On the Inherent Privacy Properties of Discrete Denoising Diffusion Models

Anonymous authors
Paper under double-blind review

Abstract

Privacy concerns have led to a surge in the creation of synthetic datasets, with diffusion models emerging as a promising avenue. Although prior studies have performed empirical evaluations on these models, there has been a gap in providing a mathematical characterization of their privacy-preserving capabilities. To address this, we present the pioneering theoretical exploration of the privacy preservation inherent in *discrete diffusion models* (DDMs) for discrete dataset generation. Focusing on per-instance differential privacy (pDP), our framework elucidates the potential privacy leakage for each data point in a given training dataset, offering insights into how the privacy loss of each point correlates with the dataset's distribution. Our bounds also show that training with s-sized data points leads to a surge in privacy leakage from $(\epsilon, \mathcal{O}(\frac{1}{s^2\epsilon}))$-pDP to $(\epsilon, \mathcal{O}(\frac{1}{s\epsilon}))$-pDP of the DDM during the transition from the pure noise to the synthetic clean data phase, and a faster decay in diffusion coefficients amplifies the privacy guarantee. Finally, we empirically verify our theoretical findings on both synthetic and real-world datasets.

1 Introduction

Discrete tabular or graph datasets with categorical attributes are prevalent in many privacy-sensitive domains (Vatsalan et al., 2013; Pourhabibi et al., 2020; Li et al., 2021; Shwartz-Ziv & Armon, 2022; Borisov et al., 2022), including finance (Clements et al., 2020; Wang et al., 2021), e-commerce (Ahmed et al., 2017; Zhang et al., 2019), and medicine (Duvenaud et al., 2015; Schork, 2015; Ulmer et al., 2020). For instance, medical researchers often collect patient data, such as race, gender, and medical conditions, in a discrete tabular form. However, using and sharing data in these domains carry the risk of revealing personal information (Abay et al., 2019). Studies have shown that it is possible to re-identify individuals in supposedly de-identified healthcare data (McGuire & Gibbs, 2006; El Emam et al., 2011). To address these types of concerns, publishing synthetic datasets with privacy guarantees has been proposed as a way to protect sensitive information and to reduce the risk of privacy leakage (Choi et al., 2017; Patel et al., 2018; Tucker et al., 2020; DuMont Schütte et al., 2021).

Previous research has explored discrete synthetic database releasing methods (Zhou et al., 2009b; Blum et al., 2013; Li et al., 2023). Many of these methods employ data anonymization techniques (Sweeney, 2002; Li et al., 2006; Liu & Terzi, 2008; Lu et al., 2012) or focus on private statistics/statistical models (Sala et al., 2011; Jorgensen et al., 2016; Balog et al., 2018; Harder et al., 2021). In the former category, k-anonymization (Sweeney, 2002) directly works on anonymizing categorical features but it can be vulnerable to the attackers with background knowledge (Machanavajjhala et al., 2007). Alternatively, methods using private statistics or models concentrate on sharing specific private statistics (Harder et al., 2021) or privatizing model parameters (Hardt et al., 2012; Zhang et al., 2017). However, these techniques can sometimes misrepresent the original distribution or reduce sample quality by adding noise directly to model parameters.

Neural network (NN)-based generative models have been leveraged in various domains on account of their ability in learning underlying distributions (Austin et al., 2021). Recently, discrete diffusion models (DDMs) (Hoogeboom et al., 2021; Austin et al., 2021; Campbell et al., 2022; Gu et al., 2022; Vignac et al., 2022), as a typical representative of diffusion models (DMs), have emerged as a powerful class of generative models

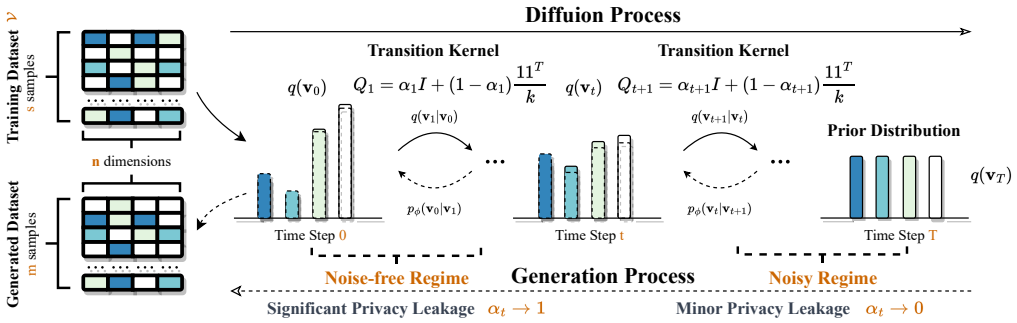

Figure 1: **An Illustration of Discrete Diffusion Models (DDMs).**

for discrete data and demonstrate great potential to generate samples with striking performance (Haefeli et al., 2022; Zheng et al., 2023). DDMs are latent variable generative models that employ both a forward and reverse Markov process (See Fig. 1). In the forward diffusion process, each discrete sample is gradually corrupted with dimension-wise independent noise. This is often implemented through the use of progressive transition kernels, which yields not only high fidelity-diversity trade-offs but also robust training objectives (Dhariwal & Nichol, 2021). On the other hand, the reverse process learns denoising neural networks that aim to predict the noise and reconstruct the original sample. Despite the impressive performance of DDMs, it is still unclear whether DDMs trained on sensitive datasets can be safely used to generate synthetic samples.

Efforts have empirically examined the privacy implications of DMs. While previous literature suggests that DMs generate synthetic training data to address privacy concerns (Jahanian et al., 2021; Carr, 2022), recent studies have shown that DMs may not be suitable for releasing private synthetic data. Specifically, Wu et al. (2022); Hu & Pang (2023) conduct membership inference attacks on DMs for text-to-image tasks and demonstrate that membership inference poses a severe threat in diffusion-based generation. Besides, studies show that DMs can memorize training samples (Somepalli et al., 2022; Carlini et al., 2023). Although there exist practical observations for privacy properties of DMs, there is limited research aimed at mathematically characterizing the privacy guarantees of data generated by DMs. Moreover, understanding privacy guarantees may guide practitioners to determine whether additional mechanisms, such as DP-SGD (Abadi et al., 2016), PATE (Papernot et al., 2016), should be incorporated to meet practical privacy requirements.

Differential privacy (DP) (Dwork et al., 2006; 2014), the most commonly used *algorithm-centric* framework to characterize the privacy guarantee of an algorithm, is derived from the worst-case dataset. However, in the context of synthetic data sharing, the characterization of privacy leakage is about the synthetic dataset (the algorithm output) rather than the generative model (the algorithm itself), and the learned data distribution to generate synthetic data strongly depends on the empirical distribution of the data points used for training. Therefore, a privacy guarantee that may incorporate *the distributional characteristics of data points* in the given training dataset may offer a far more accurate privacy characterization than the worst-case analysis. Such data-dependent analysis may help practitioners learn which data points in the training dataset tend to introduce privacy leakage concerns in the generation process and thus design the relevant protection strategy.

In this paper, we take the first step to analyze the privacy guarantees of DDMs for a fixed training dataset. Specifically, we leverage the data-dependent privacy framework termed per-instance differential privacy (pDP), which is defined upon an instance in a fixed training dataset as outlined by (Wang, 2019). The analysis of pDP allows for a fine-grained characterization of the potential privacy leakage of each data point in the training set. This offers data curators a better understanding of the sensitivity of training data.

Our analysis considers a DDM trained on s samples and generates m samples, and we keep track of the privacy leakage in each generation step. We prove that as the data generation step transits from $t = T$ (noisy regime) to $t = 0$ (noise-free regime), the privacy leakage increases from $(\epsilon, \mathcal{O}(\frac{m}{s^2\epsilon(1-e^{-\epsilon})}))$-pDP to $(\epsilon, \mathcal{O}(\frac{m}{s\epsilon(1-e^{-\epsilon})}))$-pDP where the data-dependent term is hidden in the big-$\mathcal{O}$ notation. Consequently, the

final few generation steps ($\alpha_t \to 1$ in Fig. 1) dominate the main privacy leakage in DDMs. Further, our analysis demonstrates that the privacy bound $\mathcal{O}(1/s)$ is tight when $m = 1$, emphasizing the inherent weak privacy guarantee of DDMs. Moreover, faster decay in diffusion coefficients yields better privacy preservation. Both synthetic and real dataset evaluations validate our theoretical findings.

