Quantitative Auditing of Fairness Measures with Differentially Private Synthetic Data

ANONYMOUS AUTHOR(S)

abstract

ACM Reference Format:
Anonymous Author(s). 2024. Quantitative Auditing of Fairness Measures with Differentially Private Synthetic Data. 1, 1 (December 2024), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

2 Related Work

Manuscript submitted to ACM

3 Preliminaries

Let $\prod_i x_i$ denotes the Cartesian product of x_is.

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.

Let $X \in \mathcal{X}$ be a discrete random variable and $p(x) := \Pr[X = x]$. Its *marginal Shannon entropy* is $H(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x)$. It quantifies the level of uncertainty of X. Let $Y \in \mathcal{Y}$ be another discrete random variable. Their *joint Shannon entropy* is $H(X,Y) := -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$.

Let X, Y be discrete random variables. Their *mutual information* is $I(X;Y) := H(X) + H(Y) - H(X,Y)$. It quantifies the level of dependence between X and Y. By definition[23], the joint entropy is greater than the marginal entropies $H(X,Y) \geq \max(H(X), H(Y))$. Hence, we have an upper bound of mutual information $I(X;Y) \leq \min(H(X), H(Y))$.

Let $\{X_i\}$ be a set of discrete random variables indexed by a graph $G = (V, E)$, where V represents the random variables X_is and E represents dependencies between these random variables. A *Markov random field* is a probability distribution over X_is, such that each random variable X_i, given its neighborhood in G, is conditionally independent of all other variables. Since edges represent dependencies, cliques in a Markov random field represent groups of variables that are all mutually dependent. As a machine learning model, there has been much development in the fitting and inference of Markov random fields[15].

3.1 Fairness Measures

For fairness measures[20, 27], let Y denote the ground truth of an outcome, let $\hat{Y}$ 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 1. $X \perp Y | Z$ denotes the conditional independence between X and Y conditioning on Z.

Table 1. Fairness categories.

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

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.

Table 2. Fairness measures.

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

3.2 Rényi Differential Privacy

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

Let D_1, D_2 be two databases. They are *neighbors*, denoted $D_1 \sim D_2$, if they differ in exactly one row.

Definition 3.1 (Gaussian Mechanism[3]). Let $f : \mathcal{D} \rightarrow \mathbb{R}^p$ 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 $D_1 \sim D_2$, we have

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

where $D_\alpha(P\|Q)$ is the Rényi divergence[8, 25, 26] of order α between discrete probability distributions P and Q, defined on $\mathcal{X}$:

$$D_\alpha(P\|Q) := \frac{1}{\alpha-1}\log\sum_{x\in\mathcal{X}} P(x)^\alpha Q(x)^{1-\alpha}$$

Theorem 3.3 (RDP of the Gaussian Mechanism[4, 14]). *The Gaussian Mechanism satisfies $(\alpha, \alpha\frac{\Delta_f^2}{2\sigma^2})$-RDP, where Δ_f denotes the sensitivity[3] of f, which is defined as the maximum L^2-norm difference in the output of f*

$$\Delta_f := max_{D_1\sim D_2}\|f(D_1) - f(D_2)\|_2$$

3.3 Differentially Private Synthetic Data

Let $C \subseteq \mathcal{A}$ be a subset of attributes. Let $\Omega_C = \prod_{i \in C} \Omega_i$. Let x be a row and x_C denote the restriction of x to C. A *marginal*[1, 12] of C on database D is a lookup table $\mu_D : \Omega_C \to \mathbb{N}_0$ such that each entry is a count $\mu_D(t) = \Sigma_{x \in D} \delta_{t,x_C}$ where δ is the Kronecker function; that is, it is the vector of the counts of each possible domain element. We call marginals of $|C| = n$ attributes n-way marginals.

The task of differentially private synthetic data[9, 11, 18, 24] is, given a database D, adding some noise to marginals of D such that it satisfies some differential privacy guarantees and outputting another database D', such that the L^1-norm errors between marginals of D and D' is small; that is, their marginals $\mu_D, \mu_{D'}$ 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.

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

4 Motivation

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 tool 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[27], 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 raises both security and privacy concerns.

For security, it creates a point of vulnerability of unauthorized access, and the auditor is now a target of data security attacks. The auditor may not have the necessary resources to defend against these threats. A breach at the auditor's end could result in compromises of individuals' sensitive information.

As for privacy, on the other hand, it introduces risks of information leakage. Storage of real data may inadvertently reveal sensitive information by data inference attacks. Indefinite retention exacerbates this issue because the longer the data is stored, the higher the chance of unwanted information disclosure.

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 all further data breaches. The third-party retains the ability to audit all incoming future models as needed.

5 Methodology

We employed the tools of the winner of the 2018 NIST Differential Privacy Synthetic Data Challenge competition[17] by [10, 12, 13] and the fairness checker tool from our previous research[27].

