

Synthetic Data – Anonymisation Groundhog Day

Theresa Stadler
EPFL

Bristena Oprisanu
UCL

Carmela Troncoso
EPFL

Abstract

Synthetic data has been advertised as a silver-bullet solution to privacy-preserving data publishing that addresses the shortcomings of traditional anonymisation techniques. The promise is that synthetic data drawn from generative models preserves the statistical properties of the original dataset but, at the same time, provides perfect protection against privacy attacks. In this work, we present the first quantitative evaluation of the privacy gain of synthetic data publishing and compare it to that of previous anonymisation techniques.

Our evaluation of a wide range of state-of-the-art generative models demonstrates that synthetic data either does *not prevent inference attacks* or does not retain data utility. In other words, we empirically show that synthetic data does not provide a better tradeoff between privacy and utility than traditional anonymisation techniques. Furthermore, in contrast to traditional anonymisation, the privacy-utility tradeoff of synthetic data publishing is hard to predict. Because it is impossible to predict what signals a synthetic dataset will preserve and what information will be lost, synthetic data leads to a highly variable privacy gain and unpredictable utility loss. In summary, we find that synthetic data is far from the holy grail of privacy-preserving data publishing.

1 Introduction

The rise of data-driven decision making as the prevailing approach to advance science, industrial production, and governance generates a need to share and publish data [20, 21, 65]. At the same time, growing concerns about the implications that data sharing has for individuals and communities call for data publishing approaches that preserve fundamental rights to privacy. Yet, how to share high-dimensional data in a privacy-preserving manner remains an unsolved problem. Attempts to anonymise micro-level datasets have failed across the board [11, 13, 14, 42, 44, 47, 58, 59]. A large number of publications, case studies, and real-world examples demonstrate that high-dimensional, sparse datasets are inherently vulnerable to privacy attacks. The repeated failures to protect

the privacy of microdata releases reflect a fundamental trade-off: information-rich datasets that are valuable for statistical analysis also always contain enough information to conduct privacy attacks [45].

In this landscape, practitioners and researchers see in synthetic data a promising approach to open data sharing that addresses the privacy issues of previous anonymisation attempts [2, 5, 10, 15, 16, 46, 61, 62, 64, 66–68, 72, 74]. Synthetic data is presented as “the next, best step in sanitized data release” [5] that addresses a wide variety of privacy-sensitive use cases from deriving aggregate insights [55, 73] to outlier analysis [40, 63]. Synthetic datasets are promised to preserve the statistical properties of the original data but “contain no personal data” [61] and hence “enable the protection of personally identifiable information” [15]. In this work, we present a rigorous, quantitative assessment of such claims and challenge the common perception of synthetic data as the holy grail of privacy-preserving data publishing.

Previous works. Previous studies on the privacy properties of synthetic data publishing overestimate its benefits over traditional anonymisation for multiple reasons. A common argument to support claims about the privacy benefits of synthetic data is that it is ‘artificial data’ and therefore no direct link between real and synthetic records exists. Hence, many argue, synthetic data by design protects against traditional attacks on microdata releases such as linkage [19, 58] or attribute disclosure [19, 39]. Consequently, many studies rely on similarity tests between real and synthetic records to measure the privacy leakage of synthetic datasets [10, 67, 68]. As we show in this paper, these studies severely underestimate the privacy risks of synthetic data publishing. We introduce two new privacy attacks that demonstrate that, despite its artificial nature, synthetic data does not protect all records in the original data from linkage and attribute inference.

More recent works analyse the vulnerability of generative models against model-specific extraction attacks [8, 26, 28]. Due to their focus on white-box attacks against non-parametric models for synthetic image generation these works do not provide the right framework to assess the privacy risks

of synthetic data sharing in the tabular data domain. In contrast, our attacks treat the data synthesis method as a black-box, focus on tabular data publishing, and allow us to directly compare the privacy leakage of synthetic data to that of traditional anonymisation techniques.

Other approaches rely on formal privacy guarantees for the generative model training process to prevent privacy attacks [1, 6]. While formal definitions of privacy are a clear improvement over the heuristic privacy models of traditional anonymisation, their guarantees are often hard to interpret and difficult to compare to alternative anonymisation techniques. Here, we propose a framework that enables data holders to empirically evaluate the privacy guarantees of differentially private synthetic data publishing and to directly compare its tradeoffs to that of traditional anonymisation techniques.