For the data-dependent part, we develop a practical algorithm to estimate the privacy leakage of each data point in real-world datasets according to our pDP bounds. We evaluate the data-dependent part by removing the most sensitive data points (according to our data-dependent privacy parameters) from the dataset to train a DDM, and then evaluating the ML models trained based on the synthetic dataset generated by the DDM. Interestingly, we observe that the ML models obtained after a part of data removal can even outperform others without such data removal. We attribute this to the fact that the removed data points are likely outliers which may be actually not good for ML models to learn from. This illustrates another potentially valuable usage of our data-dependent analysis.

To avoid any confusion, we provide several important explanations for considering pDP in our work. pDP, tailored to the training set, offers data curators a more accurate and fine-grained estimation of the potential privacy leakage of each data point, compared to DP which studies the worst case and keeps agnostic to the dataset (Wang, 2019). However, it is crucial to understand that pDP is not a replacement for DP. Direct application of data-dependent sensitivity for noise addition is not permissible for ensuring privacy, as the added noise may leak private information due to its data dependency. Data-dependent methods such as smooth sensitivity (Nissim et al., 2007) and propose-test-release (Dwork & Lei, 2009) may be employed, while they are beyond the scope of this paper. Our analysis is to provide insights into the inherent privacy afforded by DDMs, and to guide data curators in assessing the privacy risks associated with different parts of the dataset. We are not to develop an algorithm to match a certain privacy budget as the goal. Given this purpose, pDP is a more suitable metric than DP. In practice, the pDP assessment is expected to be kept confidential and used by the data curator herself to understand the dataset and evaluate how the potential privacy leakage if one uses DDMs to generate synthetic datasets.

1.1 More Related Work

A significant amount of research has been conducted on the subject of publishing privacy sensitive data (Ji et al., 2014; Baraheem & Yao, 2022). As of now, traditional non-deep learning techniques for preserving privacy while generating discrete data can be broadly classified into two categories: **(a) Data anonymization-based approaches.** These methods employ a variety of techniques to directly sanitize data to prevent easy re-identification (Abay et al., 2019). One most popular framework is termed k-anonymity (Sweeney, 2002) that requires each record is indistinguishable from at least $k-1$ other records with respect to certain identifying attributes. Several extensions of this framework have been proposed in (LeFevre et al., 2005; Aggarwal et al., 2005; Machanavajjhala et al., 2007; Li et al., 2006; Truta & Vinay, 2006; Machanavajjhala et al., 2008; Liu & Terzi, 2008; Liang & Samavi, 2020). However, these methods are typically prone to various privacy attacks (Machanavajjhala et al., 2007). **(b) Methods based on statistical models or private statistics.** Barak et al. (2007) employed Fourier decomposition and prior knowledge to release low-dimensional data projections. Zhou et al. (2009b) proposed a database compression procedure based on low-rank random affine transformations and publish low-dimensional data. Other works along this line include (Liu et al., 2005; Zhou et al., 2009a; Ding et al., 2011; Cormode et al., 2011; Kenthapadi et al., 2012; Cormode et al., 2012). Note that these works can work for both discrete and continuous data. Furthermore, Balog et al. (2018) introduced a framework employing kernel mean embeddings (Smola et al., 2007) in Reproducing Kernel Hilbert Space, and ensuring privacy by using synthetic data approximations to enable safe data release. Nevertheless, these methods usually suffer from poorly generated sample qualities. With regard to this, establishing NN-based private models is a promising way to enhance sample qualities due to the great expressive power of deep networks.

Hitherto, there are studies on NN-based private models but few analyze the inherent privacy of the model itself. In (Lin et al., 2021), it was shown that a vanilla GAN trained on s samples inherently satisfies a weak $(\epsilon, \mathcal{O}(\frac{m}{s\epsilon}))$-DP guarantee when releasing m samples. In this work, our results demonstrate that DDMs provide weak privacy guarantees in the same order as GANs. But note that Lin et al. (2021) did not provide a data-dependent bound. Their bounds are in the order form and cannot be explicitly computed from data

curator's side for a given training dataset. Because of such weak inherent privacy there were efforts to bring additional privacy techniques into the model, such as DP-SGD (Abadi et al., 2016). Xie et al. (2018) proposed DPGAN that integrates modified DP-SGD in WGAN to ensure privacy for GAN-generated samples. Dockhorn et al. (2022) applied DP-SGD to privatize model parameters in continuous DMs for image data without analyzing the inherent privacy of DMs. Recently, Ghalebikesabi et al. (2023) have showed that fine-tuning a pre-trained diffusion model with DP-SGD can generate verifiable private synthetic data for the dataset used for fine-tuning.

2 Preliminaries

We start by introducing notations and concepts for analysis. Let $[n] = \{1, 2, ..., n\}$ and $\mathcal{X}^n$ represent an n-dimensional discrete space with each dimension having k categories, i.e. $\mathcal{X}^n := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ with $\mathcal{X}_i = [k], i \in [n]$. We assume that training datasets $\mathcal{V}$ reside in $\mathcal{X}^n$, implying samples are vector-valued data of n entries, each from one of the k categories. Although we assume consistent categories across columns, our analysis can account for datasets with varied category counts using the maximum category count.

Per-instance Differential Privacy. DP (Dwork et al., 2006; 2014) is a de-facto standard to quantify privacy leakage. We adapt DP definition for specific adjacent datasets, introducing per-instance DP:

Definition 1 ((ϵ, δ)-Per-instance Differential Privacy (pDP) (Wang, 2019))**.** Let $\mathcal{V}_0$ be a training dataset, $\mathbf{v}^* \in \mathcal{V}_0$ be a fixed point and $\mathcal{M}$ be a randomized mechanism. Define adjacent dataset $\mathcal{V}_1 = \mathcal{V}_0 \backslash \{\mathbf{v}^*\}$. We say $\mathcal{M}$ satisfies (ϵ, δ)-pDP with respect to $(\mathcal{V}_0, \mathbf{v}^*)$ if for all measurable set $\mathcal{O} \subset range(\mathcal{M})$, $\{i, j\} = \{0, 1\}$:

$$\mathcal{P}(\mathcal{M}(\mathcal{V}_i) \in \mathcal{O}) \leq e^{\epsilon} \mathcal{P}(\mathcal{M}(\mathcal{V}_j) \in \mathcal{O}) + \delta. \tag{1}$$

It is important to highlight that pDP is uniquely defined for a specific dataset-data point pair. This capability is crucial for understanding the privacy leakage of the given dataset, as elaborated in Sec. 4. Additionally, by taking the supremum over all conceivable datasets $\mathcal{V}_0$ and points $\mathbf{v}^*$, we can obtain DP from pDP when considering model releasing scenario (Theorem E.1). A more comprehensive discussion of the DP guarantees associated with DDMs is provided in Appendix. E.

Discrete Diffusion Models. DDMs (Hoogeboom et al., 2021; Austin et al., 2021; Vignac et al., 2022; Haefeli et al., 2022) are diffusion models that can generate categorical data. Let $\mathbf{v}_t$ denote the data random variable at time t. The forward process involves gradually corrupting data with the noising Markov chain q, according to $q(\mathbf{v}_{1:T}|\mathbf{v}_0) = \prod_{t=1}^{T} q(\mathbf{v}_t|\mathbf{v}_{t-1})$, where $\mathbf{v}_{1:T} = \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T$. On the other hand, the reverse process, $p_\phi(\mathbf{v}_{0:T}) = p(\mathbf{v}_T) \prod_{t=1}^{T} p_\phi(\mathbf{v}_{t-1}|\mathbf{v}_t)$, gradually reconstructs the datasets starting from a prior $p(\mathbf{v}_T)$. The denoising neural network (NN) learns $p_\phi(\mathbf{v}_{t-1}|\mathbf{v}_t)$ by optimizing the ELBO, which comprises three loss terms: the reconstruction term (L_r), the prior term (L_p), and the denoising term (L_t) , represented in the following equation (Ho et al., 2020):

$$\underbrace{\mathbb{E}_{q(\mathbf{v}_1|\mathbf{v}_0)}[\log p_\phi(\mathbf{v}_0|\mathbf{v}_1)]}_{\text{Reconstruction Term } L_r} - \underbrace{D_{\mathrm{KL}}(q(\mathbf{v}_T|\mathbf{v}_0)\|p_\phi(\mathbf{v}_T))}_{\text{Prior Term } L_p} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\mathbf{v}_t|\mathbf{v}_0)}[D_{\mathrm{KL}}(q(\mathbf{v}_{t-1}|\mathbf{v}_t, \mathbf{v}_0)\|p_\phi(\mathbf{v}_{t-1}|\mathbf{v}_t))]}_{\text{Denoising Term } L_t}. \tag{2}$$

