Data Plagiarism Index: Characterizing the Privacy Risk of Data-Copying in Tabular Generative Models

Joshua Ward [*], Chi-Hua Wang [†], Guang Cheng[‡]

June 21, 2024

Abstract

The promise of tabular generative models is to produce realistic synthetic data that can be shared and safely used without dangerous leakage of information from the training set. In evaluating these models, a variety of methods have been proposed to measure the tendency to copy data from the training dataset when generating a sample. However, these methods suffer from either not considering data-copying from a privacy threat perspective, not being motivated by recent results in the data-copying literature or being difficult to make compatible with the high dimensional, mixed type nature of tabular data. This paper proposes a new similarity metric and Membership Inference Attack called Data Plagiarism Index (DPI) for tabular data. We show that DPI evaluates a new intuitive definition of data-copying and characterizes the corresponding privacy risk. We show that the data-copying identified by DPI poses both privacy and fairness threats to common, high performing architectures; underscoring the necessity for more sophisticated generative modeling techniques to mitigate this issue.

Key Words: Data-Copying, Trustworthy AI, Membership Inference Attack, Privacy Auditing, Synthetic Data, Tabular Generative Models.

[*]Ph.D. Candidate, Department of Statistics and Data Science, UCLA, CA, 90095. Email: joshuaward@ucla.edu

[†]Postdoctoral Scholar, Department of Statistics and Data Science, UCLA, CA, 90095. Email: chihuawang@ucla.edu

[‡]Professor, Department of Statistics and Data Science, UCLA, CA, 90095. Email: guangcheng@ucla.edu

1 Introduction

Data-copying is a particularly concerning manifestation of generative models overfitting [28, 6]. Historically, researchers have observed that deep learning models exhibit overfitting behaviors, sometimes generating unrealistic instances and at other times creating instances overly similar to samples in the training dataset. While the former can be lauded as "creative and imaginative," the latter poses a significant risk, threatening the confidentiality of training data. Indeed, previous studies [3] reveal that many prominent generative models attain high scores in terms of fidelity and diversity by memorizing or copying real samples, which compromises their effectiveness for privacy-sensitive applications. The issue of data-copying is paramount in the field of tabular generative models, especially as these models are often used in scenarios involving sensitive data and strict privacy protocols [47]. This paper focuses on the challenge of detecting and evaluating data-copying in tabular data generation, highlighting its critical role in enhancing the trust and accountability of these methods.

Various paradigms have been proposed to study data-copying in generative models, including hypothesis testing [28], non-parametric statistics [6], Membership Inference Attacks [41], and ad-hoc similarity metrics [34, 3, 38]. Each of these approaches contributes unique insights into the phenomenon of data-copying in tabular generative models but have some sort of drawback. For instance, while similarity metrics used in tabular synthetic data literature offer an understanding of the geometric relationship between training and generated data, they lack a threat model to assess the *privacy risks* associated with these geometries (see [15]). Conversely, Membership Inference Attacks provide valuable tools for understanding privacy risks but are often disconnected from the model overfitting literature and are challenging to apply to the complex, high-dimensional and mixed-type distributions that are common in tabular data applications. The disjoint nature of the data-copying literature fails to address common practitioner questions, such as:

'To what extent does a model copy training data, and how practically significant is this problem?'

Unfortunately, we show that the problem is substantial, as evidenced by Figure 1.

In this paper, we study data-copying in tabular generative model from the perspective of all three of these disconnected areas (Membership Inference Attacks, ad-hoc similarity metrics and Data-Copying measures). The goal is to craft a principled, interpretable, and privacy orientated metric that can be applied to tabular generative modelling. Here, we propose Data Plagiarism Index (DPI); a theory-motivated measure of local data-copying (Section 4.1). We argue that DPI differs from other competing metrics and show how it provides a novel geometric perspective on the data-copying behavior of generative models. With the proposed privacy metric DPI, we further develop a new type of Membership Inference Attack, named DPI MIA (Section 4.2), that bridges this data-copying measure with testable privacy risk. Our experiment (Section 5) find that DPI MIA identifies a tendency among high-fidelity tabular data generators to perform risky data-copying, raising privacy concerns.

