

Evaluating Privacy Metrics for Synthetic Tabular Data

by

Mushi Wang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

© Mushi Wang 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This paper addresses the challenge of evaluating privacy risks in synthetic tabular data by examining black-box privacy metrics that do not require detailed knowledge of the data generation process. We focus on two sorts of attacks, black-box and white-box attacks. Utilizing six datasets from the UCI Machine Learning Repository, we evaluate the effectiveness of these metrics across various synthetic data generation models, including diffusion models like TabDDPM and traditional models like PrivBayes. Our findings reveal that while DOMIAS exhibits limited sensitivity across different datasets and configurations, DCR proves to be an effective measure of similarity between synthetic and real data, offering significant insights into privacy preservation. We also introduce the Step-wise Error Comparing Membership Inference (SECMCI) attack, which assesses prediction errors at each generation step to infer membership status. The study concludes that diffusion models, such as TabDDPM, generally achieve a superior balance of utility and privacy compared to traditional models. These results highlight the need for robust, adaptable privacy metrics to reliably assess privacy risks in synthetic data, thereby ensuring its safe application across various domains.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, **Prof. Xi He**, whose expertise, understanding, and patience significantly contributed to my graduate experience. Prof. He's exceptional guidance, persistent support, and insightful feedback have been instrumental in shaping the direction and quality of this thesis. Her commitment to excellence and her ability to inspire and challenge me have made this journey an enriching and rewarding one. Without her insightful guidance and persistent help, this thesis would not have been possible.

I am also deeply thankful to the members of my thesis committee, **Prof. Florian Kerschbaum**, and **Prof. Xiao Hu**, for their valuable feedback and suggestions, which greatly improved the quality of my work.

I am profoundly grateful to my colleagues and fellow researchers in the Data System Group. In particular, I would like to thank **Prof. Semih Salihoglu**, **Wei Pang** for their inspiration and motivation. The countless hours spent brainstorming ideas, troubleshooting issues, and celebrating successes together have been some of the most memorable moments of my academic journey. A special thanks to my virtual colleague, ChatGPT, in assisting with research, providing insights, and refining the language and structure of this thesis.

In addition, I would like to acknowledge my close friends **Chenxin Zhang**, **Ruifeng Wang**, **Youwei Ma**, and **Yong Qin**, whose friendship and encouragement have been invaluable. Their support and understanding have helped me maintain a balanced and positive outlook throughout my studies.

Lastly, I would like to extend my sincere thanks to my family. My parents, **Junyan Huang**, and **Lijie Wang**, have been a constant source of encouragement and strength. I also would like to express my appreciation to, my grandparents, **Yugong Huang**, **Zhijie Xu**, **Guie Luan**, and **Bingli Wang**, and all the other families, , who supported me along my journey to success. Their supports guide me through the toughest moments.

Dedication

This thesis is dedicated to my parents, Junyan Huang and Lijie Wang, whose unwavering support, encouragement, and love have been the foundation of my academic and personal achievements.

To my grandparents, Yugong Huang, Zhijie Xu, Guie Luan, and Bingli Wang, for their wisdom, strength, and the countless lessons they have imparted.

And to my friends and colleagues, who have provided constant support and companionship throughout this journey.

Table of Contents

Author’s Declaration	ii
Abstract	iii
Acknowledgments	iv
Dedication	v
List of Figures	ix
List of Tables	x
1 Introduction	1
2 Related Work	3
2.1 Synthetic Data Generators	3
2.1.1 Statistical Approach with DP Guarantees	3
2.1.2 GAN-based Models	4
2.1.3 Diffusion Models	5
2.2 Privacy Attacks	6
2.2.1 Membership Inference Attacks	6
2.2.2 Privacy Attacks on GANs	7
2.2.3 Privacy Attacks on Diffusion Models	8

3	Research Problem	10
4	Preliminaries	11
4.1	Diffusion Models	11
4.1.1	Denoising Diffusion Probabilistic Models	11
4.2	Membership Inference Attacks (MIAs)	13
4.2.1	Black-box MIAs	13
4.2.2	White-box MIAs	14
5	Methodology	15
5.1	Black box privacy metrics	16
5.1.1	DOMIAS	16
5.1.2	Distance to Closest Record (DCR)	17
5.2	White-box Privacy Metrics	17
5.2.1	Integrating TabDDPM to SECMi attack	17
5.3	Synthetic Data utility Metrics	19
5.3.1	α -precision	19
5.3.2	β -Recall	19
5.3.3	Detection score	20
5.3.4	One-way Marginal/Column Shapes	20
5.3.5	Two-way Marginal/Column Pair Trends	20
6	Evaluation Results	22
6.1	Datasets	22
6.2	SECMi Attack Effectiveness	23
6.2.1	Models	23
6.2.2	Findings	24
6.3	SECMi Attack on Image vs. Tabular Data	25
6.4	Utility and Privacy of Synthetic Data	30

7 Conclusion	34
References	35

List of Figures

4.1	Structure of DDPM [13]	11
6.1	median of relative t-errors on TabDDPM on Adult dataset	27
6.2	median of relative t-errors on image DDIM on CIFAR10 dataset	27
6.3	t-error distributions of training set and hold-out set at timestep 1 of the first column in Adult dataset	28
6.4	t-error distributions of training set and hold-out set at timestep 1 of the first column in CIFAR-10 dataset	29

List of Tables

6.1	Statistics of datasets	23
6.2	accuracy and AUC of SECMI attack on TabDDPM _{one-hot} on six datasets. TP, TP, FP, FN are true positive rate, true negative rate, false positive rate, and false negative rate respectively	24
6.3	accuracy and AUC of SECMI attack on TabDDPM _{label} on six datasets. TP, TP, FP, FN are true positive rate, true negative rate, false positive rate, and false negative rate respectively	24
6.4	SECMI attack on iamge data	26
6.5	privacy and utility metrics for different models on the Adult dataset	31
6.6	privacy and utility metrics for different models on the Default dataset	31
6.7	privacy and utility metrics for different models on the Beijing dataset	32
6.8	privacy and utility metrics for different models on the Magic dataset	32
6.9	privacy and utility metrics for different models on the News dataset	32
6.10	privacy and utility metrics for different models on the Shoppers dataset . . .	33

Chapter 1

Introduction

Synthetic data generation model [38, 34, 39] has emerged as a promising solution to this challenge by creating artificial datasets that mimic the statistical properties of real data while ensuring individual privacy. However, assessing the privacy of these synthetic datasets remains a critical concern [15]. Effective privacy metrics are essential to evaluate and guarantee the extent to which synthetic data can protect the confidentiality of the original dataset.

Privacy metrics are crucial tools in the field of data privacy, enabling researchers and practitioners to quantify the privacy risks associated with the use of synthetic data. These metrics help in identifying potential vulnerabilities that could lead to privacy breaches, such as membership inference attacks, where an adversary attempts to determine whether a particular data point was included in the training dataset. Addressing these vulnerabilities is essential for ensuring that synthetic data can be safely used in applications ranging from healthcare to finance.

The importance of synthetic data has grown in parallel with the expanding volume and sensitivity of data collected by organizations. Synthetic data can facilitate data sharing and collaboration across institutions while mitigating the risks of exposing sensitive information. For instance, in the healthcare sector, synthetic patient records can be used for research and development without compromising patient privacy. Similarly, in finance, synthetic transaction data can be used to test fraud detection systems without exposing real transaction details.

Despite its potential, the generation of synthetic data poses significant challenges. One primary concern is ensuring that synthetic data retains the utility of the original data while providing robust privacy protection. This balance is delicate; overly anonymized

data may lose its analytical value, while insufficient anonymization may fail to protect privacy. Therefore, developing reliable privacy metrics that can accurately assess this balance is of paramount importance.

This paper focuses on evaluating privacy metrics for synthetic tabular data, with a particular emphasis on black-box privacy metrics. Black-box metrics assess privacy without needing detailed knowledge of the underlying data generation process, making them versatile and applicable across various generative models. We investigate the effectiveness of two prominent black-box privacy metrics: Density Overfitting Membership Inference Attack with Synthetic Data (DOMIAS) [35] and Distance to Closest Record (DCR). These metrics are evaluated based on their ability to measure the risk of membership inference attacks and the similarity between synthetic and real data.

To provide a comprehensive analysis, our study includes extensive experiments using six real-world datasets from the UCI Machine Learning Repository. These datasets represent a diverse set of domains, including census income, credit card default, gamma telescope, online shoppers' purchasing intentions, news popularity, and air quality measurements. We compare the performance of different synthetic data generation models, including diffusion models like TabDDPM [20] and traditional models like PrivBayes [39], under various configurations. Additionally, we explore the utility-privacy trade-offs presented by these models to provide a comprehensive understanding of their effectiveness in real-world applications.

