) & \frac{1}{n}X^T Y \\ \frac{1}{n}Y^T X & \frac{1}{n}(YY^T + z_y) \end{bmatrix})$. $\|x_i\| = 1$ For simplicity, we assume Z_x is positive semidefinite diagnol matrix. Any none zero element in Z iid drawn from $Lap((2\sqrt{p} + 4a\sqrt{p} +$

$a^2)/\sqrt{n}\epsilon_{\sum})$. Then, the variance of y is less than $\frac{1}{n}\epsilon^T\epsilon + \frac{1}{n}\beta^{-1}Z_X\beta + z_Y$ and the distance between y and $X\beta$ is less than $z_{max}\beta^T(XX^T)^{-1}$ where z_{max} is the largest absolute value in diagonal matrix Z

Proof:

$$Avg(y|x) = \frac{1}{n}Y^TX(\frac{1}{n}XX^T + Z_x)^{-1}x$$

$[(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}]$

$$= Y^TX((XX^T)^{-1} - (XX^T)^{-1}Z_x(XX^T + Z_x)^{-1})x$$
$$\leq Y^TX((XX^T)^{-1} - (XX^T)^{-1}Z_x(XX^T)^{-1})x$$
$$= Y^TX(XX^T)^{-1}x - Y^TX(XX^T)^{-1}Z_x(XX^T)^{-1}x$$

$[\beta = (X^TX)^{-1}X^TY$ and $(X^TX)^{-1T} = (X^TX)^{-1}]$

$$= \beta^Tx - \beta^TZ_x(XX^T)^{-1}x$$

Since $\|x_i\| = 1$ and Z_x is a diagonal matrix , $Avg(y|x) - \beta^Tx \leq z_{max}\beta^T(XX^T)^{-1}$

$$Var(y|x) = \frac{1}{n}(YY^T - Y^TX(X^TX + Z_X)^{-1}(X^TY)) + z_Y$$

[if (x, y) is multivariate gaussian, where mean is 0 and variance is $\begin{bmatrix}\sum_{11} & \sum_{12}\\ \sum_{21} & \sum_{22}\end{bmatrix}$, then $Var(y|x) = \sum_{22} - \sum_{21}\sum_{11}^{-1}\sum_{12}$]

$$(X^TX + Z_x)^{-1} = (X^TX)^{-1} - (X^TX)^{-1}Z_x(X^TX + Z_x)^{-1}$$
$$\geq (X^TX)^{-1} - (X^TX)^{-1}Z_x(X^TX)^{-1}$$

$$\frac{1}{n}(Y^TY - Y^TX(XX^T + Z_X)^{-1}X^TY) + z_Y$$
$$\leq \frac{1}{n}(Y^TY - Y^TX((XX^T)^{-1} - (XX^T)^{-1}Z_x(XX^T)^{-1})X^TY + z_y$$

$[H = X(X^TX)^{-1}X^T$ and H is symmetric and $\beta = (X^TX)^{-1}X^TY]$

$$= \frac{1}{n}(Y^TY - Y^THY + \beta^{-1}Z_X\beta) + z_Y$$

[Since $H * H = H$, we have $I - H = I - HH = I - H - H + HH = (I-H)(I-H)$ since I - H is symmetric, $I - H = (I-H)^T(I-H)$]

$$= \frac{1}{n}(Y^T(I-H)^T(I-H)Y + \beta^{-1}Z_X\beta) + z_Y$$
$$= \frac{1}{n}\epsilon^T\epsilon + \frac{1}{n}\beta^{-1}Z_X\beta + z_Y$$

Thus, as $n- > \infty$, z_{max} and z_y go to 0, both difference and variance goes to 0. $y|x$ convergence to β^Tx in probability.

For the same setting as Theorem 4, our proposed method is not individually biased under definition 1 as sample goes to infinity. y—x is sampled from $N(\beta^Tx + \epsilon, \sigma)$, where ϵ and σ go to 0. Thus, for any samples such that $|x_i - x_i^{'}| \leq \epsilon_j$, we have $|f(x) - f(x^{'})| \rightarrow \beta^T|x - x'| \leq \beta_{max}\sum_{j=1}^{t}\epsilon_j$ where β_{max} is the largest absolute value in β. We can find constant C, as sample size goes to infinitiy, $|f(x) - f(x^{'})| \leq C + \beta_{max}\sum_{j=1}^{t}\epsilon_j$. Thus, the difference of outcome variable is bounded. Our analysis is based on definition 1. It may not hold in other definitions of individual bias.[Dwork et al. (2011a) Dwork & Ilvento (2018) Kearns et al. (2019)]