The Name of the Title Is Hope

Chih-Cheng Rex Yuan
hello@rexyuan.com
Institute of Information Science, Academia Sinica
Taipei, Taiwan

Bow-Yaw Wang
bywang@iis.sinica.edu.tw
Institute of Information Science, Academia Sinica
Taipei, Taiwan

Abstract

abstract

ACM Reference Format:
Chih-Cheng Rex Yuan and Bow-Yaw Wang. 2024. The Name of the Title Is Hope. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

2 Related Work

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 Preliminaries

A *row* r_i is a lookup table or dictionary. A *database* $D = \{r_1, r_2, ...\}$ is a collection of rows. The set of all databases is denoted $\mathcal{D}$. The attributes of a D is $\mathcal{A} = \{A_1, A_2, ...\}$. The domain of A_i is Ω_i.

3.1 Fairness Measures

For fairness measures[11, 15], let Y to denote the ground truth of an outcome, let $\hat{Y}$ to denote the predicated result of an outcome, let S denote the protected attribute, and let ϵ denote some threshold. For non-binary prediction, such as a score, we use $\hat{V}$.

Fairness measures can be broadly categorized into independence, separation, and sufficiency, which are defined by conditional independence in Table 6.

Table 1: Fairness categories.

Category	Definition
Independence	$S \perp \hat{Y}$
Separation	$S \perp \hat{Y} \mid Y$
Sufficiency	$S \perp Y \mid \hat{Y}$

These categories can be expanded into forms of probability. For example, the definition of separation is expanded to

$$P[\hat{Y} = 1 | S = 1, Y = 1] = P[\hat{Y} = 1 | S \neq 1, Y = 1]$$
$$P[\hat{Y} = 1 | S = 1, Y = 0] = P[\hat{Y} = 1 | S \neq 1, Y = 0]$$

The definition can be relaxed. Its relaxation, for some parameter ϵ, is

$$|P[\hat{Y} = 1 | S = 1, Y = 1] - P[\hat{Y} = 1 | S \neq 1, Y = 1]| \leq \epsilon$$
$$|P[\hat{Y} = 1 | S = 1, Y = 0] - P[\hat{Y} = 1 | S \neq 1, Y = 0]| \leq \epsilon$$

which is also the definition of a fairness measure called equalized odds.

We consider in this work various fairness measures listed in Table 2.

3.2 Differential Privacy

A *randomized mechanism* is a randomized algorithm $M : \mathcal{D} \to R$ that takes a database and, after introducing noise, outputs some results.

Definition 3.1 (Gaussian Mechanism[2]). Let $f : \mathcal{D} \to \mathbb{R}^p$ be a function that takes a database and outputs a vector. The Gaussian Mechanism M adds i.i.d. Gaussian noise with scale σ to each of the p outputs:

$$M(D) = f(D) + \mathcal{N}(0, \sigma^2 \mathbb{I})$$

Definition 3.2 (Rényi Differential Privacy (RDP)). A randomized mechanism M satisfies (α, γ)-RDP for $\alpha \geq 1$ and $\gamma \geq 1$ if, for all databases D_1, D_2 that differ in exactly one row, we have

$$D_\alpha(M(D_1) \| M(D_2)) \leq \gamma$$

where D_α is the Rényi divergence[14] of order α.

Theorem 3.3 (RDP of the Gaussian Mechanism[3, 9]). *The Gaussian Mechanism satisfies* $(\alpha, \alpha\frac{\Delta_f^2}{2\sigma^2})$*-RDP.*

3.3 Differentially Private Synthetic Data

Let $C \subseteq \mathcal{A}$ be a subset of attributes. Let $\Omega_C = \Pi_{i \in C} \Omega_i$. A *marginal*[1, 7] of C is a vector $\mu \in \mathbb{R}^{|\Omega_C|}$, indexed by domain element $t \in \Omega_C$, such that each entry is a count $\mu_t = \Sigma_{x \in D} \mathbb{1}[x_C = t]$ where $\mathbb{1}$ is the indicator function; that is, it is the vector of the count of each possible element.