Contributions. In this paper, we *quantitatively assess* whether (differentially private) synthetic data produced by a wide range of common generative model types does provide a higher gain in privacy than traditional sanitisation at a lower cost in utility. Our results demonstrate that:

(I) Synthetic data drawn from generative models without explicit privacy protection *does not protect outlier* records from linkage attacks. Given access to a synthetic dataset, a strategic adversary can infer, with high confidence, the presence of a target record in the original data.

(II) Differentially private synthetic data that hides the signal of individual records in the raw data protects these targets from inference attacks but does so *at a significant cost in utility*. Worse, in contrast to traditional anonymisation techniques, synthetic datasets do not give any transparency about this tradeoff. It is impossible to predict what data characteristics will be preserved and what patterns will be suppressed.

(III) Our empirical evaluation of the existing implementations of two popular differentially private generative model training algorithms reveals that certain implementation decisions *violate their formal privacy guarantees* and leave some records vulnerable to inference attacks. We provide a novel implementation of both algorithms that addresses these shortcomings.

(IV) We make our evaluation framework available as an *open-source library*. Our implementation allows practitioners and researchers to quantify the privacy gain of publishing a synthetic in place of a raw or sanitised dataset and compare the quality of different anonymisation mechanisms.

2 Synthetic data and generative models

In this section, we formalise the process of synthetic data generation. Let \mathcal{R} be a population of data records where each record $\mathbf{r} \in \mathcal{R}$ contains k attributes: $\mathbf{r} = (r_1, \dots, r_k)$. We denote the unknown joint probability distribution over the data domain of the population as $\mathcal{D}_{\mathcal{R}}$. We refer to $R \sim \mathcal{D}_{\mathcal{R}}^n$, a collection of n data records sampled independently from $\mathcal{D}_{\mathcal{R}}$, as raw dataset which defines the data distribution \mathcal{D}_R .

Synthetic data generation. The goal of a generative model is to learn a representation of the joint probability distribution of data records \mathcal{D}_R . The model training algorithm $\text{GM}(R)$ takes as input a raw dataset R , learns $\mathcal{D}_{g(R)}$, a representation of the joint multivariate distribution \mathcal{D}_R , and outputs a trained generative model $g(R)$. The model $g(R)$ is a stochastic function that, without any input, generates synthetic records \mathbf{s}_i , distributed according to $\mathcal{D}_{g(R)}$. We denote the process of sampling a synthetic dataset $S = (\mathbf{s}_1, \dots, \mathbf{s}_m)$ of size m as $S \sim \mathcal{D}_{g(R)}^m$. We write $g(R) \sim \text{GM}(R)$ instead of $g(R) \leftarrow \text{GM}(R)$ to indicate that the training algorithm can be a stochastic and non-deterministic process.

Approximation by features. It is tempting to assume that the model $\mathcal{D}_{g(R)}$ provides a perfect representation of the data distribution \mathcal{D}_R and that synthetic data “carries through all of the statistical properties, patterns and correlations in the [input] data” [27]. The trained model, however, only provides a *lower-dimensional approximation* of the true data distribution. It retains *some* characteristics but can never preserve all of them. Which characteristics are captured, and how they approximate \mathcal{D}_R , is determined by the generative model choice. *Statistical models*, such as Bayesian networks [35], or Hidden Markov models [23], provide an explicit, parametric model of \mathcal{D}_R . The features these models extract from their training data is determined upfront. *Non-parametric models*, such as generative adversarial networks (GANs) [25] or variational auto encoders (VAEs) [33], do not estimate a parametric likelihood function to generate new samples from $\mathcal{D}_{g(R)}$. Which features of the input data are most relevant and how the model approximates \mathcal{D}_R is implicitly determined during training [24].

The features a generative model uses to approximate \mathcal{D}_R define which of the statistical properties of the raw data R are replicated by a synthetic dataset $S \sim \mathcal{D}_{g(R)}$ sampled from the trained model. Statistical models provide some control over what features will be preserved. However, it is not possible to *exclude* that a synthetic dataset reproduces characteristics of the original data other than the features explicitly captured by the model. For instance, synthetic data generated through independent sampling from a set of 1-way marginals is expected to preserve a dataset’s independent frequency counts. However, if the raw data contains strong correlations between attributes, these correlations are likely to be replicated in the synthetic data even under independent sampling.

2.1 Generative models in this study

In Sections 4.3 and 6.2, we empirically evaluate the privacy gain of synthetic data publishing for five existing generative model training algorithms. We implemented three generative models without explicit privacy protection and two models with differential privacy guarantees. We chose models relevant to the tabular data sharing use case and to cover a wide range of model architectures. We further considered their computational feasibility for high-dimensional datasets and

whether a working implementation was available. Table 1 in Appendix 8.1 lists our parametrisation of these models.