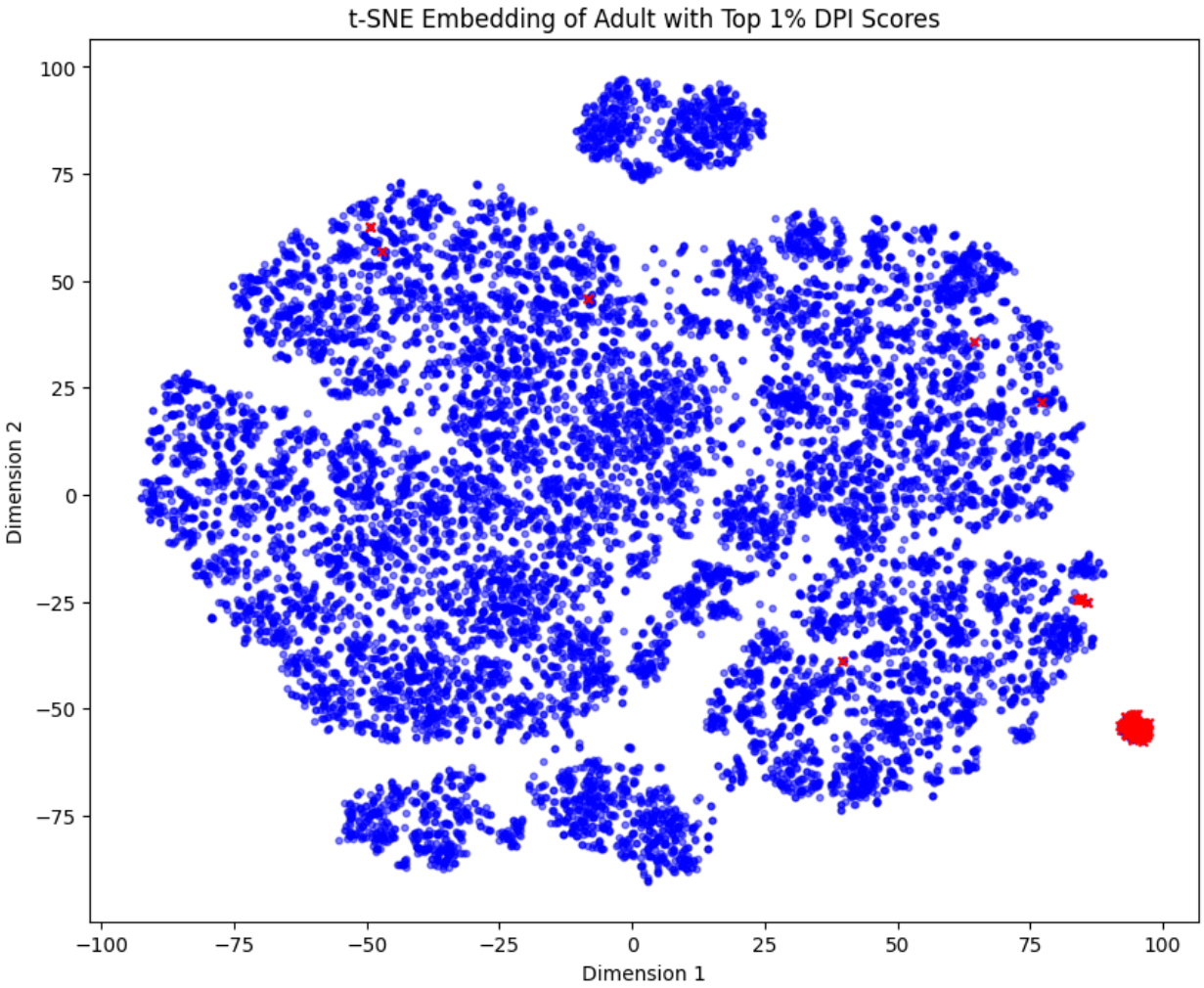

Figure 1: A t-SNE plot of Tab-DDPM's training data with corresponding top 1% DPI Scores in red on the Adult dataset. DPI identifies an outlier region in the bottom right corresponding to an extreme privileged class (married, white, middle aged, high capital gains, private industry, respondents making >50k in income). This provides evidence that Tab-DDPM copies the training data of outlier and privileged classes, creating serious fairness concerns for practitioners who use synthetic data in their downstream machine learning tasks. See Sec. 5.3 for detailed discussions.

Furthermore, our empirical evidence (Figure 1) reveals that tabular generative models **disproportionately favor certain privileged sub-populations** within the synthetic sample of the classic Adult dataset, raising serious fairness concerns when practitioners use synthetic data in downstream machine learning tasks. This troubling finding underscores that the proposed Data Plagiarism Index (DPI) is not merely a technical curiosity but a critical issue that can compromise both the privacy and fairness of tabular generative models.

Overall, we summarize our contribution to the Trustworthy Generative Modeling literature as follows:

1. We propose a novel privacy measure called **Data Plagiarism Index (DPI)** and a corresponding Membership Inference Attack called DPI MIA that measures the privacy risk of synthetic tabular data from data-copying.

2. We provide empirical evidence (Figure 4) that DPI identifies a positive correlation between tabular generative models' utility and privacy risks.

3. We provide empirical evidence (Figure 1) that reveals that Tab-DDPM [23] drastically copies the source training data of privileged sub-classes, creating a source of structural unfairness in the synthetic data.

4. We show (Figure 4, Table 1) that DPI MIA can identify a different kind of data-copying undetectable by existing Membership Inference Attacks with comparable threat performance, providing a new way to deploy privacy attacks to audit synthetic data privacy.

2 Related Work

In this section, we review three major literature to measure the overfitting and data-copying phenomenon of tabular generative models.

2.1 Measure Data-Copying in Generative Models

Generative models' data-copying in the literature is identified when synthetic data is excessively similar to training samples [28]. This is typically evaluated using a reference dataset, sampled from the same distribution as the training dataset. The common method involves an "appropriate distance function" to check whether synthetic data is closer to the training data or the reference data [34]. [28] for example focuses on identifying a suitable distance function to determine whether a synthetic dataset or a reference dataset is closer to the training dataset, thus testing if generative models exhibit data-copying. They devised a non-parametric method that divides the instance space into cells, tests each cell individually, and combines the results to understand the overall degree of data-copying. This concept inspired our more sophisticated data-copying measure (See Section 3.1). [6] introduces a more advanced data-copying test, capable of detecting data-copying behaviors not identified by [28]. Both of these studies however have limitations in high-dimensional complex distributions and also do not evaluate data-copying from a privacy perspective. DPI is more practical and accurately captures the degree of local data-copying, aiding in the assessment of synthetic data privacy risks (See Fig 1).

2.2 Similarity Metrics between Real and Synthetic Data

A variety of ad-hoc metrics have been proposed to evaluate the privacy of tabular synthetic data from a model overfitting perspective. Broadly speaking, these metrics focus on determining the level of similarity between the training and synthetic datasets ideally hoping to find that the generated data is in a 'Goldilocks' zone: not too similar to the training

data, but also not too dissimilar. These metrics will often be posed as a comparison between the training and a reference set and training and synthetic sets creating a sort of 'Null Distribution' in which to test privacy.

A common example of an often used similarity metric is Distance to Closest Record (DCR) [31, 27, 46, 51, 16, 25] where for each training point the distance to its closest neighbor in the synthetic dataset is compared with the distance of the closest neighbor in the reference dataset. Another is Identical Matching Score [1, 2, 27] which compares the proportion of identical records in the training and synthetic observations.

There are a variety of problems with this style of privacy evaluation. The first is that these similarity metrics do not actually guarantee nor are they evaluated by any idea of privacy protections against some sort of attack. For example, passing or failing a DCR or IMS test does not provide any information for how well a privacy attack may or may not do. Secondly, posing privacy as a hypothesis testing problem is a cardinal Statistics sin in that if the null hypothesis is: "the synthetic data are private" and the alternative is: "the synthetic data are not private", passing the test does not imply confirmation of the null, only that there is a failure to find sufficient evidence in which to reject the null [7]. However, while not being theoretically motivated, these metrics do provide an intuitive tool to study the geometric relationships between the training and synthetic datasets. Technically, DPI can be classified as a similarity metric and used in this way but these problems motivate DPI to adopt a Membership Inference Attack paradigm to overcome these dubious privacy interpretations.

2.3 Membership Inference Attacks for Generative Models

Membership Inference Attacks have traditionally been studied in a supervised learning context and only relatively recently have become applied to generative models. Here, the goal of the attack is to determine whether some sample of test data $x^* \in X^*$ was used in the training of the model based on some information about that model [37]. The scenario of what information is available for an attacker is called the threat model to which there are a variety of identified contexts for generative model MIAs. These include black box attacks [17, 18, 8] in which only generated synthetic data is available, white box attacks in which both synthetic data and the internals of a model are known, and calibrated (also called shadow) attacks in which both a synthetic dataset and then a reference dataset from the same or similar distribution as the training set are given [8, 17, 41]. Most attacks generally fall into the black box and calibrated paradigms as white box attacks are usually specific to the model architecture in question [36]. The value of MIAs is that they provide a tangible, practical scenario in which to study privacy risks. We will later frame DPI as an MIA in order to study how identifying local data-copying corresponds to privacy leakage.

3 Preliminaries

3.1 Formal definition of Data-Copying

We first introduce a formal definition of data-copying motivated by [28] to provide theoretical context for Data Plagiarism Index. Given a data distribution $\mathbb{P}$ and a region C, consider the following probability measure: $\mathbb{P}|_C(A) \equiv \mathbb{P}(A \cap C)/\mathbb{P}(C)$ for all measurable set A. Let R denote the distribution of reference data points and S denote the distribution of synthetic data points.

Given a neighborhood $D(x)$ of target data point x, define a one-dimensional distribution by $L(R) \equiv I(R \in D(x))$ and $L(S) \equiv I(S \in D(x))$ to denote separately the event of the reference data distribution and synthetic data distribution belonging to the neighborhood $D(x)$. By definition, $L(R) \sim R|_{D(x)}$ and $L(S) \sim S|_{D(x)}$

Define $\Delta(R, S) = P(B > A|B \sim L(S), A \sim L(P)$ to be the event that the synthetic data points have a greater probability in belonging to neighborhood $D(x)$ than the reference data points.