Let $M_C(D)$ be the function that computes the marginal of C on D, i.e., $\mu = M_C(D)$. We call marginals of $|C| = n$ attributes n-way marginals.

The task of differentially private synthetic data is, given a database D, adding some noise such that it satisfies differential privacy guarantee and outputting another database D', such that the L_1 errors between some selected marginals $C_1, C_2, ...$ of D and D' is small; that is, their marginals $(M_{C_1}(D), M_{C_1}(D')), (M_{C_2}(D), M_{C_2}(D')), ...$ are similar.

For example, suppose we have a database with attributes sex and race. The 2-way marginals of the original database and the synthetic database are shown in Table 3. The marginals of the synthetic data is supposed to be similar to that of the original database.

4 Auditing Framework

Our auditing framework is tripartite. It consists of three parties: the data provider, the model maker, and the third-party auditor.

The data provider is responsible for supplying the raw datasets which should originate from trustworthy sources, such as government agencies like a census bureau.

The model maker develops AI models. These are AI companies or research labs specialized in training and optimizing AI models.

The third-party auditor acts as an evaluator, using our framework to audit the AI models for fairness issues by combining both the datasets and the models. These may be investigative journalists or regulatory bodies.

In the framework of our previous work[15], after obtaining real data from the data provider, the 3rd party auditor holds onto the real data for performing fairness audits, and it supposedly retains it indefinitely for the possibility of any future audits.

However, this practice introduces security concerns. It creates a vulnerability to data security threats. A breach at the auditor's end could result in compromises of individuals' privacy.

Moreover, the storage of the datasets also raises privacy concerns. Holding large amounts of sensitive data for an extended period opens the door to the risk of misuse. The auditor may misuse the data for unauthorized purposes.

Thus, we introduce a new framework where the auditor generates synthetic data based on real data upon retrieval of the real data, and then holds onto the synthetic data and discards the real data, preventing further privacy breaches.

5 Methodology

We employed the tools of the winner of the 2018 NIST Differential Privacy Synthetic Data Challenge competition[10] by Ryan McKenna[5–8, 13] and the fairness checker tool from our previous research[15].

This research is implemented in Python Jupyter notebooks and is publicly available.

Table 2: Fairness measures.

Category	Fairness Measure	Definition
Independence	Disparate Impact	$\frac{P[\hat{Y}=1 \mid S\neq 1]}{P[\hat{Y}=1 \mid S=1]} \geq 1-\epsilon$
	Demographic Parity	$\lvert P[\hat{Y}=1 \mid S=1] - P[\hat{Y}=1 \mid S\neq 1]\rvert \leq \epsilon$
	Conditional Statistical Parity	$\lvert P[\hat{Y}=1 \mid S=1, L=l] - P[\hat{Y}=1 \mid S\neq 1, L=l]\rvert \leq \epsilon$
	Mean Difference	$\lvert E[\hat{Y} \mid S=1] - E[\hat{Y} \mid S\neq 1]\rvert \leq \epsilon$
Separation	Equalized Odds	$\lvert P[\hat{Y}=1 \mid S=1, Y=0] - P[\hat{Y}=1 \mid S\neq 1, Y=0]\rvert \leq \epsilon$ $\lvert P[\hat{Y}=1 \mid S=1, Y=1] - P[\hat{Y}=1 \mid S\neq 1, Y=1]\rvert \leq \epsilon$
	Equal Opportunity	$\lvert P[\hat{Y}=1 \mid S=1, Y=1] - P[\hat{Y}=1 \mid S\neq 1, Y=1]\rvert \leq \epsilon$
	Predictive Equality	$\lvert P[\hat{Y}=1 \mid S=1, Y=0] - P[\hat{Y}=1 \mid S\neq 1, Y=0]\rvert \leq \epsilon$
Sufficiency	Conditional Use Accuracy Equality	$\lvert P[Y=1 \mid S=1, \hat{Y}=1] - P[Y=1 \mid S\neq 1, \hat{Y}=1]\rvert \leq \epsilon$ $\lvert P[Y=0 \mid S=1, \hat{Y}=0] - P[Y=0 \mid S\neq 1, \hat{Y}=0]\rvert \leq \epsilon$
	Predictive Parity	$\lvert P[Y=1 \mid S=1, \hat{Y}=1] - P[Y=1 \mid S\neq 1, \hat{Y}=1]\rvert \leq \epsilon$
	Equal Calibration	$\lvert P[Y=1 \mid S=1, \hat{V}=v] - P[Y=1 \mid S\neq 1, \hat{V}=v]\rvert \leq \epsilon$
N/A	Overall Accuracy Equality	$\lvert P[Y=\hat{Y} \mid S=1] - P[Y=\hat{Y} \mid S\neq 1]\rvert \leq \epsilon$
	Positive Balance	$\lvert E[\hat{V} \mid Y=1, S=1] - E[\hat{V} \mid Y=1, S\neq 1]\rvert \leq \epsilon$
	Negative Balance	$\lvert E[\hat{V} \mid Y=0, S=1] - E[\hat{V} \mid Y=0, S\neq 1]\rvert \leq \epsilon$