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

5.1 Data Synthesis

The synthesis framework is three-fold, namely, select-measure-generate[11]. 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 inference 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, [10, 12] 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, [10, 12] further manually added certain cliques based on his investigation of the competition dataset. For example, they manually added the clique (sex,city,income). In addition, they would add some edges based on some sophisticated heuristics tailored to that particular dataset. Meanwhile, his other apporach where the auditor does access the real dataset does not fit our framework.

We hence developed an alternative heuristic for the general-purpose workflow. As mentioned in Section 3, the mutual information of two random variables is 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 [12], this procedure satisfies $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP for all $\alpha \geq 1$.

5.2 Fairness Checking

After synthesizing the dataset, we used the fairness checker from [27] to compute the fairness measures of any incoming AI model.

The fairness checker is an open-sourced public domain Python package that computes various fairness measures, such as those mentioned in Table ??.

The checker is designed to be user-friendly and agnostic to the underlying AI model. It is also designed to be easily extensible to accommodate new fairness measures.

The checker simply iterates through the given database D and computes the results based on some given predicates on the rows r_is, and finally outputs the fairness measure values.

Protected groups S, predicted outcomes $\hat{Y}$, and ground truths Y are all formulated as these predicates. These are straightforward logical boolean expressions. Specifically, they are given as Python functions that output boolean values.

For example, if the sensitive attribute is sex and the protected group is female, the protected group predicate would be $S := r_i(\text{“sex”}) = \text{“Female”}$. This can be easily implemented in Python as a comparison function.

The interpretation of the resulting fairness measure values is dependent on the third-party auditors. The auditors may have different thresholds ϵ for different fairness measures or different AI models.

6 Experiments

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 [20].

We looked at several publicly available datasets, such as adult[2, 22], COMPAS[7, 19], and diabetes[6, 16]. The adult dataset comes from the 1994 census in the United States and contains about 30000 individuals. The COMPAS dataset comes from an investigative report by ProPublica of the COMPAS criminal recidivism assessment system and contains about 7000 individuals. The diabetes dataset comes from the hospital readmission data published in the 1994 AI in Medicine journal and contains about 100000 individuals.

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 Section 5, 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.

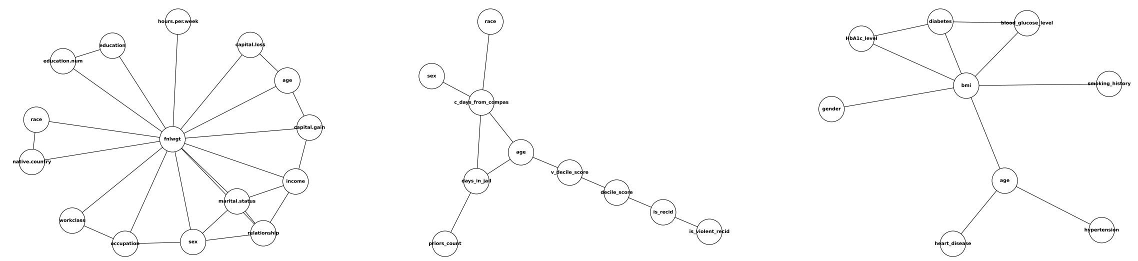
Next, several AI models are extracted from various real-life authors from Kaggle. They are finetuned to perform well on the original dataset. For one, a random forest model is finetuned by searching hyperparameters settings[5]. For another, a logistic regression model is finetuned by performing principal component analysis[21].

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 [20]. 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 of their perspective fairness measure values. The average of the differences serves as a summary of the analysis.

6.1 Adult Income Dataset

For the adult income dataset, the shape of the maximum spanning tree is very shallow, almost resembling a star; it has one internal node and all but one of the leaves have a depth of one. After introducing edges according to our heuristic,



(a) Markov random field for marginals of the adult dataset.

(b) Markov random field for marginals of the COMPAS dataset.

(c) Markov random field for marginals of the diabetes dataset.

Fig. 1. Markov random field for marginals of the experimented datasets.

we observed an increase in the pairwise edges of the leaves, forming many 3-cliques and two 4-cliques. The resulting graph is shown in Figure 1a.

Marginals based on this graph are then passed to the synthetic data generator for model fitting.

After generating ten rounds of synthetic data and passing them to the checker, their fairness measure values are averaged. Then, we compare them against the values of the original data. The results are shown in Table 4.

We observed that across all examined fairness measures, their difference all fall below 0.1. The average of their differences is 0.0431, which we consider quite satisfactory.

Table 4. Fairness measures experiment results of the adult dataset. Average of differences is 0.0431.

Measure	Original	Synthetic	Difference
Demographic Parity	0.172	0.104	0.067
Accuracy Eqaulity	0.047	0.117	0.069
Equalized Odds 1	0.057	0.079	0.021
Equalized Odds 2	0.166	0.122	0.044
Accuracy Equality 1	0.100	0.132	0.031
Accuracy Equality 2	0.119	0.146	0.027

6.2 COMPAS Dataset

For the COMPAS dataset, the initial spanning tree has a long tail, which is not surprising because, upon closer inspection, they all are related to the original COMPAS risk scores. The heuristic edge addition did not change the graph significantly. It only introduced one 3-clique triangle. The resulting graph is shown in Figure 1b.