IndHist. The IndHist training algorithm from Ping et al. [50] extracts marginal frequency counts from each data attribute and generates a synthetic dataset S through independent sampling from the learned marginals. Continuous attributes are binned. The number of bins is a configurable model parameter.

BayNet. Bayesian networks capture correlations between attributes by factorising the joint data distribution as a product of conditionals. The degree of the network model is a model parameter. The trained network provides an efficient way to sample synthetic records from the learned distribution (see Zhang et al. [71] for details). We use the GreedyBayes implementation provided by Ping et al.’s DataSynthesizer [50].

PrivBay. PrivBayes [71] is a differentially private Bayesian network model. Both, the Bayesian network and the conditional distributions, are learned under ϵ -differentially private algorithms. A synthetic dataset can be sampled from the trained model without any additional privacy budget cost. We use the GreedyBayes procedure provided by Ping et al. [50] to train a differentially private version of BayNet such that a PrivBay-trained model with $\epsilon \rightarrow \infty$ corresponds to a BayNet-model without formal guarantees.

CTGAN. CTGAN [66] uses mode-specific normalisation of tabular data attributes to improve the approximation of complex distributions through GANs. CTGAN further uses a conditional generator and training-by-sampling to get better performance on imbalanced datasets.

PATEGAN. PATEGAN builds on the Private Aggregation of Teacher Ensembles (PATE) framework [48] to achieve DP for GANs [32]. PATEGAN replaces the discriminator’s training procedure with the PATE mechanism. The trained model provides (ϵ, δ) -DP with respect to the discriminator’s output.

3 Quantifying the privacy gain of synthetic data publishing

The promise of synthetic data is that it allows data holders to publish (synthetic) datasets that are useful for analysis while, at the same time, protect the privacy of individuals in the raw data against powerful adversaries [2, 5, 10, 16, 66–68]. The increasing number of applications of synthetic data tools shows that this has become an appealing proposition [15, 46, 61, 62, 64, 72, 74]. Here, we introduce a novel evaluation framework that allows data holders to quantitatively assess claims about the privacy benefits of synthetic data sharing.

Synthetic data as an anonymisation mechanism. Synthetic data providers often present synthetic data as a novel “data anonymisation solution” [74] that addresses the shortcomings of traditional sanitisation techniques, such as generalisation [39, 58] or perturbation [43]. Data holders are promised

that publishing a synthetic in place of the raw dataset prevents the leakage of private information about individuals in the raw data previously observed in sanitised datasets [43, 58].

To evaluate this claim, that synthetic data generation is an effective anonymisation mechanism, we hence need to assess whether synthetic data addresses the privacy risks that originally motivated the use of data anonymisation techniques. These are the risk of *linkability* and *inference* [3]. Previous anonymisation methods such as k-anonymity [58] or l-diversity [39] have failed to provide robust protection against these attacks for high-dimensional, sparse datasets [43]. So far, however, there is no evidence that synthetic data provides *better protection* against these attacks at a lower cost in utility than traditional sanitisation techniques.

A number of recent papers have tackled related problems but focus primarily on non-parametric models for synthetic image generation and adversaries with white-box or query access to the model [8, 26, 28]. However, like sanitisation, synthetic data is primarily seen as a tool for privacy-preserving tabular data sharing, i.e., to enable data holders to publish a single copy of synthetic data as opposed to the trained model or a set of statistics [15, 46, 64, 72, 74]. In our framework, we hence assume that the adversary only has access to a synthetic dataset and can not repeatedly query the trained model or observe its parameters.

3.1 Evaluation framework

The goal of our framework is to quantitatively assess whether publishing a synthetic dataset S in-place of the raw data R reduces the privacy risks for individuals in the raw data with respect to the relevant privacy concerns. We model each privacy concern as an adversary \mathcal{A} that given a raw or synthetic dataset aims to infer a secret about a target record \mathbf{r}_t from the population \mathcal{R} . For each adversary, we define an advantage measure $Adv^{\mathcal{A}}$ that captures by how much including an individual’s record in the published data increases this individual’s privacy risk. In Section 4.1 and 6.2, we define adversaries and advantage measures that model the risk of linkability and inference, respectively.

Privacy gain. We assess the privacy gain of publishing a synthetic dataset S in place of the raw data R for target record \mathbf{r}_t as the reduction in the adversary’s advantage when given access to S instead of R

$$PG \triangleq Adv^{\mathcal{A}}(R, \mathbf{r}_t) - Adv^{\mathcal{A}}(S, \mathbf{r}_t). \quad (1)$$

A high privacy gain indicates that publishing S in place of R substantially reduces the privacy risk modelled by adversary \mathcal{A} for target record \mathbf{r}_t . A low gain, in contrast, implies that the data holder’s decision to publish S or R has no impact on the privacy loss for target record \mathbf{r}_t , i.e., the adversary’s advantage remains the same.