Table 3: Example marginals.

(a) Marginal of original data.

Attributes	Count
Male,White	24
Female,White	33
Male,Black	13
Female,Black	47

(b) Marginal of synthetic data.

Attributes	Count
Male,White	22
Female,White	35
Male,Black	10
Female,Black	46

5.1 Data Synthesis

The synthesis framework is three-fold, namely, select-measure-generate[6]. We first select the important marginals to preserve, measure them by adding differentially private noise, and then generate synthetic data.

Underneath the hood, the tool employs a Markov random field. The select step corresponds to marking cliques in a Markov random field, and the generate step corresponds to sampling from the fitted Markov random field.

By default, all 1-way marginals are selected to preserve the quantity of each attribute element. We can further preserve correlations by adding n-way marginals. For example, if we want to preserve the relationship between sex and race, we may add the clique (sex,race).

In a perfect world where all correlation information is to be preserved, we may wish to make a completely connected graph. However, this was found to be intractable as the complexity of the problem would skyrocket.

To circumvent the complexity explosion, instead, Ryan McKenna devised a technique where the mutual information of all the database attribute pairs is calculated, and then a maximum spanning tree algorithm was run with edge weights being the mutual information to obtain a skeleton spanning-tree-shaped Markov random field.

For the competition, Ryan McKenna further manually added certain cliques based on his investigation of the competition dataset. For example, he manually added the clique (sex,city,income). In addition, he would add some edges based on some sophisticated heuristics tailored to that particular dataset.

We developed an alternative heuristic for a general-purpose workflow. From the definition of mutual information, we can obtain that they are bounded by the pair's respective Shannon entropy. Using this property, we add additional edges with weights exceeding a fraction of the minimum of these upper bounds. As a rule of thumb, we have found setting the fraction to be 0.1 to be effective.

For the measure step, we followed the examples provided in the tool's repository. Gaussian noises are added to the selected marginals. Half of the privacy budget is spent on all 1-way marginals and the other half on the selected cliques. These marginals are then fed to the tool to fit the Markov random field. By [7], this procedure satisfies $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP for all $\alpha \geq 1$.

5.2 Fairness Checking

After synthesizing datasets, we used the fairness checker from [15] to compute their fairness measures. To test the viability of our method, we compare the metrics computed from the synthetic dataset against those of the original dataset. We used various datasets with fairness concerns mentioned in [11].

The fairness checker evaluates datasets based on multiple fairness metrics, such as demographic parity and equalized odds. These metrics are computed on some sensitive attributes, predicted outcomes, and ground truths. Examples of sensitive attributes are race and sex. Examples of predicted outcomes and ground truths are loan approval and criminal recidivism.