By examining both the privacy and utility of synthetic data, this paper aims to offer a balanced view of how well current models and metrics perform. Our goal is to identify the strengths and weaknesses of existing privacy metrics and propose directions for future research to enhance the assessment of synthetic data privacy. This investigation is crucial for advancing the safe and effective use of synthetic data in practice, fostering innovation while protecting individual privacy.

Furthermore, this study contributes to the broader discourse on data privacy by highlighting the practical implications of using synthetic data in various sectors. It underscores the necessity of developing robust and adaptable privacy metrics that can reliably assess the privacy risks associated with synthetic data across diverse applications. Through rigorous experimentation and analysis, we aim to provide actionable insights that can guide the development of more secure and effective synthetic data generation methods, ultimately supporting the widespread adoption of privacy-preserving technologies.

Chapter 2

Related Work

2.1 Synthetic Data Generators

Synthetic data can be produced using a variety of methods, including statistical models, generative adversarial networks (GANs), and other machine learning algorithms. These data sets are invaluable for training and validating models, enhancing data diversity, and preserving privacy by reducing the need for actual personal or confidential information. Moreover, synthetic data generation allows for the exploration of scenarios and edge cases that may not be present in the original data, thereby improving the robustness and generalizability of analytical models.

2.1.1 Statistical Approach with DP Guarantees

Statistical approaches with differential privacy (DP) guarantees offer a robust framework for balancing the trade-off between data utility and privacy. These methods aim to protect individual data entries while still allowing for meaningful statistical analysis and insights. By introducing controlled noise into statistical computations, differential privacy ensures that the inclusion or exclusion of a single data point does not significantly affect the outcome, thus preserving privacy. This approach has become a cornerstone in privacy-preserving data analysis, enabling the release of aggregated information, synthetic data, and machine learning models with quantifiable privacy assurances. Through techniques such as Bayesian networks, generative models, and local perturbation, statistical methods with DP guarantees provide versatile solutions for a wide range of applications, from healthcare to finance, where the confidentiality of sensitive information is paramount.

PrivBayes [39] and PrivSyn [41] are two popular models with DP guarantees. PrivBayes uses a Bayesian network to model the correlations among attributes in a high-dimensional dataset. By injecting noise into the low-dimensional marginals of this network, PrivBayes effectively mitigates the curse of dimensionality, making it feasible to publish useful high-dimensional synthetic data under differential privacy guarantees. This approach is particularly effective in preserving the statistical properties of the original dataset, as the Bayesian network captures complex dependencies between attributes. However, constructing and managing a Bayesian network can be computationally intensive, especially for very large datasets.

On the other hand, PrivSyn employs local differential privacy (LDP) by perturbing each data point individually before any synthesis occurs. This two-phase approach, where the first phase ensures that individual data points are privatized, and the second phase involves learning a generative model from the perturbed data, allows PrivSyn to maintain strong privacy guarantees from the onset. This method decouples data perturbation from data synthesis, potentially leading to higher utility in the synthetic data because the generative model can more accurately reflect the underlying data distribution without being directly influenced by global noise addition. However, the reliance on LDP means that the initial perturbation phase can sometimes lead to a loss of utility if not carefully managed.

2.1.2 GAN-based Models

Generative Adversarial Networks (GANs) have been a cornerstone of generative models since their introduction by Goodfellow et al. in 2014 [10]. GANs consist of two neural networks, a generator and a discriminator, that are trained simultaneously through adversarial processes: the generator aims to produce realistic data samples, while the discriminator strives to distinguish between real and generated samples. This adversarial training framework has led to advancements in generating high-fidelity images, realistic video frames, and even coherent text sequences [24].

DPGAN [36] and dp-GAN [40] are pioneering GAN-based approaches designed to operate under differential privacy (DP) settings, addressing both privacy and data utility concerns. These models introduce carefully calibrated noise to the gradients during the training process, which helps in maintaining the privacy of the individual data points in the training set. By leveraging this technique, DPGAN and dp-GAN are able to generate high-fidelity synthetic data while ensuring that the privacy of the original data is preserved.

PATE-GAN [18] modifies the standard approach of updating the discriminator using differentially private stochastic gradient descent (DPSGD). The authors identify that

adding noise solely to the discriminator updates disrupts the balance between the generator and discriminator. To address this, they propose taking more discriminator steps between generator steps, using larger batch sizes, and adapting the discriminator update frequency. These adjustments significantly enhance the quality of generated data, outperforming previous GAN privatization schemes on standard image synthesis benchmarks.

2.1.3 Diffusion Models

Diffusion models have gained significant attention in recent years due to their efficacy in generating high-quality samples across various domains, particularly in image synthesis. Unlike traditional GANs, diffusion models operate by iteratively refining samples from a noise distribution towards the data distribution through a sequence of denoising steps [14]. This iterative approach, inspired by non-equilibrium thermodynamics, allows for more stable training and improved sample quality [31]. Recent advancements in this field have demonstrated the potential of diffusion models in achieving state-of-the-art results in image generation tasks, surpassing the performance of GANs in terms of both fidelity and diversity of generated samples [5].

Diffusion models have been successfully applied to a diverse range of generative modeling tasks, including image generation [32], [13], [33] and image super-resolution [26], [27]. These applications showcase the versatility and effectiveness of diffusion models in producing high-quality synthetic data across various domains, highlighting their potential for broader adoption in generative tasks. Recent research has also demonstrated the potential of diffusion models for tabular data synthesis, further expanding their applicability and showcasing their potential for broader adoption in generative tasks.

TabDDPM [20] explores the use of denoising diffusion probabilistic models (DDPMs) for generating synthetic tabular data, which includes heterogeneous features such as numerical and categorical data. The introduced model, TabDDPM, leverages both multinomial and Gaussian diffusions to handle different types of features effectively. Extensive evaluations show that TabDDPM outperforms GAN and VAE-based alternatives in terms of generating high-quality synthetic data while also being well-suited for privacy-sensitive applications, ensuring better separation between synthetic and real data points.

TabSyn [38] takes a step further and achieves a better performance than its predecessor TabDDPM. TabSyn leverages a hybrid model combining Variational Autoencoders (VAEs) and score-based diffusion model to capture complex data distributions.

2.2 Privacy Attacks

Synthetic data generation introduces several privacy concerns that need careful consideration. Membership inference attacks aim to determine whether a specific record was included in the training dataset, potentially exposing individuals' involvement. Attribute inference attacks enable attackers to deduce sensitive attributes of individuals by analyzing patterns within the synthetic data. Furthermore, linkage attacks involve combining synthetic data with other publicly available datasets, which can lead to the identification of individuals in the training data. These risks are relevant for both tabular synthetic data and generative models, underscoring the necessity of ensuring that synthetic data is sufficiently anonymized to prevent re-identification through linkage with external data sources.

In this work, we choose membership inference attacks (MIAs) as our primary privacy metric. MIAs are the most common and straightforward privacy metric for synthetic data generators, offering a clear and direct measure of privacy risks. By focusing on MIAs, we aim to evaluate the extent to which synthetic data generation methods protect against the exposure of individual records within the training dataset. This approach provides a tangible benchmark for assessing the effectiveness of privacy-preserving techniques in synthetic data generation and highlights areas where further improvements are necessary to enhance data privacy. Moreover, MIAs allow for a standardized comparison across different synthetic data generation methods, facilitating the identification of best practices and the development of more robust privacy-preserving algorithms.

2.2.1 Membership Inference Attacks

Membership inference attacks represent a significant privacy threat in machine learning, where an adversary seeks to determine whether a particular data point was part of a model's training dataset. These attacks exploit overfitting and the model's sensitivity to specific data points to infer membership information.

In their seminal work, Shokri et al. [30] introduced the concept of membership inference attacks against machine learning models. They demonstrated that an adversary with access to the model's predictions can train a shadow model to simulate the target model's behavior, thereby inferring the membership status of specific data points. This approach leverages differences in the target model's output distributions for training and non-training data, which can be particularly pronounced in models that overfit to their training data.

Fredrikson et al. [8] extended this line of research by introducing model inversion attacks, where the adversary aims to reconstruct sensitive input data from the model’s outputs. While not purely a membership inference attack, model inversion can be used to infer sensitive details about the training data, further highlighting the privacy risks associated with machine learning models.

Recent advancements have continued to reveal vulnerabilities in various types of models. For example, Salem et al. [28] showed that membership inference attacks could be generalized across different types of models and datasets, underscoring the widespread applicability of these attacks. Similarly, Nasr et al. [21] provided a comprehensive analysis of membership inference attacks across different threat models, including white-box and black-box scenarios, and proposed defense mechanisms to mitigate these risks.

To address the privacy concerns posed by membership inference attacks, several defense strategies have been proposed. Differential privacy is a prominent approach that adds controlled noise to the training process, thereby reducing the model’s reliance on specific data points and mitigating the risk of membership inference [7]. Regularization techniques, such as dropout and weight decay, have also been shown to enhance model robustness against these attacks by reducing overfitting [17].