In the spirit of Definition 2.1 of [28], we define a generative model as **data-copying** a training data point x, if there exists a neighborhood $D(x)$ of x such that the synthetic data distribution S is systematically closer to the training data point x than the reference data distribution R, in the sense that

$$\Delta(R|_{D(x)}, S|_{D(x)}) < \frac{1}{2}$$

See Figure 2(a) as an example.

3.2 Distance to Closest Records (DCR)

The Distance to Closest Records (DCR) technique was developed to identify identical matches between synthetic and training data [31, 27]. However, it is essential to note that detecting identical matches and identifying data-copying are distinct tasks. While distance measures are appropriate for finding identical matches, they are not as effective for detecting data-copying, which is more appropriately viewed as a proportional measure. As discussed in Section 3.2 of [34], a DCR of zero signifies an identical match, but a non-zero DCR does not capture the extent of personal information disclosure. This inherent flaw renders DCR an unsuitable measure of synthetic data privacy, as it does not accurately assess privacy risk. Our proposed Data Plagiarism Index (DPI) provides a more effective metric for synthetic data privacy by quantifying privacy risk. We utilize DPI to construct a Membership Attack, whose results reveal the privacy risks of synthetic data. See Section 4.2.

3.3 Membership Inference Attacks as Privacy Auditors

Membership Inference Attacks (MIA) for synthetic data aim to determine if a given sample was a member of the original training set used to train a generative model. Consider a random variable X defined on the domain $\mathcal{X}$, following a probability distribution $p_X(X)$. Let D_{train} represent a training dataset consisting of independently sampled observations from $p_X(X)$.

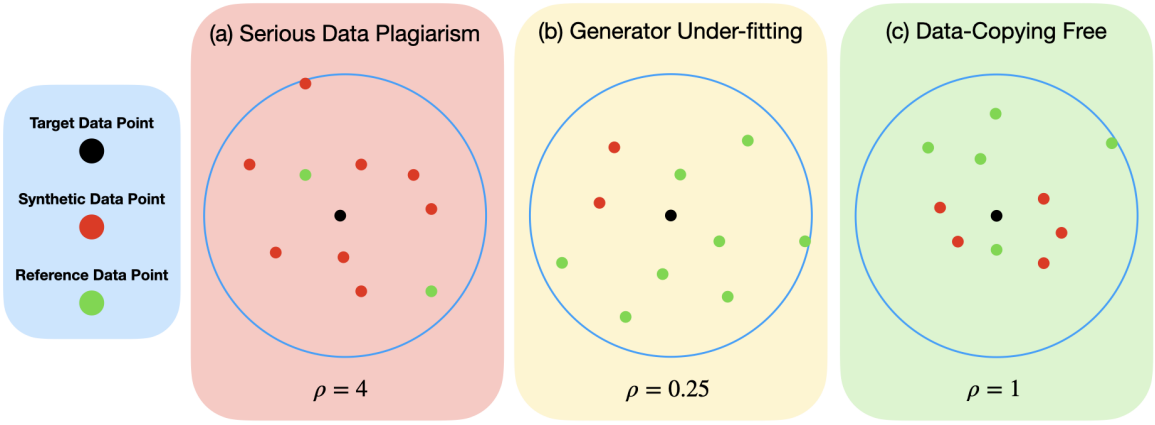

Figure 2: Data Plagiarism Index (DPI): We propose a novel privacy metric named Data Plagiarism Index (DPI). For each target data point (black point), we calculate the Data Plagiarism Index ρ by first construct a k-nearest neighborhood (blue circle) around the target data point on the space with reference data points (green points) and synthetic data points (red points). The Data Plagiarism Index ρ is defined simply as the ratio of number of synthetic data points to the number of reference data points. See Sec. 4.1 for whole details.

A generative model G is then trained on D_{train} and used to produce a synthetic dataset D_{syn}. An attacker model, $\mathcal{A} : X \to 0, 1$, has access to the synthetic dataset D_{syn}, a test data point x^* drawn from the same distribution as X, and depending on the threat model, potentially other information. The attacker's objective is to determine whether the test point x^* belongs to the original training set D_{train}. An ideal attacker would output $\mathcal{A}(x^*) = 1$ if $x^* \in D_{train}$, and 0 otherwise. By considering privacy in this framework, MIAs allow for the practical evaluation of the risk of synthetic data release, relative to an adversarial scenario [37].

In this paper, we study a special case of Membership Inference Attacks in which the attacker has access to an additional reference dataset D_{ref} sampled from same distribution as D_{train} but not used in the training of G. There are a variety of both practical and theoretical reasons for the inclusion of D_{ref} in the attack. First, D_{ref} represents a worse-case scenario in which the attacker has access to the most information possible in which to build an attack which makes this situation important to study from a conservative, privacy conscious data publisher's perspective. Second, it is a real scenario that particularly tabular-orientated data publishers face in that D_{ref} can be a leaked dataset, historic data, data from a competitor, or even a dataset built from publicly available data. Lastly, the inclusion of D_{ref} is theoretically motivated from the growing body of literature involved with testing for data-copying in a generative model in that in order most frameworks contextualize model miss-specification between D_{train} and D_{syn} as being relative to a holdout dataset [28, 6, 34].

4 Measuring Data-Copying Misbehavior

4.1 Data Plagiarism Index (DPI)

We propose a novel privacy metric called **Data Plagiarism Index** (`DPI`), as illustrated in Figure 2. For each training data point x, we calculate the Data Plagiarism Index ρ by first constructing a K-Nearest Neighborhood $D(x)$ around x on the space with reference data points and synthetic data points. The Data Plagiarism Index ρ is defined simply as the ratio of number of synthetic data points to the number of reference data points; that is

$$\rho(x) \equiv \frac{\sum_{x_i \in D(x)} I(x_i \in S)}{\sum_{x_i \in D(x)} I(x_i \in R)} \tag{1}$$

The DPI ranges from 0 to infinity, capturing a spectrum of data generation scenarios:

- **DPI = 0**: Indicates no synthetic data points in the neighborhood, suggesting potential *under-fitting* by the generative model.
- **DPI = 1**: Represents an equal number of synthetic and reference data points, implying no data plagiarism or under-fitting, signifying a balanced data generation process.
- **DPI > 1**: Highlights a greater presence of synthetic data points compared to reference data points, with increasing values pointing towards significant *data plagiarism*.

This comprehensive range allows the DPI to effectively capture various data generation scenarios, providing critical insights into the quality and privacy of synthetic datasets.

We provide 3 toy examples to help better understand how to interpret DPI values. In each example, consider the $K = 10$ nearest neighborhood for the target data point x, denoted by $D(x) = \{x_1, \cdots, x_n\}$. Each element x_i in the neighborhood may come from the synthetic dataset S or reference dataset R:

Toy Example 1 (Data Plagiarism, Figure 2.(a)). Say 8 elements come from the synthetic dataset and 2 elements come from the reference dataset. Then the Data Plagiarism Index $\rho(x) = 8/2 = 4$, which means the number of synthetic data points is 4 times that of the reference data points! This marks serious data-copying behavior from that generative model.