By comparing these measures between the synthetic and original datasets, we aim to ensure that the synthetic data preserves the fairness properties of the original data. The comparison process is three-fold. It goes as follows.

The dataset is first processed so it can be fed into the synthetic data generator. Some marginals are selected as described in the previous section, and the synthetic data generator model is fitted to the original data according to the marginals. Then the generator is run multiple times to obtain multiple sets of synthetic data.

Next, several AI models are extracted from various real life authors from Kaggle. They are finetuned to perform well on the original dataset. For one, a random forest model is finetuned by searching hyperparameters settings[4]. For another, a logistic regression model is finetuned by performing principal component analysis[12].

Several AI models and both the original dataset and the rounds of synthetic datasets are fed to the fairness checker. Sensitive attributes are identified based on manual examination with common sense or by referring to [11]. Then, all applicable fairness measures are computed using the checker for both the original and the synthetic.

Finally, we analyze the discrepancies between the fairness properties of the original and the synthetic by calculating the difference and the ratio of their perspective fairness measure values. The sum of the difference and the average of the ratio serve as a summary of the analysis.

6 Results

6.1 Adult Income Dataset

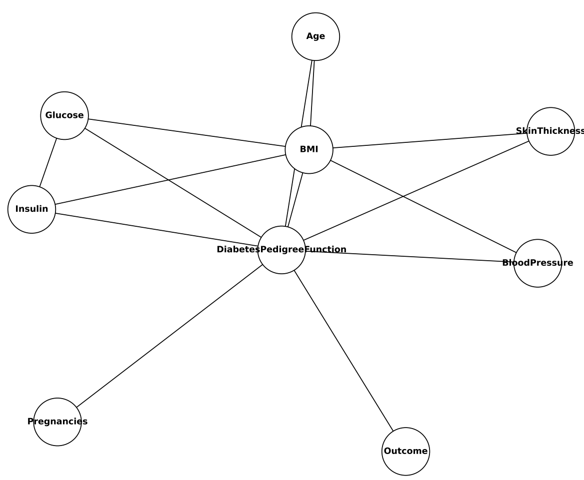

Figure 1: Markov random field for marginals of the adult dataset.

Table 4: Fairness measures experiment results of the adult dataset. Sum of difference is 0.259. Average of ratio is 0.952.

Measure	Original	Synthetic	Diff	Ratio
Demographic Parity	0.172	0.104	0.067	1.651
Accuracy Eqaulity	0.047	0.117	0.069	0.404
Equalized Odds 1	0.057	0.079	0.021	0.723
Equalized Odds 2	0.166	0.122	0.044	1.365
Accuracy Equality 1	0.100	0.132	0.031	0.759
Accuracy Equality 2	0.119	0.146	0.027	0.814

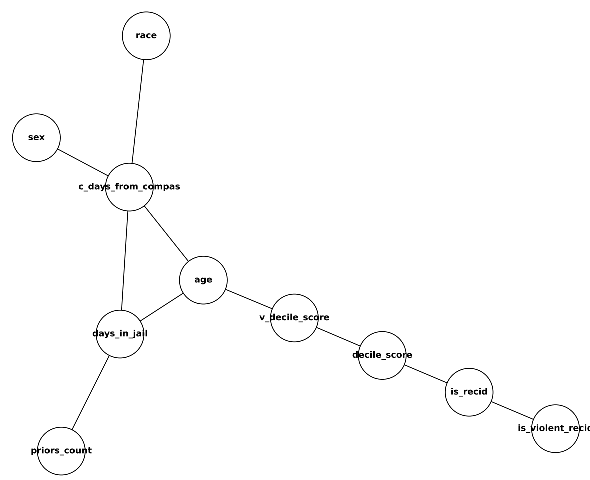

Figure 2: Markov random field for marginals of the COMPAS dataset.

Table 5: Fairness measures experiment results of the COMPAS dataset. Sum of difference is 0.419. Average of ratio is 1.745.