In conclusion, membership inference attacks expose a critical vulnerability in machine learning models, particularly those trained on sensitive data. As machine learning models become increasingly integrated into various applications, developing robust defense mechanisms to protect training data privacy remains an urgent priority.

2.2.2 Privacy Attacks on GANs

While Generative Adversarial Networks (GANs) have shown great potential, concerns about privacy attacks have also emerged, particularly in sensitive applications such as healthcare. These models can sometimes inadvertently reveal information about the training data. For instance, membership inference attacks allow an adversary to determine if a specific data point was included in the training set [30]. These attacks leverage the fact that GANs may generate samples closely resembling the training data, which can pose privacy risks. Similarly, model inversion attacks aim to reconstruct input data from the model’s output, potentially exposing sensitive information [8]. Studies have highlighted the vulnerabilities of even advanced GANs like StyleGAN to such privacy issues [11]. To address these concerns, various defense mechanisms have been proposed. For example, differential privacy introduces noise during training to obscure individual data contributions [36], and adversarial regularization aims to make models more resilient to privacy

attacks [3]. Specific applications like medGAN [4] and healthGAN [37] have focused on enhancing privacy-preserving features in GANs for healthcare data, balancing the need for data utility with robust privacy safeguards.

2.2.3 Privacy Attacks on Diffusion Models

Diffusion models, while celebrated for their stability and high-quality generative capabilities, have also raised privacy concerns, particularly regarding the potential leakage of training data. A notable type of privacy attack against these models is the membership inference attack. In such attacks, an adversary attempts to determine whether a specific data point was part of the training dataset used to train the diffusion model.

Recent studies have highlighted the vulnerability of diffusion models to membership inference attacks. For example, Duan et al. [6] explores the susceptibility of diffusion models to such attacks, demonstrating that these models can indeed leak membership information under certain conditions. This research shows that the iterative nature of diffusion models, which involves refining samples from noise, can inadvertently reveal details about the training data.

Further examination by Carlini et al. [2] delves into how training data can be extracted from diffusion models, indicating that these models can expose more information than previously anticipated. This extraction capability raises significant privacy concerns, especially in sensitive applications where training data confidentiality is paramount.

The specific case of membership inference attacks via quantile regression has been explored by Tang et al. [34], showing another vector through which diffusion models can be probed for membership information. This study employs statistical methods to infer membership with a high degree of accuracy, thereby highlighting a critical vulnerability in the current design of diffusion models.

Other works, such as [23], investigate white-box membership inference attacks, where the adversary has access to the internal mechanisms of the diffusion model. This access allows for a more precise inference of training data membership, emphasizing the need for robust defense mechanisms.

To mitigate these privacy risks, various techniques have been proposed. For instance, TabSyn[38] and TabDDPM [20] suggest modifications to diffusion models aimed at enhancing their robustness against membership inference attacks. These approaches involve altering the training procedures and introducing noise to obscure the contributions of individual data points, thereby protecting the privacy of the training data.

In conclusion, while diffusion models offer significant advancements in generative modeling, their vulnerability to privacy attacks, particularly membership inference attacks, necessitates the development of more secure training protocols and defense mechanisms. Ensuring the privacy of training data remains a critical challenge in the deployment of diffusion models in real-world applications.

Chapter 3

Research Problem

The research questions in this thesis address critical concerns in the field of synthetic data generation and privacy preservation. The first question explores whether synthetic data generation models for tabular data are vulnerable to privacy attacks. This inquiry is crucial because, while synthetic data is often considered a safer alternative to real data, it is essential to understand if these models can inadvertently expose sensitive information through various types of privacy attacks, such as membership inference or re-identification attacks. The second question investigates the effectiveness of privacy metrics, aiming to discern how one can evaluate whether a privacy metric genuinely reflects the protection of sensitive information. This question is vital because not all privacy metrics are created equal—some may give a false sense of security, while others might effectively quantify the risk of exposure. Together, these questions seek to deepen our understanding of the balance between data utility and privacy, which is foundational in the development of secure synthetic data generation techniques.

Chapter 4

Preliminaries

4.1 Diffusion Models

The inspiration of diffusion models comes from non-equilibrium thermodynamics. Diffusion model is a type of generative model that estimates the target distribution by Markov chain. These models utilize a sequence of noise injections and subsequent denoising steps to generate high-fidelity synthetic data from random noise inputs.

4.1.1 Denoising Diffusion Probabilistic Models

An example of diffusion models in the field of generative modeling is the Denoising Diffusion Probabilistic Models (DDPMs) [13]. The Markov chain framework is utilized in DDPMs, and an example of a DDPM is illustrated in Figure 4.1.

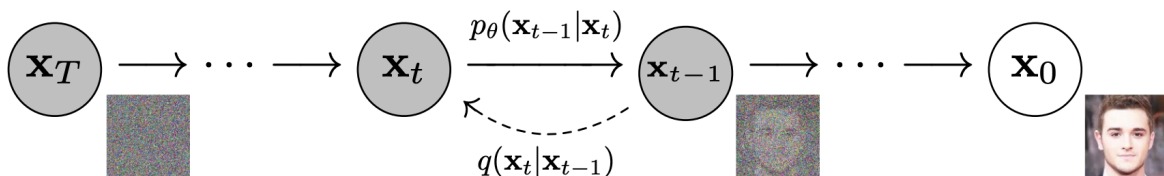


Figure 4.1: Structure of DDPM [13]

Given the initial data distribution $x_0 \sim q(x_0)$, the forward Markov process produces a sequence of random variables x_1, x_2, \dots, x_t with a transition kernel $q(x_t|x_{t-1})$. By applying the chain rule of probability and utilizing the Markov property, the joint distribution of x_1, x_2, \dots, x_t conditioned on x_0 , denoted as $q(x_1, \dots, x_t|x_0)$ can be factorized as follows:

$$q(x_1, \dots, x_t|x_0) = \prod_{t=1}^t q(x_t|x_{t-1}) \quad (4.1)$$

One common choice for the transition kernel is Gaussian distribution and it is defined as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (4.2)$$

where β_t is a variance schedule, and \mathcal{N} denotes a Gaussian distribution. By utilizing the property of the diffusion process, we can derive a closed form for x_t at any time step t . Let $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I) \quad (4.3)$$

The reverse process, which aims to denoise the data, is modeled as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (4.4)$$

Here, μ_θ and Σ_θ are parameterized functions learned during the training phase.

During the training process, the model learns μ_θ and Σ_θ by minimizing the variational lower bound on the negative log-likelihood. The variational lower bound can be decomposed into a sum of KL divergence terms and a reconstruction term:

$$\begin{aligned} & \mathbb{E}_q \left[\underbrace{D_{KL}(q(x_t|x_0)||p(x_t))}_{L_T} \right. \\ & \quad \left. + \sum_{t=2}^T \underbrace{D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))}_{L_{t-1}} \right. \\ & \quad \left. - \log p_\theta(x_0|x_1) \right] \quad (4.5) \\ & \quad \underbrace{\hspace{10em}}_{L_0} \end{aligned}$$

where $D_{KL}(q||p)$ denotes the Kullback-Leibler (KL) divergence between two probability distributions q and p . Furthermore, $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t) = \sigma_t^2 I)$. We can represent the loss term L_t at time step t as:

$$L_t = \mathbb{E}_q[\frac{1}{2\sigma_t^2} \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2] + C \quad (4.6)$$

where C is a constant that does not depend on θ .

4.2 Membership Inference Attacks (MIAs)

A membership inference attack is a type of adversarial attack on machine learning models where the adversary seeks to determine whether a specific data point was included in the model’s training set. This attack exploits the model’s responses or output probabilities to infer the presence or absence of individual data points in the training dataset. The fundamental vulnerability that membership inference attacks leverage is the discrepancy in the model’s behavior between data it has been trained on and data it has not seen before.

Consider a random variable X with distribution $p(X)$. Let $\mathcal{D}_T \stackrel{i.i.d.}{\sim} p(X)$ represent the training set of a generator model, and $\mathcal{D}_H \stackrel{i.i.d.}{\sim} p(X)$ be the holdout set, where $\mathcal{D}_T \cap \mathcal{D}_H = \emptyset$. Given a generator model f_θ with trainable parameters θ , and a dataset $\mathcal{D} = \mathcal{D}_T \cup \mathcal{D}_H$, a membership inference attack model \mathcal{M} on $\vec{x} \in \mathcal{D}$ and f_θ is defined as:

$$\mathcal{M}(f_\theta, \vec{x}) = \mathbb{1}(\vec{x} \in \mathcal{D}_T) \quad (4.7)$$

4.2.1 Black-box MIAs

Black-box MIAs operate under the assumption that the attacker has only limited access to the model, typically through an API that provides predictions or confidence scores for input data. The attacker does not have direct access to the model’s internal parameters or architecture. Instead, they rely solely on the model’s outputs to infer membership.