Toy Example 2 (Generator Under-fitting, Figure 2.(b)). In this situation, say 2 elements come from the synthetic dataset and 8 elements from the reference dataset. Then the Data Plagiarism Index $\rho(x) = 2/8 = 1/4$, which means the number of synthetic data points is 25% of that of the reference data points! This marks little data-copying behavior, but also suggests a potential underfitting issue around the target data point x.

Toy Example 3 (Data-Copying Free, Figure 2.(c)). Lastly, say there are 5 items from the synthetic dataset and 5 from the reference dataset. The Data Plagiarism Index is then $\rho(x) = 5/5 = 1$, meaining the number of synthetic data points is equal to the number of reference data points! This marks no data plagiarism and no underfitting behavior around the target data point x.

In summary, the Data Plagiarism Index offers a robust and intuitive metric for evaluating the balance between synthetic and reference data points within a specified neighborhood

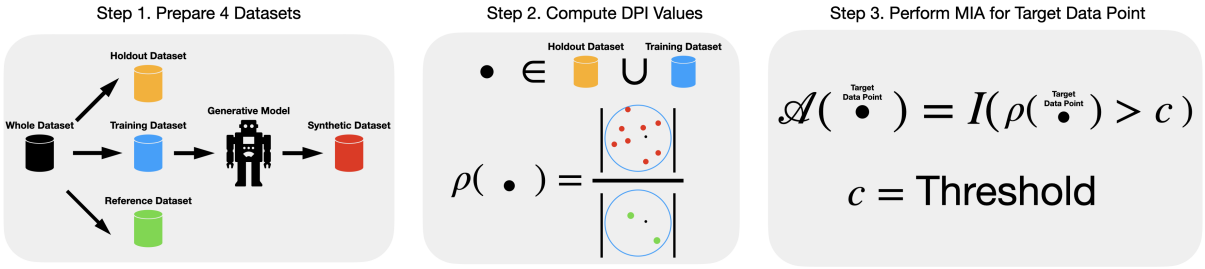

Figure 3: Data Plagiarism Index Membership Inference Attack (DPI MIA) See Sec. 4.2 for full details.

around a target data point. By calculating the ratio of synthetic to reference data points, the DPI effectively highlights critical instances of data plagiarism, under-fitting, or a balanced data generation process. This metric is invaluable for rigorously assessing the performance of generative models, ensuring data integrity, and proactively identifying potential issues in synthetic data generation.

4.2 Data Plagiarism Membership Inference Attack

While the Data Plagiarism Index defined at Equation (1) provides insight into the local behavior of generative models, on its own it does not characterize the actual *privacy risk* of data-copying in tabular generative models. We therefore show that a Membership Inference Attack can be derived from DPI which is designed to measure and then attack the local data-copying of generative models to evaluate the privacy risk.

Note that, in the literature of similarity metrics (Sec. 2.2), only the training dataset D_{train} is compared to the synthetic D_{syn} and reference D_{ref} sets. In the MIA paradigm however (Sec. 2.3), a fourth holdout dataset $D_{holdout}$ is taken and combined with the training dataset post generator training: $X = D_{train} \cup D_{holdout}$. The goal of an attacker $\mathcal{A}(x)$ is to effectively discriminate which set each record $x \in X$ originated from.

The Membership Inference Attack based on the Data Plagiarism Index can divided into three steps (See Figure 3):

Step 1: Prepare 4 Datasets. In the initial step, four distinct datasets are prepared to facilitate an attack construction. The foundation begins with the original dataset (denoted in black), which is divided into 3 equal sized sub-sets: the *Training Dataset* (blue), the *Holdout Dataset* (Yellow) and the *Reference Dataset* (green). The Training Dataset is used to train the generative model, which subsequently generates the *Synthetic Dataset* (red). In parallel, the Holdout Dataset serves as an independent test set, deliberately excluded from the training phase.

Step 2: Compute DPI Values. The second step involves the computation of the *Data Plagiarism Index (DPI)* values defined at (1) for each data point within both the Holdout and Training datasets. This identifies data-copying in the local neighborhoods of test points from both datasets. The intuition here is that scores for test points from the Training set

should theoretically be higher than the Holdout set if a model is vulnerable to copying data.

Step 3: Perform MIA for Each Test Point In Step 3, a *Membership Inference Attack (MIA)* is executed to evaluate the privacy risk associated with a specific target data point. This uses the DPI score of the target data point, which is compared against a predefined threshold c. The attack can be written as:

$$\mathcal{A}(\cdot) = I(\rho(\cdot) > c), \tag{2}$$

where $\mathcal{A}$ denotes the attack function, $\rho(\cdot)$ signifies the DPI value, and I is an indicator function assessing if the DPI surpasses the threshold c. Should the DPI value of the target data point exceed c, it indicates that the data point is likely to have been part of the training data, thereby exposing a potential privacy breach. In practice, MIAs are often benchmarked using AUCROC and so the choice of c is largely irrelevant. Our implementation chooses c as the median of $\rho(x)$ for x for all test data.

The proposed DPI MIA defined at Equation (2), though straightforward, is a fast and easily interpretable attack, well-suited for auditing privacy in high-dimensional, mixed-type datasets. Indeed, as it only uses a K-Nearest Neighbor Search, it can be ran with linear or logarithmic time complexity (see [14], [21]). Moreover, DPI MIA is compatible with various definitions of distance, making it versatile to many specific applications.

5 Results

5.1 Experiment Setup

We are interested in studying how DPI compares with other Membership Inference Attacks, to what extent different tabular data generator architectures copy data, and if there are trends in the kind of data copied. To investigate these research questions, we benchmark a variety of model architectures and MIAs on the Adult Census dataset [5]. We provide descriptions of each model and MIA in appendix B and C respectively. In each experiment, Adult is randomly split into 3 equal sized training, holdout and reference sets where the generated synthetic size is also equal to these set. We replicate each experiment 5 times and visualize or report the means and standard deviations of the measures of interest where applicable.

For DPI, K has to be chosen as a hyperparameter and for data visualization purposes we plot the best performing DPI attack (Using L2 distance and K=20). We refer to appendix **??** for an ablation study of Section 5.4 containing plots with various distance metrics and K levels but in practice, DPI is fairly stable in its results regardless of K and we observed no extreme changes in behavior based on these hyperparameters.

5.2 Data-Copying in Tabular Data Generators

We first wish to confirm that there is a privacy risk to data copying in common tabular generators. Here, we benchmark a wide range of tabular generator strategies including: CTGAN and TVAE [45], Normalizing Flows (NFlow) [12], Bayesian Network (BN) [4], Adversarial Random Forests (ARF) [44], Tab-DDPM [23], PATEGAN [49], and Ads-GAN [48].