Measure	Original	Synthetic	Diff	Ratio
Demographic Parity	0.131	0.098	0.032	1.329
Accuracy Eqaulity	0.007	0.013	0.005	0.575
Equalized Odds 1	0.024	0.099	0.074	0.249
Equalized Odds 2	0.017	0.097	0.079	0.182
Accuracy Equality 1	0.170	0.082	0.088	2.082
Accuracy Equality 2	0.169	0.027	0.141	6.058

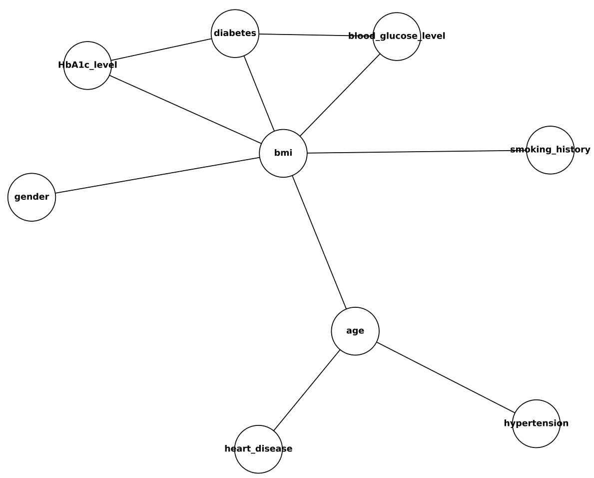

Figure 3: Markov random field for marginals of the diabetes dataset.

Table 6: Fairness measures experiment results of the COMPAS dataset. Sum of difference is 0.189. Average of ratio is 0.489.

Measure	Original	Synthetic	Diff	Ratio
Demographic Parity	0.013	0.015	0.002	0.867
Accuracy Eqaulity	0.007	0.009	0.002	0.782
Equalized Odds 1	0.000	0.003	0.003	0.000
Equalized Odds 2	0.008	0.106	0.097	0.081
Accuracy Equality 1	0.013	0.097	0.084	0.136
Accuracy Equality 2	0.021	0.020	0.001	1.068

6.2 COMPAS Dataset

6.3 Diabetes Dataset

7 Discussion

7.1 Accuracy

7.2 Impossibility

8 Conclusion

References

[1] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. 2007. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 273–282.
[2] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
[3] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. 2018. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 521–532.
[4] Ipbyrne. 2023. Income Prediction (84.369% Accuracy). https://www.kaggle.com/code/ipbyrne/income-prediction-84-369-accuracy Accessed: 2024-11-10.
[5] Ryan McKenna. 2023. private-pgm: A library for private probabilistic graphical models. https://github.com/ryan112358/private-pgm Accessed: 2024-11-10.
[6] Ryan McKenna and Terrance Liu. 2022. A Simple Recipe for Private Synthetic Data Generation. https://differentialprivacy.org/synth-data-1/ Accessed: 2024-11-10.
[7] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. 2021. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *arXiv preprint arXiv:2108.04978* (2021).
[8] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. 2019. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*. PMLR, 4435–4444.
[9] Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 263–275.
[10] National Institute of Standards and Technology. 2018. 2018 Differential Privacy Synthetic Data Challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic Accessed: 2024-11-10.
[11] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.
[12] Prashant111. 2023. EDA, Logistic Regression, PCA. https://www.kaggle.com/code/prashant111/eda-logistic-regression-pca#Introduction-to-Principal-Component-Analysis-(PCA) Accessed: 2024-11-10.
[13] Jonathan Ullman. 2022. What is Synthetic Data? https://differentialprivacy.org/synth-data-0/ Accessed: 2024-11-10.
[14] Tim Van Erven and Peter Harremos. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory* 60, 7 (2014), 3797–3820.
[15] Chih-Cheng Rex Yuan and Bow-Yaw Wang. 2024. Ensuring Fairness with Transparent Auditing of Quantitative Bias in AI Systems. In *2024 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*. IEEE, 25–32.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009