In black-box MIAs, attackers often employ statistical and machine learning techniques to analyze the model’s behavior. By training shadow models, which are designed to mimic the target model’s behavior, attackers can create datasets that help distinguish between

members and non-members. These shadow models are trained on auxiliary datasets with a known membership status, and the attack model is then used to infer the membership status of data points in the target model.

4.2.2 White-box MIAs

White-box MIAs, on the other hand, assume that the attacker has full access to the model’s internal parameters, architecture, and potentially the training algorithm. This level of access allows the attacker to leverage more detailed information to infer membership status.

With access to the model’s gradients, weights, and other internal states, white box attackers can conduct more precise and effective attacks. For instance, they can analyze the model’s gradients during the training process to identify patterns that indicate whether a particular data point was included in the training set. This detailed information can significantly enhance the accuracy of membership inference, as it provides insights into how the model’s parameters were adjusted in response to specific training data.

White-box MIAs are generally more powerful than black-box attacks due to the additional information available to the attacker. However, they also require a higher level of access, which may not always be feasible in practice.

Chapter 5

Methodology

A good privacy metric should contain the following properties:

They should correlate with data quality. Consider the following example: Assume we generate completely random synthetic data. Although this method would preserve complete privacy, as the synthetic data bears no resemblance to the original data, it also offers very low utility. This is because random data lacks any meaningful patterns or relationships that can be used for analysis or decision-making.

In contrast, high-quality synthetic data should maintain the statistical properties and relationships present in the original data, enabling it to be used effectively for analytical purposes. Thus, a good privacy metric should reflect this balance between privacy and utility. It should penalize methods that, while offering high privacy, result in synthetic data that is too dissimilar from the original data to be useful. Conversely, it should reward methods that achieve a reasonable compromise, maintaining data utility while still providing a robust level of privacy protection.

They should be robust and consistent against various attack models and datasets. A privacy metric should be model and data invariant. In other words, it should maintain its reliability and accuracy across different types of synthetic data generation models and various datasets. This robustness is crucial because synthetic data can be generated using diverse methodologies and applied to a wide range of datasets.

They should be scalable to large datasets. Compared to datasets from years ago, current datasets are becoming increasingly larger and more complex. This trend is driven by the exponential growth in data generation across various fields such as healthcare, finance, social media, etc. As datasets grow, it is essential that privacy metrics for synthetic data scale effectively to handle this increased volume and complexity.

They should be sensitive to different levels of privacy. In privacy-preserving frameworks like differential privacy, a privacy budget (often denoted as ϵ) is provided, quantifying the trade-off between privacy and utility. Therefore, a good privacy metric should accurately reflect varying levels of privacy protection, offering a nuanced view of how different privacy settings impact both privacy and utility.

5.1 Black box privacy metrics

A black box privacy metrics offer a method to evaluate the privacy of synthetic data without needing detailed knowledge of the underlying data generation process. These metrics treat the synthetic data generator as a "black box," focusing on the observable outcomes rather than the internal mechanics. By examining the relationships between the original and synthetic data, black-box privacy metrics assess how well privacy is preserved. This approach can include analyzing the statistical properties, patterns, and distributions of the synthetic data in comparison to the original data [12]. Additionally, black-box privacy metrics often involve conducting various privacy attacks, such as re-identification or membership inference attacks, to determine the resilience of the synthetic data against potential breaches [35]. The advantage of black-box metrics lies in their applicability to a wide range of data generation techniques and their ability to provide practical insights into the privacy-utility trade-offs, making them an essential tool for practitioners aiming to evaluate and improve the privacy of synthetic datasets.

In this paper, we choose two black-box privacy metrics, DOMIAS (Density Overfitting Membership Inference Attack with Synthetic Data) [35] and Distance to Closest Record (DCR). DOMIAS effectively measures the likelihood of membership inference attacks, providing a robust indication of privacy risks. DCR provides a straightforward yet powerful indication of how closely synthetic data mimics real data, which can highlight potential re-identification risks.

5.1.1 DOMIAS

The DOMIAS model introduces an approach to Membership Inference Attacks targeting generative models. This model uses density estimation to detect local overfitting in synthetic data, thereby providing enhanced accuracy compared to traditional methods.

The model employs density estimators $p_X(x)$ and $p_G(x)$ to approximate the real data distribution and the synthetic data distribution, respectively. By comparing $p_X(x)$ and

$p_G(x)$, the model detects local overfitting. If $p_G(x)$ is significantly higher than $p_X(x)$ for a given sample x , this suggests that the sample x was likely part of the training set, indicating overfitting.

The membership score for a sample x is computed as the ratio $\frac{p_G(x)}{p_X(x)}$. A higher score implies a greater likelihood that x is part of the training data. The model ensures that this score is invariant to bijective transformations of the data space, maintaining robustness and consistency under different data representations.

The DOMIAS model demonstrates improved accuracy in MIAs by incorporating knowledge of the real data distribution. Its ability to effectively detect local overfitting is crucial for identifying membership of underrepresented samples. Additionally, the model provides a clear metric for balancing data utility and privacy in synthetic data generation.

5.1.2 Distance to Closest Record (DCR)

We introduce another black-box privacy metric, Distance to Closest Record (DCR). In our setup, DCR is defined by the median of the distances between each synthetic data record and the closest real data record. This metric helps to quantify how closely a synthetic dataset mimics the real dataset while maintaining privacy. A lower DCR indicates that the synthetic records are similar to the real records, which is desirable for utility. However, to ensure privacy, DCR should not be too low, as it might indicate that the synthetic data is too similar to the real data, potentially risking privacy breaches. By balancing DCR, we aim to achieve an optimal trade-off between data utility and privacy preservation.

5.2 White-box Privacy Metrics

5.2.1 Integrating TabDDPM to SECMI attack

White-box attacks refer to a type of adversarial attack where the attacker has complete knowledge of the internal workings of the system being targeted. Unlike black-box attack, white-box attacks have access to the model’s architecture, parameters, training data, and any other relevant information. In the context of synthetic data generation and privacy evaluation, white-box attacks are particularly concerning because they allow the attacker to exploit specific vulnerabilities within the data generation process. For example, an attacker might use knowledge of the synthetic data model to craft inputs that reveal sensitive

information about individuals in the original dataset. White-box attacks can include re-identification attempts, where the attacker tries to link synthetic data points back to real individuals, or membership inference attacks, where the goal is to determine whether a particular individual’s data was included in the training set. The complete transparency available in white-box scenarios enables more sophisticated and targeted attacks, making it crucial for privacy-preserving methods to be robust even under these challenging conditions. Evaluating synthetic data against white-box attacks ensures that the privacy protection mechanisms are strong enough to withstand adversaries with significant knowledge and resources.

We chose the Step-wise Error Comparing Membership Inference (SECM) attack [6] as our white-box attack for several reasons. Firstly, diffusion models have demonstrated significant potential and effectiveness in generating high-quality image data. Due to their success with image data, recent research [38, 20] has adapted diffusion models for tabular data generation, leveraging their robust capabilities as foundational models. The SECM attack paper provides a comprehensive and detailed methodology for conducting membership inference attacks specifically on diffusion models, particularly DDIM, in a white-box setting.

A reasonable assumption about membership exposure is that sample x_m from the training (member) set \mathcal{D}_T may exhibit smaller estimation errors at step t compared to samples from the hold-out set \mathcal{D}_H :

$$L_{t,x_m} \leq L_{t,x_h} \tag{5.1}$$

Inspired by deterministic reversing and sampling from diffusion models, approximate ℓ_{t,x_0} with deterministic processes:

$$x_{t+1} = \phi_\theta(x_t) = \sqrt{\bar{\alpha}_{t+1}}f_\theta(x_t) + \sqrt{1 - \bar{\alpha}_{t+1}}\epsilon_\theta(x_t) \tag{5.2}$$

$$x_{t-1} = \psi_\theta(x_t) = \sqrt{\bar{\alpha}_t}f_\theta(x_t) + \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t) \tag{5.3}$$

Where:

$$f_\theta(x_t) = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t)}{\sqrt{\bar{\alpha}_t}} \tag{5.4}$$

We then define $x_t = \Phi_\theta(x_s, t) = \phi_\theta(\dots \phi_\theta(\phi_\theta(x_s, s), s + 1), t - 1)$ and $x_s = \Psi_\theta(x_t, s) = \psi_\theta(\dots \psi_\theta(\psi_\theta(x_t, t), t - 1), s + 1)$. For a given sample x_0 and $\tilde{x}_t = \Phi_\theta(x_0, t)$ at timestep t , the t -error is defined as the approximation posterior estimation error at step t :

$$\tilde{L}_{t,x_0} = \|\psi_\theta(\phi_\theta(\tilde{x}_t, t)) - \tilde{x}_t\|^2 \tag{5.5}$$

The paper [6] later showed that \tilde{L}_{t,x_0} is converged to the learning objective, i.e., $\tilde{L}_{t,x_0} \rightarrow L_{t,x_0}$.