Specifically, the **forward process** can be described by a series of transition kernels $\{Q_t^i\}_{t\in[T], i\in[n]}$ where for any entry $\mathbf{v}^i$, $[Q_t^i]_{lh} = q(\mathbf{v}_t^i = h|\mathbf{v}_{t-1}^i = l)$ represent the probability of a jump from category l to h on the i-th entry at time t. Since for each entry i the number of categories is the same, we can rely on the same transition kernels for all dimensions and use Q_t instead of Q_t^i. Let $\overline{Q}_t = Q_1 Q_2 ... Q_t$ denote the accumulative transition matrix from time 1 to time t. We use a uniform prior distribution $p(\mathbf{v}_T)$. The corresponding doubly stochastic matrices is determined by a series of important parameters termed **diffusion coefficients** ($\{\alpha_t, t \in [T]|\alpha_t \in (0, 1)\}$) which control the transition rate from original distribution to uniform measure. Specifically, define $Q_t = \alpha_t I + (1 - \alpha_t)\frac{\mathbb{1}\mathbb{1}^T}{k}$ and then $\bar{Q}_t = \bar{\alpha}_t I + (1 - \bar{\alpha}_t)\frac{\mathbb{1}\mathbb{1}^T}{k}$ where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t$. In the **reverse process**, denoising networks are leveraged to predict $p_\phi(\mathbf{v}_{t-1}|\mathbf{v}_t)$ in hope of approximating $q(\mathbf{v}_{t-1}|\mathbf{v}_t, \mathbf{v}_0)$. In practice, instead of directly predicting $p_\phi(\mathbf{v}_{t-1}|\mathbf{v}_t)$, denoising networks are learned to predict a clean data $\mathbf{v}_0$ at time 0 with a noisy $\mathbf{v}_t$ as input, i.e. $p_\phi(\mathbf{v}_0|\mathbf{v}_t)$. To train the denoising network,

one needs to sample noisy points from $q(\mathbf{v}_t|\mathbf{v}_0)$, and feed them into the denoising network ϕ_t and obtain $p_\phi(\mathbf{v}_0|\mathbf{v}_t)$. Specifically, we adopt

$$L_{\text{train}} = D_{\text{KL}}(q(\mathbf{v}_0|\mathbf{v}_t)||p_\phi(\mathbf{v}_0|\mathbf{v}_t)) = \frac{1}{|\mathcal{V}|}\sum_{\mathbf{v}_0\in\mathcal{V}} \mathbb{E}_{\mathbf{v}_t\sim q(\mathbf{v}_t|\mathbf{v}_0)}\left[\sum_{i=1}^{n} L_{\text{CE}}(\mathbf{v}_0^i, p_\phi(\mathbf{v}_0^i|\mathbf{v}_t))\right] \tag{3}$$

This loss serves as the basis for our later sufficient training Assumption 1. In the generation process, we need to bridge the connection of $p_\phi(\mathbf{v}_{t-1}|\mathbf{v}_t)$ and $p_\phi(\mathbf{v}_0|\mathbf{v}_t)$, which in practice depends on a dimension-wise conditional independence condition (Vignac et al., 2022):

$$p_\phi(\mathbf{v}_{t-1}|\mathbf{v}_t) = \prod_{i\in[n]} p_\phi(\mathbf{v}_{t-1}^i|\mathbf{v}_t) = \prod_{i\in[n]}\sum_{l\in\mathcal{X}_i} q(\mathbf{v}_{t-1}^i|\mathbf{v}_t, \mathbf{v}_0^i = l)p_\phi(\mathbf{v}_0^i = l|\mathbf{v}_t). \tag{4}$$

Other Notations. Given two samples $\mathbf{v}$ and $\tilde{\mathbf{v}}$, let $\bar{\omega}(\mathbf{v}, \tilde{\mathbf{v}})$ represent the count of differing entries, i.e., $\bar{\omega}(\mathbf{v}, \tilde{\mathbf{v}}) = \#\{i|\mathbf{v}^i \neq \tilde{\mathbf{v}}^i, i \in [n]\}$. For $\eta \in [n]$ and $\mathbf{v} \in \mathcal{V}_1$, define $N_\eta(\mathbf{v}) = |\{\mathbf{v}' \in \mathcal{V}_1 : \bar{\omega}(\mathbf{v}, \mathbf{v}') \leq \eta\}|$ and $\mathcal{V}_1^{i|l} = \{\mathbf{v} \in \mathcal{V}_1|\mathbf{v}^i = l\}$ the set of data points with a fixed-valued entry. We use $\mathcal{D}_{\text{KL}}(\cdot\|\cdot)$ and $\|\cdot\|_{TV}$ for KL-divergence and total variation. Let $\mu_t^+ = \frac{1+(k-1)\alpha_t}{k}$ and $\mu_t^- = \frac{1-\alpha_t}{k}$ represent one-step transition probabilities to the same and different states respectively at time t while $\bar{\mu}_t^+ = \frac{1+(k-1)\bar{\alpha}_t}{k}$ and $\bar{\mu}_t^- = \frac{1-\bar{\alpha}_t}{k}$ are the accumulated transition probabilities. Transition probability ratios are defined as $R_t = \frac{\mu_t^+}{\mu_t^-}$ and $\bar{R}_t = \frac{\bar{\mu}_t^+}{\bar{\mu}_t^-}$. A larger ratio indicates a higher likelihood of maintaining the same feature category in the diffusion process. Moreover, define $(\cdot)_+ = \max\{\cdot, 0\}$.

3 Main Results

3.1 Inherent Privacy Guarantees of DDMs

First, we define the mechanism under analysis. Let $\mathcal{M}_t(\mathcal{V}; m)$ represent the mechanism where, for an input dataset $\mathcal{V}$, it outputs m samples generated at time t using the DDM's generation process. Specifically, $\mathcal{M}_0(\mathcal{V}; m)$ signifies the final generated dataset by DDM. In the paper, we focus on the behavior of $\mathcal{M}_t$ in the generation process. Below, we outline the assumptions:

Assumption 1 (Sufficient training of ϕ). Given dataset $\mathcal{V}$, let $\mathbf{v}_0$ denote the predicted random variables at time 0. Let ϕ denote denoising NNs trained on dataset $\mathcal{V}$. We say Assumption 1 is satisfied if there exist small constants $\gamma_t > 0$ such that $\forall \mathbf{v}_t \in \mathcal{X}^n$:

$$\mathcal{D}_{\text{KL}}(q(\mathbf{v}_0^i|\mathbf{v}_t)\|p_\phi(\mathbf{v}_0^i|\mathbf{v}_t)) \leq \gamma_t, \forall i \in [n], \forall t \in [T]. \tag{5}$$

Assumption 2 (Gap between Forward and Backward Diffusion Paths). Given dataset $\mathcal{V}$, let $\mathbf{v}_t$ denote the random variable sampled from intermediate distributions at time t in both the forward process (following $q(\mathbf{v}_t)$) and backward process (following $p_\phi(\mathbf{v}_t)$). We say the Assumption 2 is satisfied if there exists small positive constant $\tilde{\gamma}_t \ll 2$ such that

$$\|q(\mathbf{v}_t) - p_\phi(\mathbf{v}_t)\|_{\text{TV}} \leq \tilde{\gamma}_t, \forall t \in [T]. \tag{6}$$

Assumption 1 states that denoising networks, when trained using the loss function in Eq. (3), can effectively infer clean data from intermediate noisy data distributions. Given a sufficiently expressive model, we expect γ_t to be small. Assumption 2 asserts that diffusion and generation paths are close, which is a reasonable assumption due to the recent analysis (Campbell et al., 2022). However, one cannot use Eq. (6) to derive privacy bound directly as closeness in total variation does not imply DP in general though the reverse could be true (Bassily et al., 2016).

With above assumptions, we investigate the flow of privacy leakage along generation process. Our analysis centers around the inherent privacy guarantees of DDM-generated samples at specific release step, denoted as T_{rl}. Later, we will show that our privacy bound is tight when generating a single sample ($m = 1$).