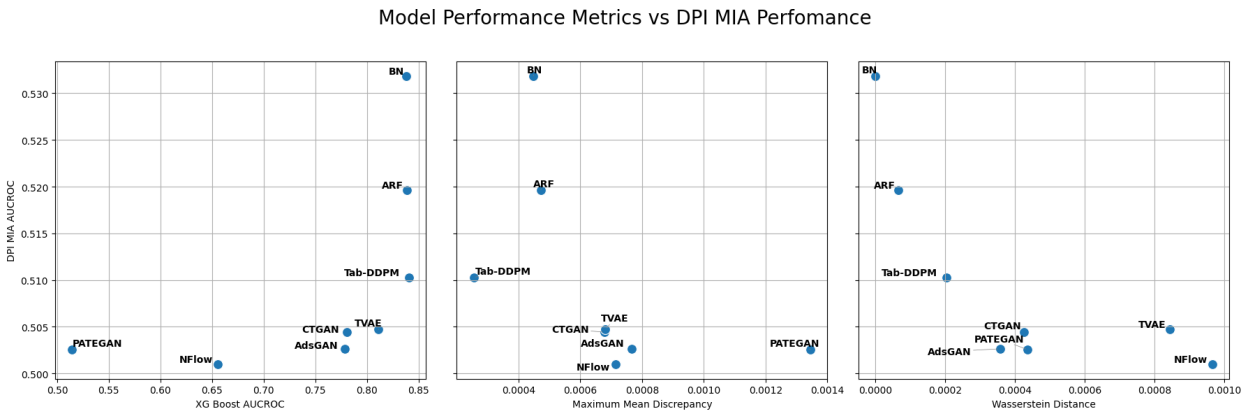

Figure 4: Classifier AUCROC, Maximum Mean Discrepancy, and Wasserstein Distance plotted with corresponding DPI MIA AUCROC for various common architectures. Interestingly, Bayesian Network, Adversarial Random Forest, and Tab-DDPM outperform other models in these performance metrics but have higher privacy risk. See Sec. 5.2 for full details.

We evaluate these models on several metrics of synthetic data quality: the Maximum Mean Discrepancy between the training and generated data (lower is better), the Wasserstein distance between the marginals of the training and generated data (lower is better), and finally the AUCROC of an XGBoost [9] classifier trained on the synthetic set and tasked to predict on a holdout set (higher is better). We deploy the DPI MIA attack as described in Section 4.2 and report its AUCROC.

We visualize the mean values of utility metrics plotted against the AUCROC of DPI MIA in Figure 4. Very interestingly, we observe that there is a clear trend between model performance and privacy risk. This suggests that data-copying could be a reason for why these models perform better, but additional research would need to be conducted to see if this is a causal or correlative relationship.

5.3 Properties of Training Data with High DPI Scores

We also investigated the training data egregiously copied according to DPI. Here, we identified the top 1% highest scored training data based on synthetic data generated by Tab-DDPM [23] as it is widely used as a benchmark in the current tabular generation literature [39, 22, 50]. We visualize these data in Figure 1 as a t-SNE plot [42] with the top 1% highly scored indexes by DPI in red. Worryingly, DPI identifies an outlier region of the distribution as being subject to this extreme top percentile data-copying (the bottom right of the plot). When analyzing these observation themselves, we found that an extreme majority were all examples of married, white, middle aged, high capital gains, private industry, respondents making ¿50k in income. In the algorithmic fairness literature that often uses Adult in benchmarking, this is considered to be a minority but very privileged class [30]. This suggests that not only is there a technical and privacy concern with data-copying, but it could also exacerbate unfairness.

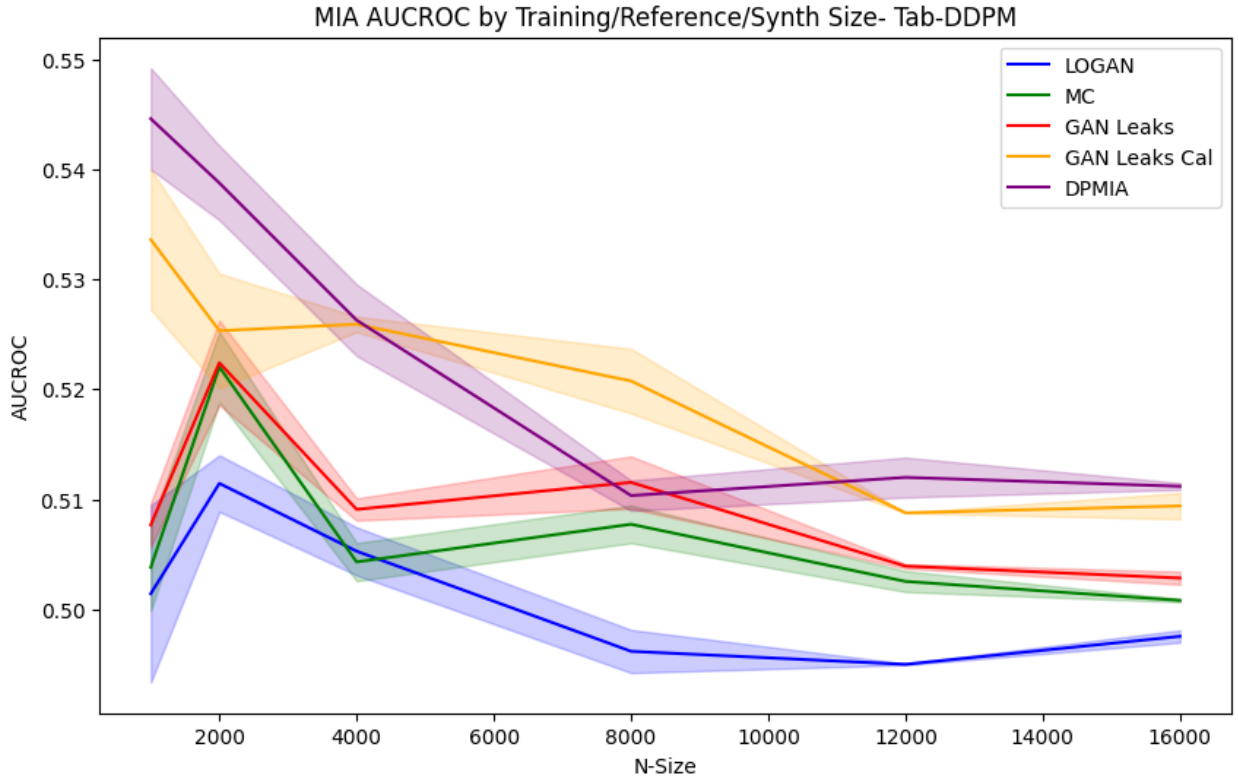

Figure 5: MIA AUCROC Benchmarks by Training Set Size on Tab-DDPM. This shows that DPI MIA is more effective than other existing MIAs described at Sec. 2.3. See Sec. 5.4 for full details.