The privacy gain hence allows us to assess whether synthetic data is, as promised, an effective anonymisation mech-

anism. A good anonymisation mechanism should result in a high privacy gain for all records in the population and under any potential privacy adversary. A low gain in privacy indicates that the anonymisation mechanism does not provide a *significant improvement over publishing the raw data*.

Comparison to previous evaluation approaches. In contrast to model-specific evaluation techniques [30], our framework treats the data generating mechanism as a *complete black-box* and evaluates the privacy risks of *synthetic data publishing* rather than an adversary’s inference power when given query or white-box access to a model [26, 53]. As opposed to privacy evaluations based on similarity metrics [10, 67, 68], the framework provides data holders with a direct measure of how well the synthetic data defends against the privacy risks of data sharing. This makes the evaluation results easily interpretable and relatable to relevant data protection regulations [3]. We design and implement the framework in a modular fashion. This ensures that the framework is not limited to a specific threat model [53] or the privacy risks modelled in this paper. Instead, it can be adapted to any privacy concern specific to the data holder’s use case. The proposed evaluation method is independent of the data generation method. Thus, it can be used to evaluate the privacy gain of synthetic data generated by models trained without any explicit privacy protection, models trained under formal privacy guarantees [1, 6], or traditional anonymisation techniques.

Worst-case vs. average-case evaluation. Finally, previous studies have shown that the privacy risks of data sharing are not uniformly distributed across the population [36, 38, 54]. While individuals that are representative of a large majority of the population are often protected from privacy attacks, outliers or members of minorities largely remain vulnerable. Our framework allows to assess privacy risks both at an aggregate population-level and on a per-record basis. This enables us to demonstrate that synthetic data provides disparate privacy gain across population subgroups.

4 Does synthetic data mitigate the risk of linkability?

A major privacy concern in the context of privacy-preserving data sharing is the risk of linkability. Linkage attacks aim to link a target record to a single record, or group of records, in a sensitive dataset. Linkage enables adversaries to attach an identity to a supposedly de-identified record [43, 58] or to simply establish the fact that this particular record is present in a sensitive dataset [29].

Related work. The risk of linkability has been demonstrated, in theory and practice, for a large variety of data types: tabular micro-level datasets [43, 58], social graph data [42, 44], aggregate statistics [51], and statistical models [57]. Linkage attacks on tabular microdata usually intend to link a target

record (connected to an identity) to a single record in a sensitive database from which direct identifiers have been removed.

Membership inference attacks (MIAs) are linkage attacks which target the output of statistical computations run on sensitive datasets, such as aggregate statistics [29, 51] or trained ML models [57]. ML-oriented MIAs have been extensively studied on predictive models, such as binary or multi-label classifiers [37, 56, 57, 69]. Recently, this work has been extended to GANs and VAEs [8, 26, 28].

4.1 Formalizing linkability as membership inference

In a linkage attack, the adversary aims to learn whether a record is present in a sensitive dataset. Following works by Yeom et al. [69] and Pyrgelis et al. [51], we hence model the risk of linkability as a membership privacy game between an adversary \mathcal{A} and a challenger \mathcal{C} . The challenger plays the role of a data holder that publishes a dataset X that is made available to the adversary. This dataset could either be a raw dataset R or a sanitised or synthetic version of R , denoted as S . The goal of the adversary \mathcal{A} is to infer whether a target record \mathbf{r}_t , chosen by the adversary, is present in the sensitive dataset R based on the published dataset X and some prior knowledge \mathcal{P} .

Fig. 1 presents the linkability game for the case where S is a synthetic dataset sampled from a generative model trained on R . Later, we discuss how the challenger’s protocol changes when the game models sanitisation. First, \mathcal{A} picks a target record \mathbf{r}_t and sends it to \mathcal{C} . \mathcal{C} draws a raw dataset R of size $n - 1$ from the distribution defined by the population \mathcal{R} , and a secret bit $s_t \sim \{0, 1\}$. If $s_t = 0$, \mathcal{C} draws a random record \mathbf{r}_* from the population (excluding the target) and adds it to the raw dataset. If $s_t = 1$, \mathcal{C} adds the target \mathbf{r}_t to the raw dataset. Then, \mathcal{C} trains a generative model on the raw data R , and samples a synthetic dataset S of size m from the trained model. \mathcal{C} picks at random whether to send back to the adversary the raw data R or the synthetic data S . \mathcal{A} receives the dataset and makes a guess about the target’s presence in R , $\hat{s}_t \leftarrow \mathcal{A}^\mathcal{L}(X, b, \mathbf{r}_t, \mathcal{P})$. The adversary wins the game if $\hat{s}_t = s_t$.