Theorem 1 (**Inherent pDP Guarantees for DDMs**). *Given a dataset $\mathcal{V}_0$ with size $|\mathcal{V}_0| = s+1$ and a data point $\mathbf{v}^* \in \mathcal{V}_0$ to be protected, denote $\mathcal{V}_1$ such that $\mathcal{V}_1 = \mathcal{V}_0 \backslash \{\mathbf{v}^*\}$. Assume the denoising networks trained on $\mathcal{V}_0$ and $\mathcal{V}_1$ satisfy Assumption 1 and Assumption 2. Given a specific time step T_{rl}, the mechanism $\mathcal{M}_{T_{rl}}(\cdot; m)$ satisfies (ϵ, δ)-pDP with respect to $(\mathcal{V}_0, \mathbf{v}^*)$ such that given ϵ,*

$$\delta(\mathcal{V}_0, \mathbf{v}^*) \leq m \Bigg[\underbrace{\sum_{t=T_{rl}}^{T} \min\left\{ \frac{4N_{(1+c_t^*)\eta_t}(\mathbf{v}^*)}{s}, 1 \right\} \cdot \frac{n}{s^{\psi_t}} + \frac{n(1-\frac{1}{\bar{R}_{t-1}})}{s^2}}_{\textit{Main Privacy Term}} + \underbrace{\mathcal{O}\left(\sqrt{\gamma_t} + \tilde{\gamma}_t\right)}_{\textit{Error Term}} \Bigg] / (\epsilon(1-e^{-\epsilon})). \tag{7}$$

*where ψ_t, η_t, c_t^* are **data-dependent quantities** determined by $\mathbf{v}^*$ and $\mathcal{V}_1$. Define a similarity measure $Sim(\mathbf{v}^*, \mathcal{V}) = \sum_{\mathbf{v} \in \mathcal{V}} \bar{R}_t^{-\bar{\omega}(\mathbf{v}, \mathbf{v}^*)}$. Then, ψ_t, η_t, c_t^* follow*

$$\frac{n}{s^{\psi_t}} = \frac{(\overline{\alpha}_{t-1} - \overline{\alpha}_t)/(k\bar{\mu}_t^+ \bar{\mu}_t^-)}{1 + Sim(\mathbf{v}^*, \mathcal{V}_1)} \cdot \sum_{i=1}^{n} \log\left(1 + \frac{\bar{R}_{t-1}^2 - 1}{\bar{R}_{t-1}^2 Sim(\mathbf{v}^*, \mathcal{V}_1^{i|\mathbf{v}^{*i}}) + Sim(\mathbf{v}^*, \mathcal{V}_1) + 1}\right). \tag{8}$$

And, η_t, c_t^* are the <u>smallest</u> $\eta_t \in \{1, 2, ..., n\}$, $c_t^* \in \{0, \frac{1}{\eta_t}, \frac{2}{\eta_t}, ..., \frac{n-\eta_t}{\eta_t}\}$ which satisfy

$$\eta_t \geq \frac{\log \vartheta(\eta_t)}{\log \frac{1}{n(1-\bar{\mu}_t^+)}} + \left(\frac{\log\left(\vartheta(\eta_t) \frac{\overline{\alpha}_{t-1} - \overline{\alpha}_t}{k\bar{\mu}_t^+ \bar{\mu}_t^-} \cdot s^{\psi_t}\right)}{2 \log \bar{R}_t} - 2 \right)_+, \quad c_t^* \geq \frac{\frac{1}{\eta_t} \log \vartheta((1+c_t^*)\eta_t) + \frac{3}{2}}{\log \frac{1}{\bar{\mu}_t^-} - 1}. \tag{9}$$

where $\vartheta(\eta) = (s - N_\eta(\mathbf{v}^*))/N_\eta(\mathbf{v}^*)$ that represents the ratio between the numbers of points outside the η-ball and inside it.

Theorem 1 quantifies the privacy leakage of a specific point $\mathbf{v}^*$ in training set $\mathcal{V}_0$. The privacy bound comprises a main privacy term that represents the inherent pDP guarantees for DDMs, highlighting the data-dependent nature of our bound, and an error term stemming from denoising network training and path discrepancies. Those data-dependent quantities are complex to maintain a tight measurement for a dataset-data point pair. Next, we will further explain these quantities.

First, as the generation process forms a Markov chain where the transition probability $p_\phi(\mathbf{v}^{(t-1)}|\mathbf{v}^{(t)})$ is learned from training, each generation step will leak some information from the training dataset. It can be shown that the majority of such leakage, represented in the pDP bound (in the appendix) follows

$$\mathbb{E}_{\mathbf{v} \sim p_\phi(\mathbf{v}_{t|0} = \mathbf{v})} d^{(t)}(\mathbf{v}) \tag{10}$$

where let $\mathbf{v}_{t|\lambda}$ represents the r.v. of the generated data at time t of the generation process when the diffusion model gets trained over the dataset $\mathcal{V}_\lambda$, $\lambda \in \{0, 1\}$ and $d^{(t)}(\mathbf{v}) = \sum_{\lambda \in \{0,1\}} \mathcal{D}_{\mathrm{KL}}(p_\phi(\mathbf{v}_{t-1|\lambda}|\mathbf{v}_{t|\lambda} = \mathbf{v}) \| p_\phi(\mathbf{v}_{t-1|\bar{\lambda}}|\mathbf{v}_{t|\bar{\lambda}} = \mathbf{v}))$ which characterizes a symmetric distance between two conditional distributions characterized by the learned diffusion model. Essentially, the three data-dependent quantities ψ_t, η_t, c_t^* are to bound Eq. (10).

Quantity ψ_t: As shown in Fig. 2, $\frac{n}{s^{\psi_t}}$ quantifies $\max_{\mathbf{v}} d^{(t)}(\mathbf{v})$ where the maximum is achieved at the removed point $\mathbf{v} = \mathbf{v}^*$ (green in Fig. 2). A closer inspection reveals that ψ_t depends on the terms $\mathrm{Sim}(\mathbf{v}^*, \mathcal{V}_1)$ and $\mathrm{Sim}(\mathbf{v}^*, \mathcal{V}_1^{i|\mathbf{v}^{*i}})$. By the definition of $\bar{\omega}$, these terms assess how $\mathbf{v}^*$ aligns with the remaining points in $\mathcal{V}_1$.

Evolution of ψ_t. During the generation phase, as t progresses from T to 1, the values of $\frac{1}{s^{\psi_t}}$ increase from $\mathcal{O}_s(\frac{1}{s^2})$ to $\mathcal{O}_s(1)$. This implies that the potential privacy risk escalates as the data generation process evolves from a noisy regime to a noise-free regime.

Quantities η_t and c_t^*: It is evident that the intermediate generated measure $p_\phi(\mathbf{v}_{t|0})$ (blue in Fig. 2) diverges from the delta measure on the most sensitive point $\delta_{\mathbf{v}=\mathbf{v}^*}$ (green). Therefore, the actual privacy leakage characterized by $d^{(t)}(\mathbf{v})$ (yellow) averaged over the measure $p_\phi(\mathbf{v}_{t|0})$ is much less than its maximum. To provide a tight characterization of such, the two quantities η_t and c_t^* are introduced to define a local region $\mathcal{S} = \{\mathbf{v}' \in \mathcal{X}^n : \bar{\omega}(\mathbf{v}, \mathbf{v}') \leq (1 + c_t^*)\eta_t\}$ centered on vulnerable point $\mathbf{v}^*$, within which the privacy

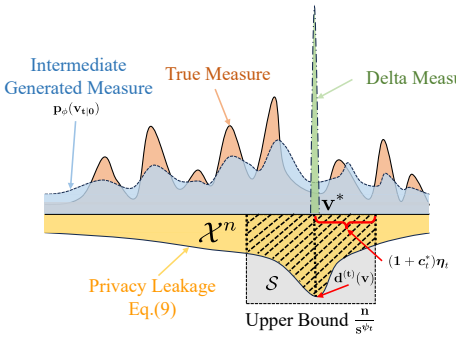

Figure 2: Illustration of Data-dependent Quantities.