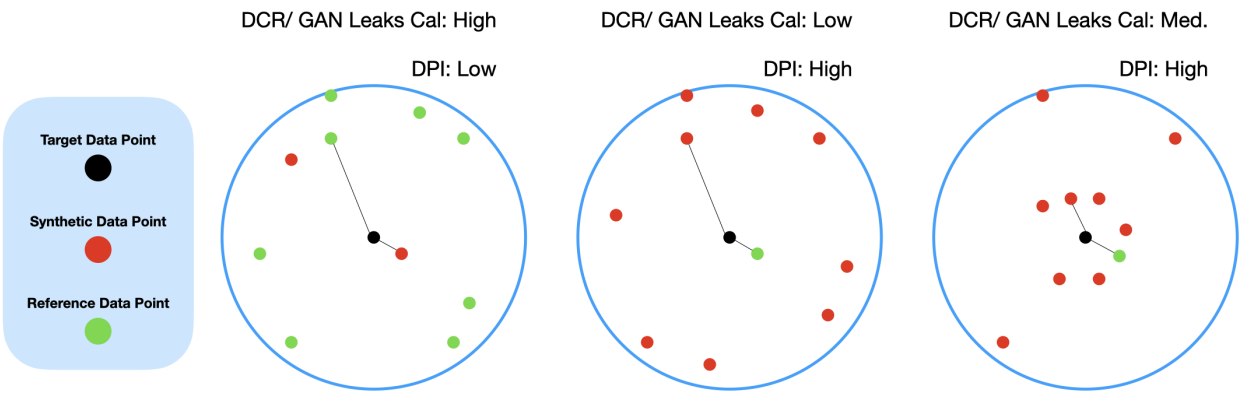

Figure 6: Distance Based Metrics/ MIAs of Overfitting vs DPI. DCR and GAN Leaks Calibrated evaluate overfitting based on a difference in relative distances whereas DPI evaluates data-copying based on the larger local neighborhood. This means that DPI evaluates data-copying differently than competing methods, explaining why predictions scores are not highly correlated (See Table 1). See Sec. 6.2 for full details.

Table 1: AUCROC for DPI and GAN Leaks Calibrated with Corresponding Pearson Correlations for the Predicted Scores. This tables shows that the proposed DPI identifies different types of data-copying than the one identified by the existing MIA attack. See Sec. 5.4 for whole details.

Model	**Metric**		
	DPI MIA	**GAN Leaks CAL**	r
Tab-DPDM	0.510±0.004	0.507±0.003	0.404±0.012
BN	0.532±0.005	0.623±0.000	0.268±0.020
ARF	0.520±0.000	0.524±0.001	0.339±0.006
AdsGAN	0.503±0.003	0.501±0.003	0.467±0.039
CTGAN	0.504±0.003	0.501±0.005	0.444±0.063
NFlow	0.501±0.006	0.500±0.000	0.482±0.023
PATEGAN	0.503±0.003	0.502±0.003	0.516±0.018
TVAE	0.505±0.005	0.502±0.003	0.466±0.023

5.4 Benchmarking the Privacy Risk of DPI

Lastly, we evaluate DPI in relation to other MIAs in order to understand its efficacy as an MIA strategy. We evaluate the black box attacks of MC [18], and GAN Leaks [8] as well as the calibrated attacks of LOGAN [17], GAN Leaks Calibrated [8] and DPI. We also benchmarked DOMIAS [41], but we found that its density estimation failed to converge and its results were always the equivalent of random guess and therefore we do not formally report it. Again, we evaluate these methods on Tab-DDPM [23]. In Figure 5, we display the results for each of these methods across various D_{train}/ D_{ref}/ D_{syn} sizes. For simplicity, each dataset is equal in size to its other.

Overall, DPI is competitive with other Membership Inference Attacks compatible with these data, dominating at most testing sizes. Interestingly, GAN Leaks Calibrated performs similarly or better than DPI. However, GAN Leaks Calibrated scores test data very differently than DPI, being based on a difference in the relative distances of the closest synthetic and reference points. In Table 1 we show that for various models, while the performance of GAN Leaks Calibrated and DPI can differ, the Pearson Correlation between their scores is relatively low. This implies that while these measures are correlated, DPI is picking up on its specific data-copying definition. Indeed, the motivation of this work is not to create a State of the Art Membership Inference Attack for all architectures, but rather to characterize the unique risk DPI implies.

6 Discussions

6.1 Implications for Privacy and Fairness in Synthetic Data

Data Plagiarism Index provides evidence that popular tabular generative models can exhibit risky data-copying behavior. This includes leaking information and favoring specific subclass outliers in the distribution of the training data. Thus, DPI can be used as a tool to audit and study model behavior. It should be noted however, that while DPI can show synthetic data are not private, it cannot prove the opposite: that particular synthetic data are private. For example, different MIAs may have various levels of success if their method targets different attributes of the synthetic data. This is proved by Table 1 where DPI and GAN Leaks Calibrated predicted scores, while correlated, were not perfectly aligned.

On top of model auditing, DPI can be connected to Differential Privacy applications [13]. Here, DPI can be used to evaluate the practical lower bound of the privacy parameter ϵ. For example, one approach is to select a pair of neighboring training datasets D_1 and D_2 and produce corresponding synthetic datasets $\widetilde{D}_1$ and $\widetilde{D}_2$ with a generative model. With DPI, we can then create corresponding data copying score distributions to find a decision rule that, given an unknown synthetic dataset $\widetilde{D}$, identifies whether its source was D_1 or D_2. If the decision rule's true positive rate is α and the false negative rate is β, based on the remark after Theorem 1 in [19], the privacy budget's lower bound can be expressed as $\epsilon \geq \log \max \left\{ \frac{\alpha}{1-\beta}, \frac{\beta}{1-\alpha} \right\}$. This approach quantifies the minimum differential privacy level that the generative model upholds.

6.2 Methodological Differences in DPI

DPI provides a new geometric definition for data-copying in the context of an available reference set and uniquely attacks this attribute relative to other MIAs. In Figure 6, we show a variety of scenarios in which hypothetical test data points are plotted with their closest synthetic and reference set neighbors. We show that under the Distance to Closest Record (DCR) metric and GAN Leaks Calibrated MIA [8] (which in many respects is the MIA version of DCR) understanding of data-copying, certain scenarios would be classified differently where they would label a positive instance of overfitting based on extreme differences in the distances of the nearest synthetic and reference points whereas DPI labels it based off of extreme differences in proportions. Thus, DPI evaluates data-copying in a fundamentally different way than DCR/ GAN Leaks Calibrated and provides additional insight into how the training and synthetic data are distributed.

6.3 Challenges in Tabular Data Generation

A key challenge of generative modelling with tabular data is the unstructured, high dimensional, mixed type nature of most datasets [45]. This poses a challenge for newer results on model data-copying and Membership Inference attacks that focus on density estimation. [6] for example proposes a scheme of comparing local densities of training data and synthetic data but does not frame their work from an MIA perspective. Similarly, they provide a proof that their method is not effective on non-smooth distributions that are characteristic

in the tabular domain. DOMIAS [41] evaluates overfitting as an MIA by comparing a test point to the probability densities of the synthetic and reference distributions. They propose two options for estimating these densities in using a Gaussian Kernel Density Estimator and a deep learning method called Block Neural Autoregressive Flow (BNAF) [11]. We found however that these methods have difficulty converging with high dimensional, mixed type datasets. Indeed, the highest dimensional tabular dataset that was benchmarked in DOMIAS was a private healthcare dataset that when one-hot-encoded was 35 columns. Adult when similarly pre-processed is 109 columns. All of our experiments with DOMIAS failed to converge, leaving its results as being the equivalent of random guess. Indeed the authors note a limitation of the work is its reliance on BNAF in that it can take several hours to train. This motivates this paper to consider a computationally easier paradigm of analyzing local neighborhoods around test points.