As in Yeom et al. [69], we assume an equal prior over the target’s membership in R and define the linkage adversary’s advantage as:

$$Adv^\mathcal{L}(X, \mathbf{r}_t) \triangleq 2\mathbb{P}\left[\mathcal{A}^\mathcal{L}(X, b, \mathbf{r}_t, \mathcal{P}) = s_t\right] - 1 \quad (2)$$

$$= \mathbb{P}[\hat{s}_t = s_t | s_t = 1] - \mathbb{P}[\hat{s}_t = s_t | s_t = 0] \quad (3)$$

where X can be a raw R or synthetic S dataset. The probability space of $Adv^\mathcal{L}$ is defined by the random choices of $R \sim \mathcal{D}_\mathcal{R}$ and $s_t \sim \{0, 1\}$ and the randomness of the synthetic data generation mechanism and the adversary’s guess.

Adversarial strategy. The adversary’s guess function $\mathcal{A}^\mathcal{L}(\cdot)$ takes as input a target record \mathbf{r}_t , the information published by the challenger, X and b , and some prior information \mathcal{P} and

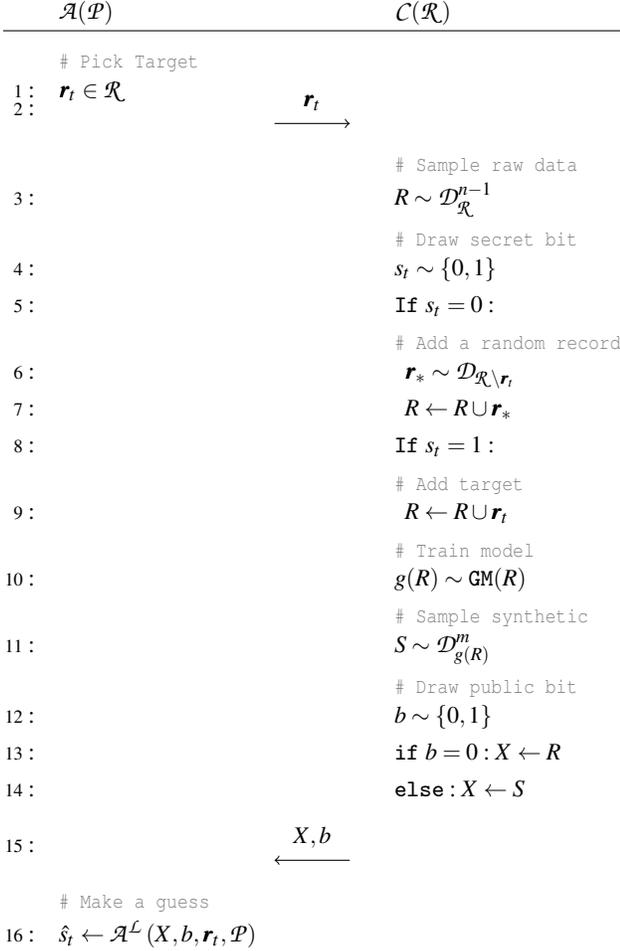


Figure 1: Linkability privacy game.

outputs a guess about the target’s presence in R . The adversary’s strategy to make a guess changes depending on the data published.

If C publishes $X = R$, the adversary simply checks whether $\mathbf{r}_t \in R$ and has a probability of 1 to win the game ($Adv^{\mathcal{L}}(R, \mathbf{r}_t) = 1$).

If $X = S$, the adversary performs a binary classification task on a set of features extracted from the synthetic data S . We explain how we implement this binary classifier in Section 4.2.

Privacy gain. For each target chosen by the adversary, we instantiate the game multiple times and measure the adversary’s advantage conditioned on the challenger’s choice of X . Under our definition of privacy gain in Eq. 1 and with $Adv^{\mathcal{L}}(R, \mathbf{r}_t) = 1$, the privacy gain of publishing a synthetic dataset S in place of the raw data with respect to the risk of linkability is given as:

$$PG = 1 - Adv^{\mathcal{L}}(S, \mathbf{r}_t) \quad (4)$$

A privacy gain of $PG = 0$ indicates that the adversary infers the target’s presence in R with perfect accuracy regardless of whether given access to the raw or synthetic data. If on the

other hand, observing the synthetic data S gives the adversary no advantage in inferring the target’s presence ($Adv^{\mathcal{L}}(S, \mathbf{r}_t) = 0$), then $PG = 1$.