5.3 Synthetic Data utility Metrics

Data synthesization aims to generate artificial data that maintains the statistical properties of the original dataset while preserving privacy. In this context, three important metrics are α -precision [1], β -recall [1], and detection score which help in evaluating the quality and utility of the synthesized data.

5.3.1 α -precision

α -precision measures the probability that a generated sample is supported by the real distribution. This is defined more rigorously as the probability that a synthetic sample resides in the α -support of the real distribution. The α -support is the minimum volume subset of the support of the real distribution that supports a probability mass of α . Mathematically, alpha-precision P_α is given by:

$$P_\alpha = \mathbb{P}(\tilde{X}_g \in S_r^\alpha)$$

where \tilde{X}_g represents the embedded synthetic data, and S_r^α denotes the α -support of the real data distribution \mathbb{P}_r . High alpha-precision indicates that the synthetic data accurately captures the most densely packed probability mass of the real data, ensuring that the synthetic data appears realistic and typical.

5.3.2 β -Recall

β -Recall evaluates the ability of the synthetic data to capture all relevant instances from the original data, focusing on the diversity of the generated samples. It is defined as the fraction of real samples that reside within the β -support of the generative distribution. The β -support is the minimum volume subset of the support of the generative distribution that supports a probability mass of β . Formally, beta-recall R_β is given by:

$$R_\beta = \mathbb{P}(\tilde{X}_r \in S_g^\beta),$$

where \tilde{X}_r represents the embedded real data, and S_g^β denotes the β -support of the synthetic data distribution \mathbb{P}_g . High β -recall signifies that the synthetic data distribution covers the diversity of the real data, ensuring that significant patterns and correlations are preserved.

By balancing α -precision and β -recall, researchers can ensure that synthetic data is both accurate and comprehensive. This balance is crucial for applications requiring both high fidelity and diversity, such as in medical data synthesis, where false positives and missing significant data points can lead to erroneous conclusions.

5.3.3 Detection score

The detection score in the Synthetic data Vault library evaluates the quality of synthetic data by determining how distinguishable it is from real data. This metric involves training a machine learning classifier to differentiate between real and synthetic samples, then measuring its performance, typically using the area under the receiver operating characteristic curve (AUC-ROC). A lower detection score indicates higher quality synthetic data, as it suggests the classifier struggles to tell the two apart. This score is essential for validating synthetic data, ensuring privacy-preserving data generation, and testing the robustness of models trained on synthetic datasets. Proper feature engineering, balanced data, and experimenting with different classifiers are crucial for accurate detection scores.

5.3.4 One-way Marginal/Column Shapes

The "Column Shapes" metric evaluates the statistical similarity between the real and synthetic data for individual columns. This involves comparing the marginal distribution of each column in the real dataset to its counterpart in the synthetic dataset. The closer the synthetic data mimics the statistical properties of the real data, the higher the quality score it receives. This metric ensures that each column in the synthetic dataset follows the same distribution patterns as the original data.

5.3.5 Two-way Marginal/Column Pair Trends

The "Column Pair Trends" metric assesses the relationship between pairs of columns in both the real and synthetic datasets. It focuses on the bivariate distributions, or correlations, between column pairs. By comparing how pairs of columns interact in the synthetic

data to how they do in the real data, this metric provides insight into the preservation of relationships and dependencies within the dataset. High scores in this metric indicate that the synthetic data maintains similar trends and patterns between column pairs as observed in the real data.

Chapter 6

Evaluation Results

6.1 Datasets

In our experiments, we used the six real-world datasets from UCI Machine Learning Repository¹: Adult, Default, Magic, Shoppers, News, and Beijing. The overall statistics of these six datasets are provided in table 6.1. The detailed description of each dataset is as follows:

- **Adult**²: The "Adult Census Income" dataset includes demographic and employment-related features of individuals. The objective is to predict whether an individual's income exceeds \$50,000.
- **Default**³: The "Default of Credit Card Clients Dataset" contains information on default payments, demographic factors, credit data, payment history, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The goal is to predict whether a client will default on payment the following month.
- **Magic**⁴: The "Magic Gamma Telescope" dataset simulates the registration of high-energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using imaging techniques. The objective is to classify high-energy gamma particles in the atmosphere.

¹<https://archive.ics.uci.edu/datasets>

²<https://archive.ics.uci.edu/dataset/2/adult>

³<https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>

⁴<https://archive.ics.uci.edu/dataset/159/magic+gamma+telescope>

- **Shoppers**⁵: The "Online Shoppers Purchasing Intention Dataset" includes information on user webpage visiting sessions. The task is to predict if a user's session ends with a purchase.
- **News**⁶: The "Online News Popularity" dataset contains a heterogeneous set of features about articles published by Mashable over two years. The aim is to predict the number of shares (popularity) on social networks.
- **Beijing**⁷: The "Beijing PM2.5 Data" dataset provides hourly PM2.5 data from the US Embassy in Beijing and meteorological data from Beijing Capital International Airport. The goal is to predict the PM2.5 value.

Dataset	Rows	Numerical	Categorical
Adult	48842	6	9
Default	30000	14	11
Magic	19019	10	1
Shoppers	12330	10	8
news	39644	46	2
Beijing	43824	7	5

Table 6.1: Statistics of datasets

6.2 SECMi Attack Effectiveness

6.2.1 Models

We implemented the SECMi attack in TabDDPM. We adopted the approach outlined in ClavaDDPM [22], which involves encoding categorical columns by labels and mapping encoded categorical columns to numerical spaces rather than using one-hot encoding. This method provides several benefits, including improved scalability and more efficient representation of categorical data in high-dimensional spaces. We denote this approach by, $\text{TabDDPM}_{\text{label}}$ and the original TabDDPM approach by, $\text{TabDDPM}_{\text{one-hot}}$.

⁵<https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset>

⁶<https://archive.ics.uci.edu/dataset/332/online+news+popularity>

⁷<https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data>

	Adult	Default	Magic	Shoppers	News	Beijing	average
Accuracy	0.5019	0.5070	0.5105	0.5134	0.5049	0.5073	0.5075
AUC	0.4971	0.4972	0.5042	0.5086	0.4994	0.5048	0.5019
TP	35.21%	40.90%	29.07%	43.96%	4.26%	25.20%	-
TN	14.99%	9.80%	21.98%	7.38%	46.23%	25.92%	-
FP	35.01%	40.20%	28.02%	42.62%	3.77%	23.69%	-
FN	14.79%	9.10%	20.93%	6.04%	45.74%	25.20%	-

Table 6.2: accuracy and AUC of SECMI attack on TabDDPM_{one-hot} on six datasets. TP, TP, FP, FN are true positive rate, true negative rate, false positive rate, and false negative rate respectively

Metric	Adult	Default	Magic	Shoppers	News	Beijing	Average
Accuracy	0.5022	0.5065	0.5058	0.5109	0.5093	0.5085	0.5072
AUC	0.4969	0.4967	0.4981	0.5104	0.5077	0.5035	0.5022
TP	35.15%	37.37%	43.06%	46.03%	14.31%	23.95%	-
TN	15.07%	13.28%	7.52%	5.07%	36.62%	26.90%	-
FP	34.93%	36.72%	42.48%	44.93%	13.38%	23.10%	-
FN	14.85%	12.63%	06.94%	3.97%	35.69%	26.05%	-

Table 6.3: accuracy and AUC of SECMI attack on TabDDPM_{label} on six datasets. TP, TP, FP, FN are true positive rate, true negative rate, false positive rate, and false negative rate respectively

6.2.2 Findings

The results are significantly lower compared to previous implementations on image data. In these experiments, the test set was evenly split between training data and non-training data. The results are provided in table 6.2, and table 6.3. We observed significantly different results compared to previous implementations on image data. The original SECMI attack reported an AUC of around 0.8, indicating a strong performance. However, when applied to tabular datasets, the AUC values observed were substantially lower, averaging around 0.5019 in table 6.2 across six diverse datasets: Adult, Default, Magic, Shoppers, News, and Beijing. The results were very similar to those obtained from TabDDPM_{label}. The average accuracy was 0.5072 and AUC was 0.5022, which is very close to the accuracy and AUC on TabDDPM_{label}, indicating consistent performance across different encoding approaches.

In Table 6.2, the average accuracy across the six datasets was 0.5075, with individual

dataset accuracies ranging from 0.5019 to 0.5134. This indicates that while the overall model performance is consistent, there are minor variations depending on the specific characteristics of each dataset. Similarly, the accuracy of $\text{TabDDPM}_{\text{label}}$ remained close to 0.5, reflecting a generally stable performance across datasets.