7 Conclusion and Future Work

In this paper, we propose a novel measure of data-copying and connect it to the Membership Inference Attack literature for tabular generative models. This allows the unique study of how local data-copying contributes to risks in the trustworthiness of these generators. Models that perform well in generating data with high measures of utility tend to copy training data more than models of a lower quality and thus have higher privacy risk profiles. Similarly, we have shown that Tab-DDPM, a highly cited and studied architecture egregiously copies outlier training data of privileged sub-classes in a widely used fairness benchmarking dataset. This indicates that data-copying can effect a variety of concepts, such as fairness, used to evaluate the trustworthiness of generative models.

Data Plagiarism Index motivates a variety of directions for future work. The disparate nature of the data-copying literature necessitates a broader theoretical framework in which to connect data-copying to privacy. Similarly, this work has shown that data-copying can affect different axis' in which to evaluate the trustworthiness of models. It would be interesting to further explore if it also effects other aspects of Trustworthy AI such as robustness, interpretability, and reliability. Lastly, DPI can be applied to Differential Privacy Auditing where it can be used to evaluate sharper privacy lower bounds.

References

[1] Mostly AI. Truly anonymous synthetic data – evolving legal definitions and technologies (part ii). 2020.

[2] Mostly AI. How to implement data privacy? a conversation with klaudius kalcher. 2021.

[3] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[4] Ankur Ankan and Abinash Panda. pgmpy: Probabilistic graphical models using python. In *Proceedings of the Python in Science Conference*, SciPy. SciPy, 2015.

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

[6] Robi Bhattacharjee, Sanjoy Dasgupta, and Kamalika Chaudhuri. Data-copying in generative models: a formal framework. In *International Conference on Machine Learning*, pages 2364–2396. PMLR, 2023.

[7] George Casella and Roger Berger. *Statistical Inference*. Duxbury Resource Center, June 2001.

[8] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20. ACM, October 2020.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Yinan Cheng, Chi-Hua Wang, Vamsi K Potluru, Tucker Balch, and Guang Cheng. Downstream task-oriented generative model selections on synthetic data training for fraud detection models. *arXiv preprint arXiv:2401.00974*, 2024.

[11] Nicola De Cao, Ivan Titov, and Wilker Aziz. Block neural autoregressive flow. *35th Conference on Uncertainty in Artificial Intelligence (UAI19)*, 2019.

[12] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows, 2019.

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[14] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3), 7 1976.

[15] Georgi Ganev and Emiliano De Cristofaro. On the inadequacy of similarity-based privacy metrics: Reconstruction attacks against "truly anonymous synthetic data", 2023.

[16] Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Matilde Karakachoff, Sophie Limou, Nicolas Vince, Matthieu Wargny, and Pierre-Antoine Gourraud. Patient-centric synthetic data generation, no reason to risk re-identification in the analysis of biomedical pseudonymised data. 05 2022.

[17] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019:133 – 152, 2017.

[18] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019:232 – 249, 2019.

[19] Florimond Houssiau, James Jordon, Samuel N Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. Tapas: a toolbox for adversarial privacy auditing of synthetic data. *arXiv preprint arXiv:2211.06550*, 2022.

[20] Din-Yin Hsieh, Chi-Hua Wang, and Guang Cheng. Improve fidelity and utility of synthetic credit card transaction time series from data-centric perspective. *arXiv preprint arXiv:2401.00965*, 2024.

[21] International Computer Science Institute and S.M. Omohundro. *Five Balltree Construction Algorithms*. Technical report (International Computer Science Institute). International Computer Science Institute, 1989.

[22] Jayoung Kim, Chaejeong Lee, and Noseong Park. STasy: Score-based tabular data synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[23] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models, 2022.

[24] Yuantong Li, Chi-Hua Wang, and Guang Cheng. Online forgetting process for linear regression models. In *International Conference on Artificial Intelligence and Statistics*, pages 217–225. PMLR, 2021.

[25] Tongyu Liu, Ju Fan, Guoliang Li, Nan Tang, and Xiaoyong Du. Tabular data synthesis with generative adversarial networks: design space and optimizations. *The VLDB Journal*, 33(2):255–280, aug 2023.

[26] Yucong Liu, Chi-Hua Wang, and Guang Cheng. On the utility recovery incapability of neural net-based differential private tabular training data synthesizer under privacy deregulation. *arXiv preprint arXiv:2211.15809*, 2022.

[27] Pei-Hsuan Lu, Pang-Chieh Wang, and Chia-Mu Yu. Empirical evaluation on synthetic data generation with generative adversarial network. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, WIMS2019, New York, NY, USA, 2019. Association for Computing Machinery.

[28] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A three sample hypothesis test for evaluating generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 3546–3556. PMLR, 2020.

[29] Matthieu Meeus, Florent Guepin, Ana-Maria Cretu, and Yves-Alexandre de Montjoye. *Achilles' Heels: Vulnerable Record Identification in Synthetic Data Publishing*, page 380–399. Springer Nature Switzerland, 2024.

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022.

[31] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.

[32] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083, June 2018.

[33] Elizabeth S Parker, Larry Cahill, and James L McGaugh. A case of unusual autobiographical remembering. *Neurocase*, 12(1):35–49, 2006.

[34] Michael Platzer and Thomas Reutterer. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*, 4:679939, 2021.

[35] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023.

[36] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.

[37] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, Los Alamitos, CA, USA, may 2017. IEEE Computer Society.

[38] Aivin V Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers. *arXiv preprint arXiv:2302.02041*, 2023.

[39] Namjoon Suh, Xiaofeng Lin, Din-Yin Hsieh, Mehrdad Honarkhah, and Guang Cheng. Autodiff: combining auto-encoder and diffusion model for tabular data synthesizing. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023.

[40] Lan Tao, Shirong Xu, Chi-Hua Wang, Namjoon Suh, and Guang Cheng. Discriminative estimation of total variation distance: A fidelity auditor for generative data. *arXiv preprint arXiv:2405.15337*, 2024.

[41] Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection, 2023.

[42] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[43] Chi-Hua Wang and Guang Cheng. Badgd: A unified data-centric framework to identify gradient descent vulnerabilities. *arXiv preprint arXiv:2405.15979*, 2024.

[44] David S. Watson, Kristin Blesch, Jan Kapar, and Marvin N. Wright. Adversarial random forests for density estimation and generative modeling. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5357–5375. PMLR, 25–27 Apr 2023.

[45] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Neural Information Processing Systems*, 2019.

[46] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Assessing privacy and quality of synthetic health data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, AIDR '19, New York, NY, USA, 2019. Association for Computing Machinery.

[47] Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE journal of biomedical and health informatics*, 24(8):2378–2388, 2020.