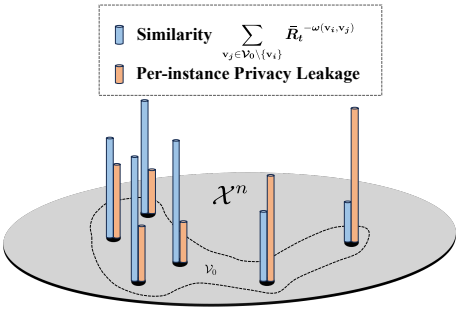

Figure 3: Illustration of the correlation between dataset similarity $(\mathrm{Sim}(\mathbf{v}_i, \mathcal{V}_0 \backslash \{\mathbf{v}_i\}), \forall \mathbf{v}_i \in \mathcal{V}_0)$ and pDP Leakage.

leakage can be bounded by the sum of (a) $p_\phi(\mathbf{v}_{t|0} \in \mathcal{S}) \max_{\mathbf{v} \in \mathcal{S}} d^{(t)}(\mathbf{v})$ with a small $p_\phi(\mathbf{v}_{t|0} \in \mathcal{S})$ and (b) $p_\phi(\mathbf{v}_{t|0} \notin \mathcal{S}) \max_{\mathbf{v} \notin \mathcal{S}} d^{(t)}(\mathbf{v})$ with a small $\max_{\mathbf{v} \notin \mathcal{S}} d^{(t)}(\mathbf{v})$. (η_t, c_t^*) shown in Eq. (9) are chosen to properly balance these two parts. η_t and c_t^* always exist: Note that when $\eta_t = n$ or $c_t^* = n/\eta_t - 1$, the right-hand side of either inequality in Equation (9) approaches $-\infty$ $(\log 0)$. In fact, both of the RHS's of the two inequalities decrease w.r.t. η_t and c_t^*. So, in practice, the smallest η_t and c_t^* can be found via binary search given the dataset $\mathcal{V}_0$ and $\mathbf{v}^*$.

Evolution of $(1 + c_t^*)\eta_t$. For each time step t, the smallest value of $(1 + c_t^*)\eta_t$ is chosen as the radius. As t progresses from T to 1, the value of $(1 + c_t^*)\eta_t$ monotonically decreases. When $\bar{\alpha}_t$ approaches 1 for smaller t values, $(1 + c_t^*)\eta_t$ tends to zero, i.e., $\mathcal{S}$ only includes $\mathbf{v}^*$. The reason is that, as smaller t values, different data points are less mixed with others (because of less noise added in the forward process), the privacy leakage of $\mathbf{v}^*$ becomes more concentrated around the changes of the likelihoods of the generated data points that look like $\mathbf{v}^*$, thus calling for a decrease of the radius. To consider the impact on the bound in Eq. (7), the number of data points in this region $N_{(1+c_t^*)\eta_t}(\mathbf{v}^*)$ will decrease from s to 1 as t changes from T to 1.

Discussion on Theorem 1. Based on the previous discussion, as t decreases from T to 1, $N_{(1+c^*)\eta_t}(\mathbf{v}^*)/s$ changes from $\mathcal{O}_s(1)$ to $\mathcal{O}_s(1/s)$, and $1/s^{\psi_t}$ changes from $\mathcal{O}_s(\frac{1}{s^2})$ to $\mathcal{O}_s(1)$. Consequently, the privacy leakage for each-step DDM-generated samples gradually weakens from $\mathcal{O}_s(\frac{m}{s^2})/[\epsilon(1 - e^{-\epsilon})]$ to $\mathcal{O}_s(\frac{m}{s})/[\epsilon(1 - e^{-\epsilon})]$.

This implies a natural utility-privacy tradeoff for the data generated by DDMs. In practice, to guarantee the data quality, we often release the data in the noise-free side $(t = 0)$, where only a weak privacy guarantee of approximately $(\epsilon, \mathcal{O}_s(m/s)/[\epsilon(1 - e^{-\epsilon})])$ can be achieved. To enhance data privacy, we may expect to release the data generated with a larger step $t \geq 1$.

This result also reveals that the inherent privacy guarantees of releasing data generated by DDMs is weak $(\propto \mathcal{O}(m/s))$, in the same order of guarantees for GAN-generated samples (Lin et al., 2021). This characterization also matches many recent empirical studies that have shown concerns on privacy leakage due to publishing data generated by DMs (Hu & Pang, 2023; Carlini et al., 2023; Dockhorn et al., 2022). While privacy budgets for all data points maintain the same order in relation to the sample size, the contacts can differ markedly across data points. Intuitively, a data point $\mathbf{v}^* \in \mathcal{V}_0$ with less similarity with the other data points tends to have higher privacy leakage. This is indicated by Eq. (8), where a smaller similarity $\sum_{\mathbf{v} \in \mathcal{V}_0 \setminus \{\mathbf{v}^*\}} \bar{R}_t^{-\omega(\mathbf{v}^*, \mathbf{v})}$, leads to a larger pDP leakage (as illustrated in Fig. 3).

The upper bound for $t = 1$ with a dependence on the dataset size $\mathcal{O}_s(\frac{1}{s})$ demonstrates a weak privacy guarantee given by DDMs, as this is on par with the privacy implication of the Strawman approach that uniformly at random samples from the original dataset and publishes the samples. However, we emphasize that this is not due to a loose analysis, as the following will provide a lower bound of such privacy leakage due to DDM generation in the same order. Beyond this, our upper bound is valuable as it elucidates that releasing a dataset at an earlier stage (with a larger t step) generated by DDMs could potentially strengthen the privacy guarantee to $\Omega_s(\frac{1}{s^2})$. Moreover, our upper bound specifies the influence of diffusion coefficients

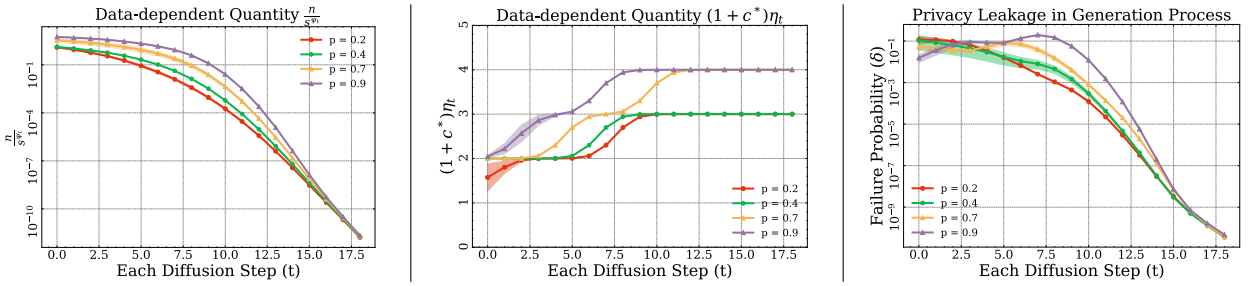

Figure 4: pDP Leakage in Eq. (7): **LEFT:** Characterization of $\frac{n}{s^{\psi_t}}$. **MIDDLE:** Characterization of $(1+c_t^*)\eta_t$. **RIGHT:** Characterization of Privacy Leakage (Main Privacy Term). **Experimental Setup:** Given specific DDM design $k=5, n=5, T=20, \epsilon=10$ trained on dataset with $s=1000$ following the distribution in Sec. 3.3 with parameter p. Fix $\mathbf{v}^*$ where each column has a non-majority category. Results are based on 5 times independent tests.

and dataset distributions on the privacy bounds, which the Strawman approach by publishing the samples from the original dataset cannot tell. More details will be discussed in Sec. 3.2.

Tightness of Privacy Bound w.r.t Sample Size. In Theorem 1, the privacy parameter of δ scales as $\mathcal{O}_s(\frac{1}{s})$ with sample size. Here we establish a lower bound for δ with respect to the sample size by evaluating the worst-case scenario and show that $\mathcal{O}(1/s)$ is the optimal bound that DDM can achieve inherently for $m=1$. For illustrative purposes, consider the case where $n=2$ with two distinct categories. Define adjacent datasets: $\mathcal{V}_0 = \{\underbrace{[0,0]^T,...,[0,0]^T}_{s-1}, [1,1]^T, [1,1]^T\}$ and $\mathcal{V}_1 = \mathcal{V}_0 \backslash \{[1,1]^T\}$.

Theorem 2 (Lower Bound on Inherent pDP Guarantees for DDMs). *Assume the denoising networks are perfectly trained. Given a diffusion model architecture design (Sigmoid Schedule* $\alpha_t = \frac{Sigmoid(3)-Sigmoid(\frac{3t}{T})}{Sigmoid(3)-0.5}, T=10$*), there exist an adjacent dataset* $\mathcal{V}_0, \mathcal{V}_1 = \mathcal{V}_0 \backslash \{\mathbf{v}^*\}$ *with feature dimension* $n=2$ *such that the mechanism* $\mathcal{M}_0(\cdot;1)$ *does not satisfies* $(0.04,\delta)$*-pDP with respect to* $(\mathcal{V}_0, \mathbf{v}^*)$ *for any* $\delta < \frac{1}{6s}$.