TabDDPM_{label} has a more stable performance than TabDDPM_{one-hot}. The true positive rates (TP) varied considerably, from as low as 4.26% for the News dataset to as high as 43.96% for the Shoppers dataset, highlighting the model’s differing ability to correctly identify positive cases in different contexts. This variability suggests that certain datasets may pose more challenges for the model in terms of recognizing true positives. Likewise, the true negative rates (TN) exhibited a wide range, from 7.38% in the Shoppers dataset to 46.23% in the News dataset. This disparity underscores the model’s varying effectiveness in correctly identifying negative cases, which could be influenced by the unique distributions and characteristics of the datasets.

Interestingly, in contrast to these fluctuations, the true positive rates (TP) in $\text{TabDDPM}_{\text{label}}$ demonstrated a much more stable performance across datasets. This consistency suggests that $\text{TabDDPM}_{\text{label}}$ might have a more balanced approach in handling diverse data, potentially making it a more reliable model in scenarios where consistent identification of positive cases is critical. Overall, these findings underscore the importance of considering dataset-specific factors when evaluating model performance, as they can significantly impact the effectiveness of the model in different applications.

SECMi attack on tabular tends to make False positive predictions. False positive rates (FP) were notably high for some datasets, such as Default (40.20%) and Adult (35.01%), whereas others, like News (3.77%), showed much lower rates. The false negative rates (FN) similarly ranged widely from 6.04% for Shoppers to 45.74% for News. These results suggest that the SECMi attack’s effectiveness is highly dataset-dependent when applied to tabular data, contrasting with the more consistent performance observed on image data.

6.3 SECMi Attack on Image vs. Tabular Data

We reproduced the SECMi attack on image diffusion models, results are shown in table 6.4. In this experiment, we randomly select 50% of the training sample as the training set \mathcal{D}_T and the other 50% as hold-out set \mathcal{D}_H .

The CIFAR-10 and CIFAR-100 datasets are popular benchmark datasets used for evaluating machine learning algorithms, particularly in the field of image recognition. They

Model	CIFAR-10		CIFAR-100	
	Accuracy	AUC	Accuracy	AUC
SECMi	0.82	0.89	0.80	0.87
SECMi _{NNs}	0.88	0.95	0.87	0.94

Table 6.4: SECMi attack on image data

were created by the Canadian Institute for Advanced Research (CIFAR) and are widely used in the computer vision community.

The CIFAR-10 dataset consists of 60000 32x32 color images divided into 10 classes, with 6,000 images per class. The 10 classes represent common objects and include airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each image is labeled with one of these classes, providing a diverse and comprehensive dataset for training and evaluating image classification models.

The CIFAR-100 dataset is an extension of CIFAR-10 and consists of 60000 32x32 color images categorized into 100 classes, with 600 images per class. The 100 classes in CIFAR-100 are grouped into 20 superclasses, with each superclass containing 5 subclasses.

Comparing Table 6.3 and Table 6.2, accuracy and AUCs on image data are significantly larger than accuracy and AUCs on tabular data generated by TabDDPM. Specifically, for the SECMi attack on CIFAR-10 and CIFAR-100 image data (shown in Table 6.4), the accuracy and AUC values are substantially higher.

For CIFAR-10, the SECMi model achieves an accuracy of 0.82 and an AUC of 0.89, while the SECMi_{NNs} variant achieves an even higher accuracy of 0.88 and an AUC of 0.95. Similarly, for CIFAR-100, the SECMi model reaches an accuracy of 0.80 and an AUC of 0.87, with SECMi_{NNs} attaining an accuracy of 0.87 and an AUC of 0.94.

In contrast, the tabular data results for TabDDPM in Table 6.3 and Table 6.2 show lower performance metrics. For instance, the average accuracy across six datasets for the one-hot encoding variant is 0.5075, with an average AUC of 0.5019. For the label encoding variant, the average accuracy is slightly lower at 0.5072, with an average AUC of 0.5022.

These comparisons highlight that the SECMi attack is more effective on image data compared to tabular data generated by TabDDPM, as evidenced by the higher accuracy and AUC values for CIFAR-10 and CIFAR-100 datasets.

To further investigate the reasons behind the differences, we plot relative t-errors of SECMi on tabular and image data in 6.2 and 6.1. The relative t-errors of the hold-out set are defined as $\frac{\Delta_{t,x_{t,\text{hold-out}}}}{\Delta_{t,x_{t,\text{train}}}}$, where $x_{t,\text{hold-out}}$, $x_{t,\text{train}}$ are median t-error at timestep t of the

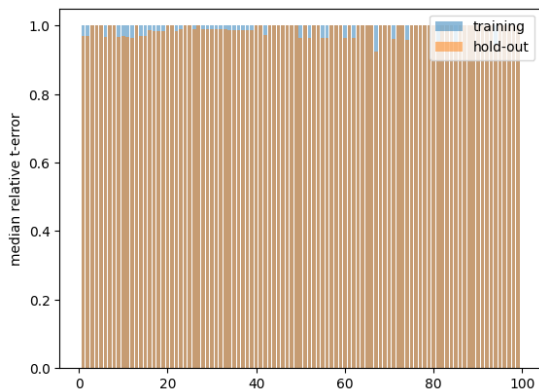


Figure 6.1: median of relative t-errors on TabDDPM on Adult dataset

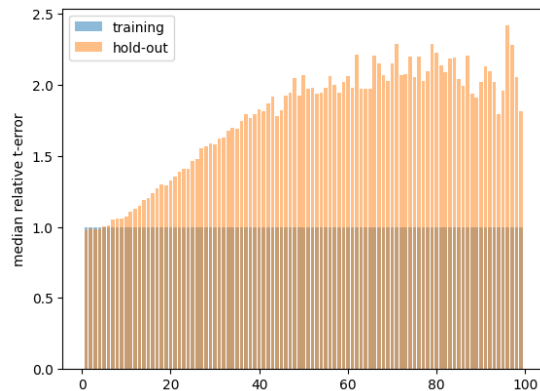


Figure 6.2: median of relative t-errors on image DDIM on CIFAR10 dataset

training set and hold set. The relative t-errors of the training set are defined as $\frac{\Delta_{t,x_t,\text{train}}}{\Delta_{t,x_t,\text{train}}} = 1$ for all timesteps t .

The low performance of the SECMI attack on tabular data can be attributed to the indistinguishability of t-errors between the training and hold-out sets. In Figure 6.1, the median relative t-error for the Adult dataset remains consistently close to 1 for both the training and hold-out sets across all timesteps. There is minimal variation in the relative t-errors between these sets, indicating that the t-errors are difficult to distinguish. This suggests a certain robustness of the TabDDPM model when applied to the Adult dataset, as the prediction errors do not significantly differ between training and hold-out data.

Conversely, Figure 6.2 illustrates the median relative t-error for the CIFAR-10 dataset with image DDIM, revealing a noticeable divergence between the training and hold-out sets as the timestep increases. The relative t-error for the hold-out set increases significantly, reaching values above 2.0 at later timesteps, while the training set remains closer to 1.0. This divergence indicates that the SECMI attack is more effective at distinguishing between the training and hold-out sets for image data compared to tabular data.

Figures 6.3 and 6.4 illustrate the t-error distributions of the training set and hold-out set at timestep 1 for the Adult dataset and the CIFAR-10 dataset, respectively. In comparing these figures, several differences become apparent in terms of distribution shape, frequency, range of errors, and the overlap between training and hold-out sets.