4.2 A black-box membership inference attack

We implement the adversary’s strategy as a *generic black-box* MIA that is independent of the generative model architecture.

Related work. Existing MIAs on generative models focus almost exclusively on non-parametric deep learning models for synthetic image generation [8, 26, 28, 41]. These works mostly investigate the privacy risks of either model-specific *white-box* attacks or *set membership* attacks that assume the adversary has access to the entire universe of training records and come to the conclusion that black-box MIAs that target specific records perform only slightly better than random baseline guessing [26, 28]. Unfortunately, previous attacks do not provide a good basis to evaluate the privacy gain of synthetic data publishing. Non-parametric models for non-tabular data cover only a very small set of use cases [15, 46, 72, 74], white-box attacks do not adequately reflect the data sharing scenario, and set inference attacks are not suitable to assess individual-level privacy gain.

Shadow model attack. In order to win the linkability game (see Fig. 1) when she receives a synthetic dataset S , the adversary needs a distinguishing function $\mathcal{A}^{\mathcal{L}}(\cdot)$ that enables her to infer the membership of \mathbf{r}_t in the raw data R used to train the generative model that output S . As in many previous works, we cast membership inference as a supervised learning problem and instantiate the adversary’s guess function with a machine learning classifier trained on data produced by generative shadow models [51, 57].

As Shokri et al. [57], we assume that, as part of her prior knowledge \mathcal{P} , the adversary has access to the training algorithm $\text{GM}(\cdot)$, the size of the raw and synthetic datasets n and m , and to a reference dataset $R_{\mathcal{A}} \sim \mathcal{D}_{\mathcal{R}}^l$ that comes from the same distribution as the target model’s training data $R \sim \mathcal{D}_{\mathcal{R}}^n$ and may or may not overlap with R . Given this prior knowledge \mathcal{P} and a target record \mathbf{r}_t , the adversary uses the following procedure to learn $\mathcal{A}^{\mathcal{L}}$: First, the adversary samples multiple training sets R_i of size n from the reference dataset $R_{\mathcal{A}}$. On each set R_i , the adversary trains a generative model $g(R_i)$ using the training procedure $\text{GM}(R_i)$. From each of the trained models, the adversary samples multiple synthetic datasets S of size m and assigns them the label $s_t = 0$. The adversary repeats the same procedure on the same training sets, this time including the target, $R'_i = R_i \cup \mathbf{r}_t$, and assigns the generated synthetic datasets the label $s_t = 1$. Finally, the adversary trains a classifier $\mathcal{A}^{\mathcal{L}}$ on the labelled datasets. The trained classifier takes as input a synthetic dataset S and outputs a guess \hat{s}_t about the target’s presence in R : $\hat{s}_t \leftarrow \mathcal{A}^{\mathcal{L}}(S, \mathbf{r}_t, \mathcal{P})$.

Feature extraction. Existing MIAs on predictive models leverage patterns in the confidence values output by a trained model that differ between two classes, members and non-

members [57]. Mounting a successful black-box MIA on a generative model is much more challenging [26]. The attacker needs to identify the influence that a *single target record* has on the high-dimensional data distribution $\mathcal{D}_{g(R)}$ as opposed to a low-dimensional confidence vector. Moreover, the output sampling process introduces additional uncertainty and the adversary only has access to a single output example.

In other words, the adversary needs to be able to distinguish between two distributions, $\mathcal{D}_{g(R \cup r_s)}$ and $\mathcal{D}_{g(R \cup r_r)}$, given a single observation $S \sim \mathcal{D}_{g(X)}^m$. To reduce the effect of high-dimensionality and sampling uncertainty, the adversary can apply *feature extraction techniques*. Instead of training a classifier directly on S , the adversary learns to distinguish feature vectors extracted from synthetic datasets produced by models trained with and without the target, respectively. A feature set can be described as a function $f(X) = \mathbf{f}$ that takes as input a set of records X from the high-dimensional data domain and outputs a numerical vector \mathbf{f} that maps X into a lower-dimensional feature space. Whether the attack using feature set \mathbf{f} is successful depends on two factors: First, whether the target’s presence has a detectable impact on any of the features in \mathbf{f} , and second, whether the synthetic dataset has preserved these features from the raw data and hence preserved the target’s signal.