Regarding lower bounds on (ϵ,δ)-pDP under general model configurations, including diffusion schedules and diffusion steps, please refer to Appendix D.

3.2 Impact of DDM Coefficients and Dataset Distributions on the Privacy Bound

Influence of Diffusion Coefficients. The privacy term is largely influenced by the proximity between $\mathbf{v}^*$ and $\mathcal{V}_1$. As time t progresses, this similarity is governed by the transition ratio $\bar{R}_t$. A faster rate of diffusion coefficients going to zero boosts this ratio, enhancing the privacy guarantee. Experiments in Sec. 4 validate this observation.

Impact of Dataset Distribution. We find that ψ_t has a major effect on the privacy bound. ψ_t is influenced by the similarity between the additional point $\mathbf{v}^*$ and $\mathcal{V}_0 \backslash \{\mathbf{v}^*\}$. If $\mathbf{v}^*$ is far away from (close to) the rest points in $\mathcal{V}_0$, then $\text{Sim}(\mathcal{V}_0 \backslash \{\mathbf{v}^*\}, \mathbf{v}^*, t)$ becomes small (large) and the corresponding term $s^{-\psi_t}$ become large (small), which indicates weaker (stronger) protection of $\mathbf{v}^*$. *This indicates that points with notably low* $Sim(\mathcal{V}_0 \backslash \{\mathbf{v}^*\}, \mathbf{v}^*, t)$ *are probably sensitive points in the dataset.*

3.3 Characterizing Data-dependent Quantities under Simple Distributions

Here, we consider the training dataset sampled from some specific distributions to further illustrate the data-dependent quantities.

Consider a distribution such that each column independently takes value $l \in [k]$ with probability p ($p \geq \frac{1}{k}$) and any other $k-1$ categories with probability $\frac{1-p}{k-1}$. Let $\mathbf{v}^*$ take non-majority category ($(\mathbf{v}^*)^i \neq l$) along all n columns (termed **non-majority points**, which thus tends to have higher privacy leakage) and the rest points in $\mathcal{V}_0 \backslash \{\mathbf{v}^*\}$ are sampled from the distribution. We have the following characterization (For detailed explanations and proofs, please refer to Appendix F).

- $\frac{1}{s^{\psi_t}}$**.** For a sufficiently large s (detailed in appendix), with high probability, $\frac{1}{s^{\psi_t-2}} \to \frac{(\overline{\alpha}_{t-1}-\overline{\alpha}_t)/(k\bar{\mu}_t^+\bar{\mu}_t^-)}{\bar{R}_{t-1}^2\cdot\tau_t^{2n-1}\cdot\frac{1-p}{k-1}+\tau_t^{2n}}$, where $\tau_t := \frac{1-p}{k-1} + \frac{\bar{\mu}_t^-}{\bar{\mu}_t^+}(1-\frac{1-p}{k-1})$. In the noisy regime (a large t, $\frac{\bar{\mu}_t^-}{\bar{\mu}_t^+} \to 1$), $\tau_t \to 1$, $\frac{1}{s^{\psi_t}} = \mathcal{O}_s(\frac{1}{s^2})$. For distribution characterized by larger skewness, i.e., larger p, we have smaller τ_t result in larger $\frac{1}{s^{\psi_t}}$. Fig. 4 (LEFT) precisely matches the above conclusions.
- $\boldsymbol{\eta_t, c_t^*}$**.** For a sufficiently large s (detailed in appendix), a sufficient condition for η_t and c_t^* to satisfy Eq. (9) is

$$\eta_t \geq n - \left(\frac{n - \log(s\sqrt{\frac{\overline{\alpha}_{t-1}-\overline{\alpha}_t}{k\bar{\mu}_t^+\bar{\mu}_t^-}}) / \log\frac{\bar{\mu}_t^+}{\bar{\mu}_t^-}}{2\log\frac{k-1}{1-p} / \log(\max\{\frac{1}{n\bar{\mu}_t^-}, 1\}) + 1} \right)_+, \quad c_t^* \geq \frac{\frac{n-\eta_t}{\eta_t}\log\frac{k-1}{1-p} - \log\frac{1}{2e}}{\log\frac{k-1}{1-p} + \log\frac{1}{e\bar{\mu}_t^-}}. \tag{11}$$

  In the noise free regime ($\alpha_t \to 1$), $\eta_t \to 0$, while in the noise full regime ($\alpha_t \to 0$), $\eta_t \to n$. From noise free regime to noisy regime, $\bar{\mu}_t$ increases, $c_t^* \to \frac{n-\eta_t}{\eta_t}$. Furthermore, as we rise in the skewness (p) of the distribution, the R.H.S of Eq. 11 monotonically increases, and results in larger values for η_t and c_t^*. Fig. 4 (MIDDLE) matches the above conclusions.

3.4 The Algorithm for Evaluating Privacy Bound in Eq. (7) on a given Dataset

In practical situations, when data curators release synthetic data, it is crucial to assess the privacy safeguards of the mechanism trained on a specific dataset. This ensures the synthetic data upholds privacy and the confidentiality of the training data's sensitive information. To this end, we introduce Algorithm 2 in Appendix C, to compute the privacy bound, enabling direct per-instance privacy leakage calculation for DDM-generated datasets given particular training sets. Specifically, for each $\mathbf{v}^*$, we determine ψ_t, η_t, and c_t^* to compute $\delta(\mathbf{v}^*, \mathcal{V}_0)$ using Eq. (7). Using this algorithm, data curator can have better assessment of the potential privacy leakage of each point in training set and may exclude sensitive points $\mathbf{v}^*$ (outliers) with high $\delta(\mathbf{v}^*, \mathcal{V}_0)$ to enhance privacy protection. This approach's efficacy is confirmed with real dataset experiments in Sec. 4.

4 Experiments

We validate our theoretical findings via computational simulations on synthetic and real-world datasets.

4.1 Synthetic Experiments

We first we study the asymptotic behavior of privacy leakage with respect to the training dataset size s . Given a DDM with 100 diffusion steps and trained with a linear schedule $\alpha_t = 1 - \frac{t}{T}$, We fix $\mathbf{v}^*$ and increase the number of samples in the training set from $1e4$ to $1e7$, ensuring that the newly added samples satisfy $\bar{\omega}(\mathbf{v}, \mathbf{v}^*) = n$, which makes $\mathbf{v}^*$ with high privacy leakage risk. From the results shown in Fig. 5 (LEFT, MIDDLE) confirm our theoretical prediction that, in noise-free regime ($t = 1$, Fig. 5 (LEFT)), the **main privacy term** in Theorem 1 is $\mathcal{O}_s(\frac{1}{s})$, which is almost a linear decay with a slope of -1 in the logarithmic scale (all lines in the figure). On the other hand, in the noisy-regime ($t = 50$, Fig. 5 (MIDDLE)), the privacy leakage term decays faster at the rate of $\mathcal{O}_s(\frac{1}{s^2})$, which is evident from the linear decay with a slope around -2. In the second experiment, we examine how decay rate of diffusion coefficients affects the privacy bound. Given specific $\mathbf{v}^*$ (non-majority categories along all entries). We sample the training set from the distribution with $p = 0.5$ in Sec. 3.3. We consider two noise schedules: linear schedule and sigmoid schedule. In Fig. 5 RIGHT, the red line denotes the linear schedule with decay rate $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and the blue line denotes the Sigmoid schedule where decay rate increases from 2.5 to 5. δ decreases along both two lines as we increase the decay rate of diffusion coefficients. This indicates that a faster decay rate in diffusion coefficients implies better privacy.

More results discussing privacy leakage and the behaviors of data-dependent quantities under various DDM configurations are given in Appendix I.