The distribution shape in Figure 6.3, representing the Adult dataset, shows a steep

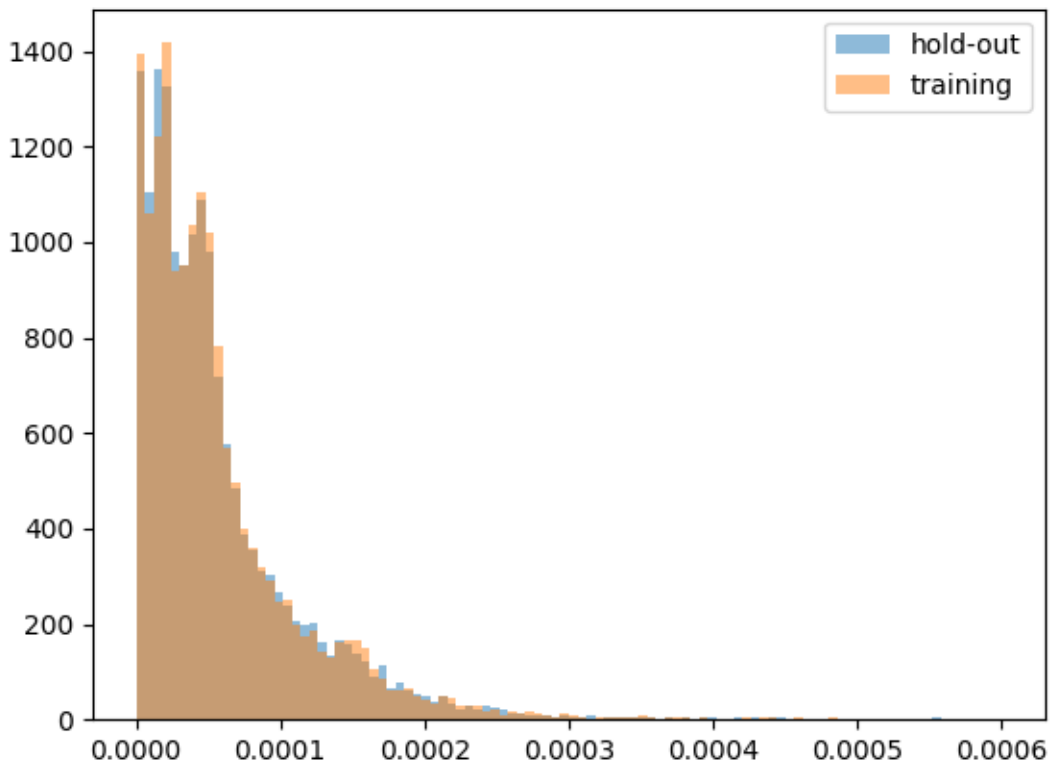


Figure 6.3: t-error distributions of training set and hold-out set at timestep 1 of the first column in Adult dataset

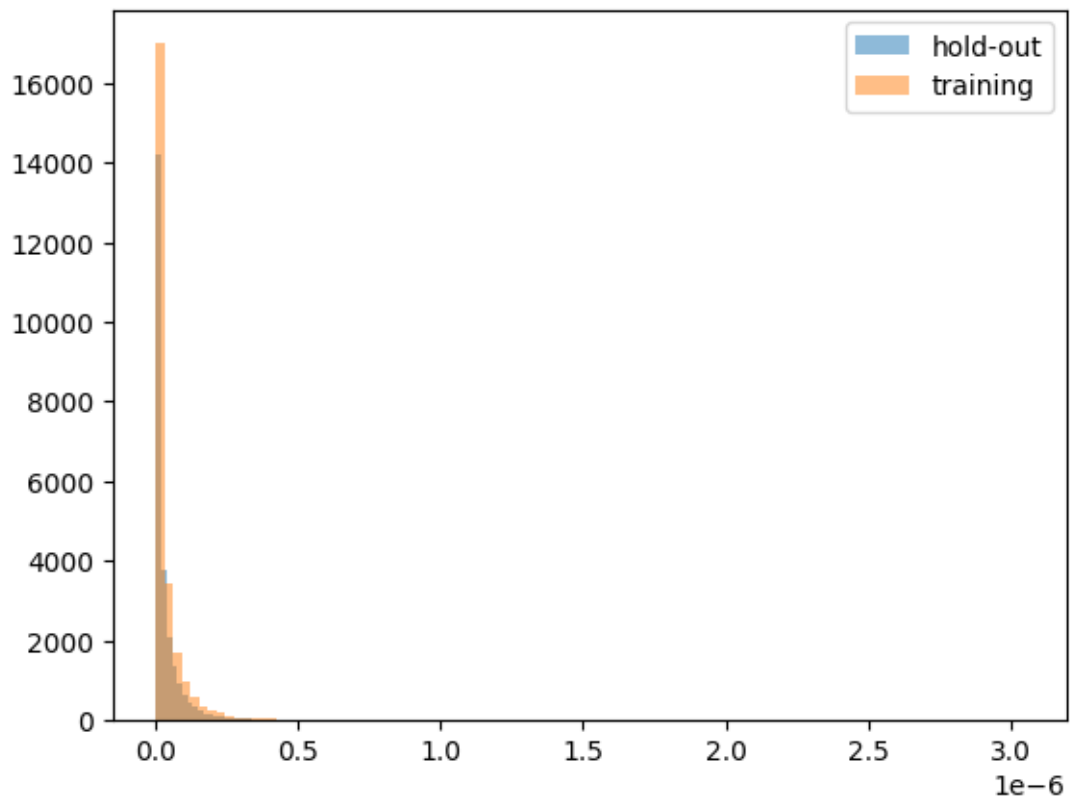


Figure 6.4: t-error distributions of training set and hold-out set at timestep 1 of the first column in CIFAR-10 dataset

decline in the frequency of t-errors as the error value increases. The t-error values are mostly concentrated near zero, with a rapid drop-off as the error values increase. Conversely, Figure 6.4, which depicts the CIFAR-10 dataset, also has a distribution heavily skewed towards zero but exhibits a much more pronounced spike at the very beginning. This indicates that the majority of t-errors for the CIFAR-10 dataset are extremely small, suggesting a more significant proportion of precise predictions.

In terms of frequency, the t-errors in Figure 6.3 for both the training and hold-out sets are relatively close to each other across different t-error values. This suggests similar error distributions for both sets in the Adult dataset. However, in Figure 6.4, the CIFAR-10 dataset shows a significantly higher frequency of very small t-errors for the training set compared to the hold-out set. This discrepancy suggests a more distinct separation between the training and hold-out error distributions for the CIFAR-10 dataset.

Examining the range of errors, Figure 6.3 indicates that the t-errors for the Adult dataset extend up to approximately 0.0006, with most errors concentrated below 0.0002. On the other hand, Figure 6.4 shows that the range of t-errors for the CIFAR-10 dataset extends slightly above $3e-6$. While the errors are generally small in both datasets, the CIFAR-10 dataset presents a broader range of t-errors compared to the Adult dataset.

The overlap between the training and hold-out sets also differs between the two figures. In Figure 6.3, there is a significant overlap between the training and hold-out sets in the Adult dataset, indicating that the t-errors do not significantly differentiate between the two sets. In contrast, Figure 6.4 shows less overlap between the training and hold-out sets in the CIFAR-10 dataset, especially at the smallest t-error values. This suggests that the t-errors can more effectively distinguish between training and hold-out data in the CIFAR-10 dataset compared to the Adult dataset.

Finally, the magnitude of errors differs between the two datasets. The Adult dataset in Figure 6.3 shows larger magnitudes of errors compared to the CIFAR-10 dataset. The errors in the Adult dataset are more spread out, while the CIFAR-10 dataset, as shown in Figure 6.4, has error magnitudes that are much smaller and more tightly clustered around zero. This indicates that the model's predictions are more precise for the CIFAR-10 dataset.

6.4 Utility and Privacy of Synthetic Data

The provided tables display various privacy and utility metrics for different models across multiple datasets, including Adult, Default, Beijing, Magic, News, and Shoppers. These

Model	DCR	DOMIAS (AUC)	1-way marginal	2-way marginal	α -precision	β -recall	Detection Score
TabDDPM _{label}	0.11	0.50	98.75%	97.73%	0.97	0.45	0.99
TabDDPM _{one-hot}	0.097	0.50	98.95%	98.42%	0.95	0.50	0.96
PrivBayes, $\epsilon = 0.1$	1.82	0.50	62.10%	64.42%	0.17	0.01	0.01
PrivBayes, $\epsilon = 0.5$	1.65	0.50	66.46%	70.84%	0.34	0.04	0.03
PrivBayes, $\epsilon = 1$	1.45	0.50	71.77%	74.53%	0.35	0.05	0.16
PrivBayes, $\epsilon = 1.5$	1.21	0.50	73.73%	80.73%	0.51	0.13	0.26

Table 6.5: privacy and utility metrics for different models on the Adult dataset

Model	DCR	DOMIAS (AUC)	1-way marginal	2-way marginal	α -precision	β -recall	Detection Score
TabDDPM _{label}	0.05	0.52	98.33%	93.10%	0.98	0.54	0.94
TabDDPM _{one-hot}	2.61	0.58	64.56%	76.06%	0.96	0.50	0.98
PrivBayes, $\epsilon = 0.1$	2.70	0.53	36.79%	68.26%	0.01	0.00	0.00
PrivBayes, $\epsilon = 0.5$	2.64	0.55	40.06%	70.68%	0.05	0.00	0.00
PrivBayes, $\epsilon = 1$	2.53	0.54	44.26%	73.40%	0.11	0.01	0.00
PrivBayes, $\epsilon = 1.5$	2.51	0.55	47.49%	74.97%	0.17	0.02	0.01

Table 6.6: privacy and utility metrics for different models on the Default dataset

metrics include Disclosure Risk (DCR), DOMIAS (AUC), 1-way and 2-way marginals, α -precision, β -recall, and Detection Score. This analysis focuses on the observed trends and relationships between these metrics, especially concerning the quality of synthetic data.