Implementation. We implement the distinguisher function \mathcal{A}^L as an instantiation of our framework’s `PrivacyAttack` class (see Appendix 8.1). We leverage the object-oriented structure of the library to create multiple attack versions that share the same training procedure but use different attack models and feature extraction techniques. As feature extractors, we implemented a naive feature set with simple summary statistics F_{Naive} , a histogram feature set that contains the marginal frequency counts of each data attribute F_{Hist} , and a correlations feature set that encodes pairwise attribute correlations F_{Corr} (see Appendix 8.1).

As attack models, we implemented a Logistic Regression, Random Forests and K-Nearest Neighbours classifier. All attack models yielded similar results with a Random Forests classifier with 100 estimators using the Gini impurity splitting criterion performing best across datasets, generative models, and feature sets. In the remainder of the paper we focus on results obtained using this classifier.

4.3 Empirical evaluation

We first evaluated the expected privacy gain with respect to the risk of linkability under the three generative models trained without any formal privacy (see Section 2.1) on two common benchmark datasets: `Adult` and `Texas`. Both are tabular datasets that contain a mix of numerical and categorical attributes. A detailed description of their characteristics can be found in Appendix 8.2.

Experiment procedure. We aim to assess whether synthetic data produced by a wide range of generative model types does,

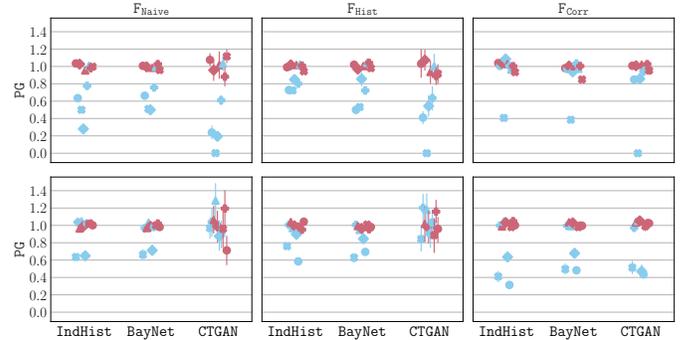


Figure 2: Expected per-record privacy gain for outliers \bullet and random records \bullet for the `Texas` (top row) and `Adult` (bottom row) datasets under three different attacks using three distinct feature sets. Error bars represent the standard deviation.

as claimed, provide robust protection against linkage attacks. If synthetic data is a “valid, privacy-conscious alternative to raw data” [5], then its privacy gain should be close to $PG = 1$ for all target records regardless of the attacker’s strategy. We evaluate two groups of targets: five records randomly chosen from the population and five manually chosen outlier records representative of population minorities and most likely to be vulnerable to linkage attacks [43]. As outliers, we selected records that either have rare categorical attribute values or numerical values outside the attribute’s 95% quantile. For instance, in the `Texas` dataset we show the privacy gain for two records that have high total charges and one record with high total non-covered charges outside the attribute’s 95% quantile, and two records with an unusually high risk mortality and illness severity. For the `Adult` dataset we followed the same procedure to select outlier records.

At the beginning of each experiment, we sample a fixed reference dataset $R_{\mathcal{A}}$ of size l from the population and use it to train the adversary’s distinguisher. For each target record, we train multiple attack models using the shadow model training procedure described in Section 4.2. To assess privacy gain, we repeatedly instantiate the linkability game described in Fig. 1 for each of our ten targets.

Disparate gain. Fig. 2 shows the average privacy gain across multiple instantiations of the linkability game for five outlier targets \bullet and five randomly chosen targets \bullet for the `Texas` (top row) and `Adult` (bottom row) datasets, respectively. Each dataset, raw and synthetic, contained $n = m = 1000$ records. The adversary was trained on a reference dataset of $l = 10,000$ records using 10 shadow models.

We find that in both datasets privacy gain is unevenly distributed across target records. While the control group of randomly chosen target records (\bullet) achieves close to perfect protection ($PG \approx 1$), other records (\bullet) remain highly vulnerable to our linkage attack. The privacy gain for the majority of outlier targets is substantially smaller than $PG = 1$ (ideal case). For instance, under an attack using the naive feature set F_{Naive}

4 out of the 5 selected targets in the `Texas` dataset achieve an average gain smaller than 0.8 across all three generative models. More worryingly, 1 out of the 5 targets tested (\star) consistently receives a privacy gain close to 0 ($PG < 0.005$) from synthetic data produced by a CTGAN-trained model. These results indicate that, contrary to claims by previous works [26], publishing the synthetic in place of the raw data does not protect outlier targets from linkage attacks.