[48] Jinsung Yoon, Lydia N Drumright, and Mihaela van der Schaar. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE journal of biomedical and health informatics*, 24(8):2378—2388, August 2020.

[49] 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.

[50] Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *The Twelfth International Conference on Learning Representations*, 2024.

[51] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. Ctab-gan: Effective table data synthesizing. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 97–112. PMLR, 17–19 Nov 2021.

A Ablation study

DPI requires a practitioner to specify a distance metric and K number of nearest neighbors in order to be deployed. This presents a hyperparameter tuning problem as to what measure of distance should be used and how large the neighbors should be. We replicate the experiment from Section 5.4 but this time with common distance metrics (L1 and L2) as well as a variety of K sizes (5, 10, 20, 30). We plot the means and standard deviations of the corresponding DPI MIA attacks in Figure 7. While the success of the attacks vary with lower sample sizes, each attack follows a clear trend with each, eventually seeing smaller deviations at the maximum sample sizes.

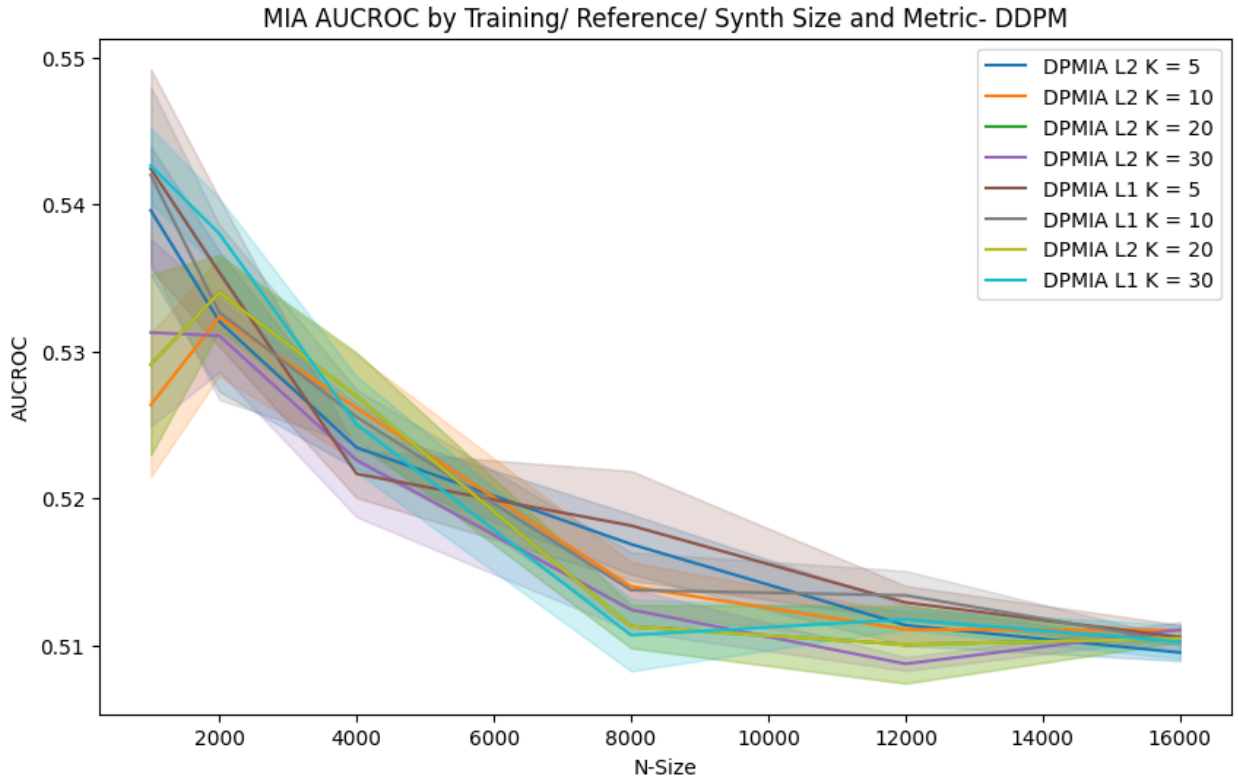

Figure 7: Ablation Results for Various Distance Measures and K Size Choices.

B Model Descriptions

In all experiments, we use the implementations of these models from the Python package Synthcity [35]. For benchmarking purposes we use the default hyperparameters for each model. A brief description for each model is as follows:

CTGAN [45]: (Conditional Tabular Generative Adversarial Network) uses a GAN framework with conditional generator and discriminator to capture multi-modal distributions. It uses mode normalization to better learn mixed-type distributions.

TVAE [45]: (Tabular Variational Auto-Encoder) is very similiar to CTGAN in its use of mode normalizing techniques, but rather than using a GAN architecture, instead employees A VAE.

Normalizing Flows (NFlow) [12]: Normalizing flows transform a simple base distribution (e.g. Gaussian) into a more complex one matching the data by applying a sequence of invertible, differentiable mappings.

Bayesian Network (BN) [4],: Bayesian Networks use a Directed Acyclic Graph to represent the joint probability distribution over variables as a product of marginal and conditional distributions. It then samples the empiric distributions estimated from the training dataset.

Adversarial Random Forests (ARF) [44]: ARFs extend the random forest model by adding an adversarial stage. Random forests generate synthetic samples which are scored against the real data by a discriminator network. This score is used to re-train the forests iteratively.

Tab-DDPM [23]: Tabular Denoising Diffusion Probabilistic Model adapts the DDPM framework from image synthesis. It iteratively refines random noise into synthetic data by learning the data distribution through gradients of a classifier on partially corrupted samples with gaussian noise.

PATEGAN [49]: The PATEGAN model uses a neural encoder to map discrete tabular data into a continuous latent representation which is sampled from during generation by the GAN discriminator and generator pair.

Ads-GAN [47]: Ads-GAN uses a GAN architecture for tabular synthesis but also adds an identifiability metric to increase its ability to not mimic training data.

C Membership Inference Attack Descriptions

A description of each of the Membership Inference Attacks referenced in the paper are as follows:

LOGAN [17]: LOGAN proposes a variety of MIA strategies. A black box version of their attack involves training a Generative Adversarial Network (GAN) on the synthetic dataset and using the discriminator to score test data. A calibrated version improves upon this by training a binary classifier to distinguish between the synthetic and reference dataset. In this paper we only benchmark the calibrated version.

GAN Leaks/ GAN Leaks Calibrated [8]: GAN Leaks is a black box attack that scores test data based on a sigmoid score of the distance to the nearest neighbor in the synthetic dataset. GAN Leaks Calibrated improves on this with the inclusion of a reference set in which this distance is subtracted from the distance to the closest record in the reference set.

MC [18]: MC is based on counting the amount of observations in the synthetic dataset that fall into the neighborhood of a test point (Monte Carlo Integration). However, they do not consider a reference dataset and the choice of distance for what to consider a neighborhood is a non-trivial hyperparameter to tune.

DOMIAS [41]: DOMIAS is a calibrated attack which scores test data by performing density estimation on the synthetic and reference datasets to then calculate the probability ratio of the test data being from the synthetic vs reference distributions.