Tables 6.5 to 6.10 present a detailed comparison of different models (TabDDPM and PrivBayes with varying epsilon values) across the specified datasets. **A notable observation is the ineffectiveness of DOMIAS as a differentiator across varying factors** such as different datasets, different training data sizes, and different model parameters (e.g., PrivBayes’s epsilon). Across all datasets, DOMIAS (AUC) values remain relatively constant, fluctuating minimally around 0.50. For instance, in Table 6.5 (Adult dataset), DOMIAS values for all models are either 0.50 or 0.51, suggesting that DOMIAS is not sensitive to the changes in model configurations or dataset variations.

DCR values exhibit significant variation across different models and datasets. For example, in Table 6.6 (Default dataset), TabDDPM with one-hot encoding has a DCR of 2.61, whereas TabDDPM with label encoding has a much lower DCR of 0.05. This trend is consistent across other datasets as well; for example, in Table 6.9 (News dataset), TabDDPM with label encoding has a DCR of 0.52 compared to 3.49 for the one-hot encoding. This indicates that higher DCR values often correlate with lower quality of synthetic data. PrivBayes with higher epsilon values (indicating less privacy) generally shows higher DCR, suggesting poorer quality of synthetic data in terms of privacy preservation.

The 1-way and 2-way marginal percentages provide insights into the utility of the synthetic data. Higher percentages indicate better utility. For example, in Table 6.7 (Beijing dataset), TabDDPM with label encoding achieves 99.25% (1-way) and 99.26%

Model	DCR	DOMIAS (AUC)	1-way marginal	2-way marginal	α -precision	β -recall	Detection Score
TabDDPM _{label}	1.00	0.53	99.25%	99.26%	0.99	0.62	0.98
TabDDPM _{one-hot}	1.77	0.50	53.93%	72.38%	0.01	0.02	0.05
PrivBayes, $\epsilon = 0.1$	1.47	0.51	65.96%	82.56%	0.07	0.20	0.00
PrivBayes, $\epsilon = 0.5$	1.37	0.51	70.21%	85.22%	0.15	0.27	0.01
PrivBayes, $\epsilon = 1$	1.33	0.50	74.44%	86.26%	0.19	0.29	0.03
PrivBayes, $\epsilon = 1.5$	1.23	0.50	79.08%	88.89%	0.29	0.30	0.07

Table 6.7: privacy and utility metrics for different models on the Beijing dataset

Model	DCR	DOMIAS (AUC)	1-way marginal	2-way marginal	α -precision	β -recall	Detection Score
TabDDPM _{label}	0.06	0.52	99.10%	97.67%	0.99	0.59	0.99
TabDDPM _{one-hot}	0.07	0.52	98.78%	98.60%	0.99	0.47	0.99
PrivBayes, $\epsilon = 0.1$	0.77	0.50	60.65%	79.17%	0.10	0.00	0.05
PrivBayes, $\epsilon = 0.5$	0.74	0.50	63.84%	79.86%	0.15	0.01	0.10
PrivBayes, $\epsilon = 1$	0.71	0.51	66.82%	80.90%	0.19	0.02	0.15
PrivBayes, $\epsilon = 1.5$	0.67	0.50	70.14%	81.88%	0.27	0.04	0.24

Table 6.8: privacy and utility metrics for different models on the Magic dataset

Model	DCR	DOMIAS (AUC)	1-way marginal	2-way marginal	α -precision	β -recall	Detection Score
TabDDPM _{label}	0.52	0.51	98.19%	98.46%	0.93	0.40	0.95
TabDDPM _{one-hot}	3.49	0.51	6.19%	1.59%	0.00	0.00	0.00
PrivBayes, $\epsilon = 0.1$	2.95	0.50	39.21%	87.86%	0.00	0.00	0.65
PrivBayes, $\epsilon = 0.5$	2.92	0.50	40.76%	88.10%	0.00	0.00	0.00
PrivBayes, $\epsilon = 1$	2.86	0.49	42.19%	88.39%	0.00	0.00	0.00
PrivBayes, $\epsilon = 1.5$	2.81	0.50	44.86%	88.67%	0.01	0.00	0.00

Table 6.9: privacy and utility metrics for different models on the News dataset

Model	DCR	DOMIAS (AUC)	1-way marginal	2-way marginal	α -precision	β -recall	Detection Score
TabDDPM _{label}	0.05	0.58	98.09%	97.45%	0.96	0.66	0.94
TabDDPM _{one-hot}	0.15	0.52	97.21%	94.46%	0.92	0.55	0.88
PrivBayes, $\epsilon = 0.1$	2.23	0.50	41.06%	56.76%	0.04	0.00	0.00
PrivBayes, $\epsilon = 0.5$	2.10	0.50	51.13%	62.28%	0.15	0.03	0.03
PrivBayes, $\epsilon = 1$	2.07	0.51	51.56%	62.87%	0.12	0.02	0.01
PrivBayes, $\epsilon = 1.5$	1.95	0.50	53.55%	65.34%	0.19	0.05	0.04

Table 6.10: privacy and utility metrics for different models on the Shoppers dataset

(2-way), indicating high utility. In contrast, PrivBayes with $\epsilon = 0.1$ shows much lower utility (53.93% for 1-way and 72.38% for 2-way). Similarly, in Table 10 (Shoppers dataset), TabDDPM with label encoding has 98.09% (1-way) and 97.45% (2-way), whereas PrivBayes with $\epsilon = 0.1$ has 41.06% (1-way) and 56.76% (2-way). This suggests that TabDDPM generally produces higher quality synthetic data compared to PrivBayes, especially when preserving marginal distributions.

α -precision and β -recall metrics further illustrate the performance differences. High α -precision indicates a higher proportion of correctly predicted positives, while high β -recall indicates a higher proportion of true positives among all actual positives. For instance, in Table 6.10 (Shoppers dataset), TabDDPM with label encoding achieves high α -precision (0.96) and β -recall (0.66), while PrivBayes with $\epsilon = 0.1$ achieves much lower values (0.04 for α -precision and 0.00 for β -recall). This further supports the superior performance of TabDDPM in generating synthetic data that balances privacy and utility.

The detection score measures the ability to distinguish between real and synthetic data. A lower detection score indicates better privacy as it is harder to distinguish synthetic data from real data. The tables show that models like TabDDPM with label encoding often have high detection scores (indicating lower privacy risk), while PrivBayes with lower epsilon values (indicating higher privacy) sometimes have lower detection scores, suggesting better privacy preservation but at the cost of utility. For example, in Table 8 (Magic dataset), TabDDPM with label encoding achieves a detection score of 0.99 compared to 0.00 for PrivBayes with $\epsilon = 0.1$.

Chapter 7

Conclusion

Our study provides critical insights into the evaluation of privacy metrics for synthetic tabular data, highlighting the varying effectiveness of these metrics across different datasets and data generation models. While DOMIAS demonstrated limited sensitivity in diverse contexts, DCR emerged as a more reliable metric for assessing the similarity between synthetic and real data. Additionally, despite the success of white-box attacks like SECMI on image data within diffusion models, its performance on tabular data was notably less effective. Diffusion models like TabDDPM consistently outperformed traditional models like PrivBayes in terms of data utility; however, this improved utility often comes at the expense of reduced privacy.

Moving forward, it is imperative for future research to focus on the development of more adaptive and robust privacy metrics that can effectively address the complexities inherent in various datasets and synthetic data generation techniques. Enhancing these metrics will be essential for the safe and efficient application of synthetic data across a wide range of fields, fostering innovation while upholding rigorous privacy standards.

References

- [1] Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models, 2022.
- [2] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.
- [3] 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*, pages 343–362, 2020.
- [4] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.
- [5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [6] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, pages 8717–8730, 2023.
- [7] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

- [8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.
- [9] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data. *arXiv preprint arXiv:2211.10459*, 2022.
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [11] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.
- [12] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–6, 2019.
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [15] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.
- [16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [17] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1895–1912, 2019.

- [18] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.
- [19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [20] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.
- [21] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, volume 2018, pages 1–15, 2018.
- [22] Wei Pang, Masoumeh Shafeinejad, Lucy Liu, and Xi He. Clavaddpm: Multi-relational data synthesis with cluster-guided diffusion models. *arXiv preprint arXiv:2405.17724*, 2024.
- [23] Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023.
- [24] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [25] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [27] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

- [28] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.
- [29] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [30] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [33] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [34] Shuai Tang, Zhiwei Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attacks on diffusion models via quantile regression. *arXiv preprint arXiv:2312.05140*, 2023.
- [35] Boris Van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. *arXiv preprint arXiv:2302.12580*, 2023.
- [36] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.
- [37] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020.
- [38] Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*, 2023.

- [39] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), oct 2017.
- [40] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model (technical report). *arXiv preprint arXiv:1801.01594*, 2018.
- [41] Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. PrivSyn: Differentially private data synthesis. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 929–946. USENIX Association, August 2021.