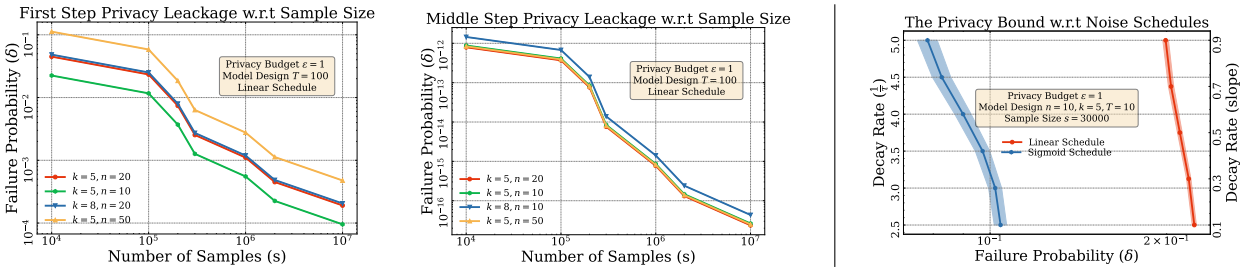

Figure 5: **LEFT:** Privacy Leakage at $t = 1$ (Noise-free Regime). **MIDDLE:** Privacy Leakage at $t = 50$ (Noisy Regime). **Right:** Privacy Leakage w.r.t Decay Rate under Linear ($\alpha_t = 1 - \text{decay rate} * \frac{t}{T}$) and Sigmoid ($\alpha_t = \frac{\text{Sigmoid}(3*\text{decay rate}) - \text{Sigmoid}(\frac{3t}{T} * \text{decay rate})}{\text{Sigmoid}(3*\text{decay rate}) - 0.5}$) Schedules. Results are based on 5 times independent tests.

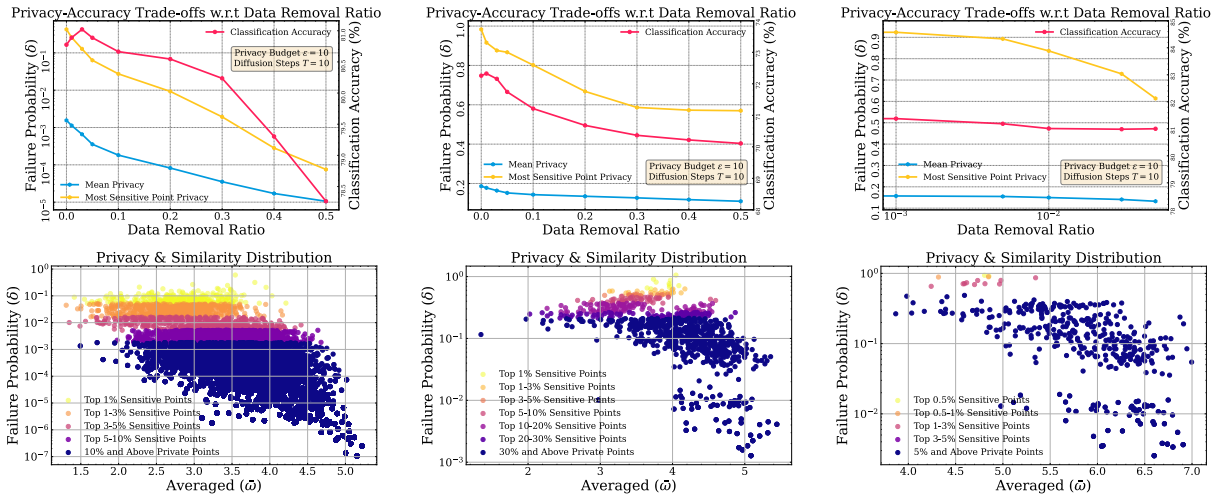

Figure 6: **First Row:** Privacy-utility trade-offs with respect to data removal ratio. **LEFT:** Adult, **MIDDLE:** German Credit, **RIGHT:** Loan. **Experimental Setup:** DDM design: T = 10, Linear Schedule. **Second Row:** Visualizing privacy budget in relation to average feature overlap. **LEFT:** Adult, **MIDDLE:** German Credit, **RIGHT:** Loan.

4.2 Experiments on Real Datasets

4.2.1 Effectiveness of Privacy Bound Algorithm for Data Sensitivity Assessment

In this series of experiments, we aim to showcase our privacy algorithm's effectiveness in assessing the sensitivity of individual data points within real-world datasets and to delineate the relationship between sample privacy leakage and dataset similarity (as illustrated in Fig. 3). Additionally, from a data curator's perspective, we explore the potential of our algorithm assisting in outlier removal, which may enhance privacy protections while maintaining utility performance. It is important to note that the use of pDP assessment itself is kept confidential to the data curator, safeguarding against any privacy concerns. We evaluate our algorithm on three benchmark datasets: Adult (Kohavi et al., 1996), German Credit (Hofmann, 1994), and Loan (ItsSuru) with (# training samples, # feature dimensions, # categories) of (30718, 9, 5), (1000, 10, 5), and (480, 11, 4) (see Appendix H for details).

Experimental Settings. Our study, approaching from the perspective of a data curator who preprocesses the dataset, investigates how varying the sensitive data removal ratio according to our per-instance privacy bound assessment can potentially assist data curators in eliminating outliers to enhance privacy protection. Specifically, we calculate the privacy budget for every point in the dataset according to Eq. (7) via Algorithm 2 and remove the most sensitive points according to the assessment in the dataset amounting to a specific portion which is controlled by the removal ratio. The removal ratio ranges from 0.01 to 0.5 for the Adult and

German Credit datasets, and between 0.001 and 0.05 for the Loan dataset. It is important to underscore that we recalculate and report the mean privacy leakage (blue line) and the most sensitive point privacy budget (yellow line) after each data removal process in Fig. 6 (First Row). We measure utility performance with respect to downstream classification task by training a binary classifier on DDM-generated samples and evaluate its performance (red line) on the original dataset. We further illustrate the sensitive points—those removed from the dataset—by graphing their potential privacy leakage alongside the average overlap with the entire dataset across all feature dimensions, denoted by $\bar{\omega}$. The visualizations are presented in Figure 6 (Second Row).

Remove privacy sensitive points with comparable utility. As depicted in Fig. 6 (First Row), eliminating a minor proportion of the most sensitive points from the dataset results in a decrease in privacy leakage. Meanwhile, the classification accuracy (red line) only gets slightly decreased: 81% $\rightarrow$ 78% for Adult, 73% $\rightarrow$ 70% for German Credit, 81.1% $\rightarrow$ 79.8% for Loan (note that for Loan we remove at most 5% data points as its size is too small). More interestingly, by removing a certain number of those most sensitive data points, the classification model trained over the pruned generated dataset may even achieve better performance over the original dataset, say removing 3% in Adult and 1% in German Credit. We attribute such gains to the fact that the most sensitive data points are often outliers in the dataset, which may be actually not good for training an ML model. In data visualization (Fig. 6, Second Row), we note that the data points prone to greater privacy leakage tend to have less feature overlap, indicating that these data points have a lower similarity to others in the dataset.

4.2.2 Evaluation of DDM Vulnerability to Black-box Membership Inference Attacks

In this subsection, we further investigate the privacy leakage of DDMs from membership inference attacks perspective.

Black-box Attacks with No Auxiliary Knowledge: In alignment with the experimental settings delineated by (Hayes et al., 2017), this study considers *black-box attacks* where the attacker has no prior knowledge or external information regarding the target model, i.e. DDM in our case, including *model parameters / hyper-parameters, model architecture, training data*, or *prediction scores*. For evaluation, it is hypothesized that the attacker possesses access to the whole dataset, denoted as $X = X_{\text{train}} \cup X_{\text{non-train}}$. Additionally, it is presumed that the adversary is aware of the size of the training set. The attack primarily utilizes the target model's generated samples to identify and exploit its vulnerabilities.

Experimental Settings: Our experiments utilize the Adult dataset. We partition this dataset by randomly selecting 20% of the records as the training set, denoted as X_{train}, while the remainder is labeled as $X_{\text{non-train}}$. We train a DDM to learn the conditional distribution of X_{train} given their corresponding labels Y_{train}. Then, the adversary employs a discriminative model $f(\cdot;\theta)$: feature space $\mathcal{X} \mapsto$ label space $\mathcal{Y}$ (detailed in Appendix. H), which is trained using two-class samples $(X_{\text{gen}}, Y_{\text{gen}})$ generated by the DDM and denote trained model weights as θ_{gen}. This trained discriminative model $f(\cdot;\theta_{\text{gen}})$ is then used to make predictions across the entire dataset $X_{\text{train}} \cup X_{\text{non-train}}$. We identify $|X_{\text{train}}|$ samples with the highest confidence values—those whose prediction scores are closest to the true labels—and designate these as training members $\tilde{X}_{\text{train}}$. The evaluation of the DDMs is centered around varying the decay rate of diffusion coefficients, influencing the degree of privacy leakage in the models. The outcomes of these evaluations, particularly the attack accuracy ($\frac{|\tilde{X}_{\text{train}} \cap X_{\text{train}}|}{|X_{\text{train}}|}$), are illustrated in Fig. 7. More specifically, we examine DDMs with total diffusion steps T set to either 20 or 30, and a linear schedule defined as $\alpha_t = 1 - \text{decay rate} \times \frac{t}{T}$. Additionally, the decay rate is varied within the range {0.1, 0.3, 0.5, 0.8, 1.0} to adjust the privacy guarantees of the DDMs, with a faster decay rate typically providing better privacy protection.