Marginals of this graph are then too passed to the synthetic data generator for fitting.

We ran the same workflow as adult dataset for the COMPAS dataset. The comparison results are shown in Table 5.

The results showed an increase in error on the sufficiency measure values. In particular, one of the measures has an error difference as high as 0.141. The average of their differences is 0.069, which is worse than the adult dataset.

Table 5. Fairness measures experiment results of the COMPAS dataset. Average of differences is 0.069.

Measure	Original	Synthetic	Difference
Demographic Parity	0.131	0.098	0.032
Accuracy Eqaulity	0.007	0.013	0.005
Equalized Odds 1	0.024	0.099	0.074
Equalized Odds 2	0.017	0.097	0.079
Accuracy Equality 1	0.170	0.082	0.088
Accuracy Equality 2	0.169	0.027	0.141

6.3 Diabetes Dataset

The tree grown from the diabetes dataset did not appear to have any particular characteristics. The root of the tree is placed in the BMI value, which reasonably captures most information. The heuristic edge addition process introduced some 3-clique triangles. The resulting graph is shown in Figure 1c.

The same process is conducted to fit the synthetic data generator model.

The same workflow was done as on previous datasets. The comparison results are shown in Table 6.

There was also an increase of error on some sufficiency measure values. The average of their differences is 0.031.

Table 6. Fairness measures experiment results of the COMPAS dataset. Average of differences is 0.031.

Measure	Original	Synthetic	Difference
Demographic Parity	0.013	0.015	0.002
Accuracy Eqaulity	0.007	0.009	0.002
Equalized Odds 1	0.000	0.003	0.003
Equalized Odds 2	0.008	0.106	0.097
Accuracy Equality 1	0.013	0.097	0.084
Accuracy Equality 2	0.021	0.020	0.001

7 Evaluation

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] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[3] 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.

[4] 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.

[5] Ipbyrne. [n. d.]. Income Prediction (84.369% Accuracy). https://www.kaggle.com/code/ipbyrne/income-prediction-84-369-accuracy Accessed: 2024-11-10.

[6] Michael Kahn. [n. d.]. Diabetes. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T59G.

[7] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. ProPublica Compas Analysis—Data and Analysis for 'Machine Bias'. https://github.com/propublica/compas-analysis.

[8] Yingzhen Li and Richard E Turner. 2016. Rényi divergence variational inference. *Advances in neural information processing systems* 29 (2016).

[9] Ryan McKenna. 2021. Marginal-based Methods for Differentially Private Synthetic Data. https://www.youtube.com/watch?v=UKzh9QgNRxA. Accessed: 2024-11-26.

[10] Ryan McKenna. 2023. private-pgm: A library for private probabilistic graphical models. https://github.com/ryan112358/private-pgm Accessed: 2024-11-10.

[11] Ryan McKenna and Terrance Liu. 2022. A Simple Recipe for Private Synthetic Data Generation. https://differentialprivacy.org/synth-data-1/ Accessed: 2024-11-10.

[12] 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).

[13] 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.

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

[15] Kevin P Murphy. 2023. *Probabilistic machine learning: Advanced topics*. MIT press.

[16] Tz Mustafa. [n. d.]. Diabetes Prediction Dataset. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset Accessed: 2024-11-26.

[17] 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.

[18] Joseph Near and David Darais. 2021. Differentially Private Synthetic Data. https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data. Accessed: 2024-11-26.

[19] Dan Ofer. [n. d.]. COMPAS Dataset. https://www.kaggle.com/datasets/danofer/compass Accessed: 2024-11-26.

[20] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[21] Prashant111. [n. d.]. EDA, Logistic Regression, PCA. https://www.kaggle.com/code/prashant111/eda-logistic-regression-pca#Introduction-to-Principal-Component-Analysis-(PCA) Accessed: 2024-11-10.

[22] UC Irvine Machine Learning Repository. [n. d.]. Adult Census Income Dataset. https://www.kaggle.com/datasets/uciml/adult-census-income/data Accessed: 2024-11-26.

[23] Madhur Tulsiani and Shubhendu Trivedi. 2014. Information and Coding Theory Lecture 2. https://home.ttic.edu/~madhurt/courses/infotheory2014/l2.pdf Accessed: 2024-11-29.

[24] Jonathan Ullman. 2022. What is Synthetic Data? https://differentialprivacy.org/synth-data-0/ Accessed: 2024-11-10.

[25] Tim van Erven and Peter Harremoës. 2010. Rényi divergence and majorization. In *2010 IEEE International Symposium on Information Theory*. IEEE, 1335–1339.

[26] Tim Van Erven and Peter Harremos. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory* 60, 7 (2014), 3797–3820.

[27] 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.