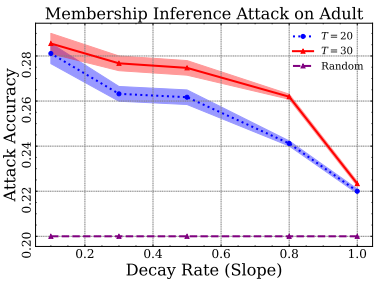

Figure 7: Black-box Attack with no auxiliary knowledge over DDMs under various designs. Results are averaged over 3 independent tests.

Increased Privacy Leakage Enhances Vulnerability of Models to Attacks: As illustrated in Fig. 7, the blue and red lines represent DDMs with total diffusion steps of 30 and 20, respectively. A noteworthy

observation is that the DDM with $T = 30$ exhibits higher attack accuracy, suggesting a stronger capability for memorizing training data in models with a greater number of diffusion steps. Furthermore, an increase in the decay rate leads to improved privacy guarantees for the DDMs. This enhancement in privacy is evidenced by a decrease in attack accuracy for both models (as represented by the blue and red lines), which diminishes from 28% / 29% to approximately 22%. It is important to note that the performance of both lines surpasses that of random guessing (indicated by the purple line), signifying that our target models retain training data memorization across all evaluated settings.

5 Conclusion

In this work, we analyzed data-dependent privacy bound for the synthetic datasets generated by DDMs, which revealed a weak privacy guarantee of DDMs. Thus, to meet practical needs, other privacy-preserving techniques such as DP-SGD (Abadi et al., 2016) and PATE (Papernot et al., 2016) may have to be corporated. Our findings well align with empirical observations over synthetic and real datasets.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pp. 510–526. Springer, 2019.

Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology*, 2005112001:400, 2005.

Mehreen Ahmed, Hammad Afzal, Awais Majeed, and Behram Khan. A survey of evolution in predictive models and impacting factors in customer churn. *Advances in Data Science and Adaptive Analysis*, 9(03): 1750007, 2017.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Matej Balog, Ilya Tolstikhin, and Bernhard Schölkopf. Differentially private database release via kernel mean embeddings. In *International Conference on Machine Learning*, pp. 414–422. PMLR, 2018.

Samah Baraheem and Zhongmei Yao. A survey on differential privacy with machine learning and future outlook. *arXiv preprint arXiv:2211.10708*, 2022.

Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 273–282, 2007.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1046–1059, 2016.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):1–25, 2013.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

Andrew Carr. Diffusion models for document synthesis. https://gretel.ai/blog/diffusion-models-for-document-synthesis, 2022.

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*, pp. 286–305. PMLR, 2017.

Jillian M Clements, Di Xu, Nooshin Yousefi, and Dmitry Efimov. Sequential deep learning for credit risk monitoring with tabular financial data. *arXiv preprint arXiv:2012.15330*, 2020.

Graham Cormode, Magda Procopiuc, Divesh Srivastava, and Thanh TL Tran. Differentially private publication of sparse data. *arXiv preprint arXiv:1103.0825*, 2011.

Graham Cormode, Cecilia Procopiuc, Divesh Srivastava, Entong Shen, and Ting Yu. Differentially private spatial decompositions. In *2012 IEEE 28th International Conference on Data Engineering*, pp. 20–31. IEEE, 2012.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Bolin Ding, Marianne Winslett, Jiawei Han, and Zhenhui Li. Differentially private data cubes: optimizing noise sources and consistency. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 217–228, 2011.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022.

August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, 4(1):141, 2021.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380, 2009.

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*, pp. 265–284. Springer, 2006.

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.

Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.

Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

Kilian Konstantin Haefeli, Karolis Martinkus, Nathanaël Perraudin, and Roger Wattenhofer. Diffusion models for graphs benefit from discrete state spaces. *arXiv preprint arXiv:2210.01549*, 2022.

Frederik Harder, Kamil Adamczewski, and Mijung Park. Dp-merf: Differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation. In *International conference on artificial intelligence and statistics*, pp. 1819–1827. PMLR, 2021.

Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*, 25, 2012.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

Hailong Hu and Jun Pang. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023.

ItsSuru. Loan Data. Kaggle. DOI: https://www.kaggle.com/datasets/itssuru/loan-data.

Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021.

Zhanglong Ji, Zachary C Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*, 2014.

Zach Jorgensen, Ting Yu, and Graham Cormode. Publishing attributed social graphs with formal privacy guarantees. In *Proceedings of the 2016 international conference on management of data*, pp. 107–122, 2016.

Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the johnson-lindenstrauss transform. *arXiv preprint arXiv:1204.2606*, 2012.

Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, 1996.

Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60, 2005.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pp. 106–115. IEEE, 2006.

Yang Li, Michael Purcell, Thierry Rakotoarivelo, David Smith, Thilina Ranbaduge, and Kee Siong Ng. Private graph data release: A survey. *arXiv preprint arXiv:2107.04245*, 2021.

Yang Li, Michael Purcell, Thierry Rakotoarivelo, David Smith, Thilina Ranbaduge, and Kee Siong Ng. Private graph data release: A survey. *ACM Computing Surveys*, 55(11):1–39, 2023.

Yuting Liang and Reza Samavi. Optimization-based k-anonymity algorithms. *Computers & Security*, 93: 101753, 2020.

Zinan Lin, Vyas Sekar, and Giulia Fanti. On the privacy properties of gan-generated samples. In *International Conference on Artificial Intelligence and Statistics*, pp. 1522–1530. PMLR, 2021.

Kun Liu and Evimaria Terzi. Towards identity anonymization on graphs. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 93–106, 2008.

Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1): 92–106, 2005.

Xuesong Lu, Yi Song, and Stéphane Bressan. Fast identity anonymization on graphs. In *Database and Expert Systems Applications: 23rd International Conference, DEXA 2012, Vienna, Austria, September 3-6, 2012. Proceedings, Part I 23*, pp. 281–295. Springer, 2012.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *2008 IEEE 24th international conference on data engineering*, pp. 277–286. IEEE, 2008.

Amy L McGuire and Richard A Gibbs. No longer de-identified. *Science*, 312(5772):370–371, 2006.

Sebastian Meiser. Approximate and probabilistic differential privacy definitions. *Cryptology ePrint Archive*, 2018.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Shreyas Patel, Ashutosh Kakadiya, Maitrey Mehta, Raj Derasari, Rahul Patel, and Ratnik Gandhi. Correlated discrete data generation using adversarial training. *arXiv preprint arXiv:1804.00925*, 2018.

Yury Polyanskiy. Lecture on information theoretic methods in statistics and computer science. University Lecture, 2020.

Tahereh Pourhabibi, Kok-Leong Ong, Booi H Kam, and Yee Ling Boo. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133:113303, 2020.

Alessandra Sala, Xiaohan Zhao, Christo Wilson, Haitao Zheng, and Ben Y Zhao. Sharing graphs using differentially private graph models. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 81–98, 2011.

Nicholas J Schork. Personalized medicine: time for one-person trials. *Nature*, 520(7549):609–611, 2015.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pp. 13–31. Springer, 2007.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

Traian Marius Truta and Bindu Vinay. Privacy protection: p-sensitive k-anonymity property. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp. 94–94. IEEE, 2006.

Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):1–13, 2020.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In *Machine Learning for Health*, pp. 341–354. PMLR, 2020.

Dinusha Vatsalan, Peter Christen, and Vassilios S Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

Jianian Wang, Sheng Zhang, Yanghua Xiao, and Rui Song. A review on graph neural network methods in financial applications. *arXiv preprint arXiv:2111.15367*, 2021.

Yu-Xiang Wang. Per-instance differential privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019.

Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*, 2023.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Compressed and privacy-sensitive sparse regression. *IEEE Transactions on Information Theory*, 55(2):846–866, 2009a.

Shuheng Zhou, Katrina Ligett, and Larry Wasserman. Differential privacy with compression. In *2009 IEEE International Symposium on Information Theory*, pp. 2718–2722. IEEE, 2009b.