Unpredictable gain. Which records remain at risk varies across generative model type and the adversary’s feature set. In the `Texas` dataset, an attack using the F_{Naive} feature set on CTGAN-produced synthetic data results in a privacy gain below $PG < 0.3$ for 3 out of the 5 outlier targets (\bullet , \blacklozenge , \star). The same attack on the same targets is less effective on synthetic data produced by IndHist-trained models. The same group of targets reaches a maximum gain of $PG = 0.77$. Attacks on the `Adult` dataset are most successful under the correlations feature set F_{Corr} . Here, IndHist-trained models provide a minimum gain of $PG = 0.64$ (\star) under F_{Naive} . The minimum gain provided by the same model drops below $PG < 0.32$ if the attacker uses F_{Corr} as input to the attack and leaves a different target (\blacklozenge) most vulnerable.

Conclusions. These results are extremely problematic from the point of view of a data holder seeking to use synthetic data generation as a privacy mechanism. Ideally, data holders should be able to predict, given a fixed dataset R and a generative model training algorithm $GM(\cdot)$, the minimum gain in privacy they can achieve. Our experiment shows, however, that this is next to impossible: The level of protection a generative model provides depends on how much information the model’s output leaks about the features targeted by the attack. This means that we can only predict privacy gain if we can (1) predict what features a potential adversary will target and (2) whether the synthetic data has preserved these features from the raw data. In practice, neither of these factors is predictable. First, like traditional linkage attacks on microdata releases, a strategic adversary might use any set of features that are likely to be influenced by the target’s presence [43]. Second, which characteristics a synthetic dataset might preserve is not constrained to the features explicitly represented by the model. For instance, even the simplest statistical model IndHist might unexpectedly preserve features targeted by the attack: Synthetic data produced through independent attribute sampling by an IndHist-model trained on the `Adult` dataset leaves some target records vulnerable to linkage attacks using the correlations feature set F_{Corr} . Non-parametric models, such as CTGAN, do not even provide a parametric specification for the data’s density function. This makes it even harder to predict what set of features the model will preserve and an attack might target.

Previous assessments of the privacy risks of synthetic data publishing based on aggregate population measurements severely underestimate the risk of linkage attacks [22, 26, 28].

Our experimental evaluation reveals that synthetic data does not provide uniform protection against strategic adversaries, and some outliers remain highly vulnerable.

5 Does differentially private synthetic data mitigate the risk of linkability?

In the previous section, we demonstrate that non-private data synthesis algorithms are largely unsuitable as privacy mechanisms. This is not an unexpected finding as none of the evaluated models were originally designed as anonymisation mechanisms. In this section, we thus extend our analysis to two model training algorithms explicitly designed to protect the privacy of a model’s training set, PrivBay and PATEGAN. We evaluate to which extent their formal privacy guarantees improve the privacy gain of synthetic data publishing with respect to the risk of linkability.

Differentially private generative models. Model training algorithms based on the differential privacy model protect the privacy of the training data through formal guarantees for the *lower-dimensional approximation of the full-dimensional data distribution* [6, 22, 32, 71]. Synthetic datasets drawn from differentially private models preserve these privacy guarantees under the post-processing guarantee [17].

The model training algorithm PrivBay learns a differentially private Bayesian network that approximates the relationship between data attributes via the exponential mechanism and computes the conditionally independent marginals in the subspaces of the Bayesian network via the Laplace Mechanism [71]. PATEGAN, a differentially private GAN, ensures that the discriminator’s decisions are not affected by the presence of a single record in the model’s training set by more than the defined ϵ -bound [32].

5.1 Empirical evaluation

We used the experimental procedure described in Section 4.3 to evaluate the privacy gain of PrivBay and PATEGAN. We integrated the implementations of these algorithms provided by Ping et al. [12] and Jordon et al. [70] into our framework and ran the linkability game defined in Fig. 1.

Differential privacy violations. Fig. 3 left shows the results of this experiment for the `Texas` (top row) and `Adult` (bottom row) datasets under an attack using the histogram feature set F_{Hist} – the overall most effective attack. Both differentially private models were trained with privacy parameter $\epsilon = 0.1$. Surprisingly, we find that neither the original implementation of PrivBay nor PATEGAN reliably prevents linkage attacks. Two out of the five outliers in the `Texas` dataset achieve close to no gain (\bullet and \blacklozenge with $PG < 0.1$). This low gain violates the theoretical lower bound on privacy provided by Yeom et al. [69] (shown as a dashed line in Fig. 3). The bound given by Yeom et al. [69] limits the expected advantage of the

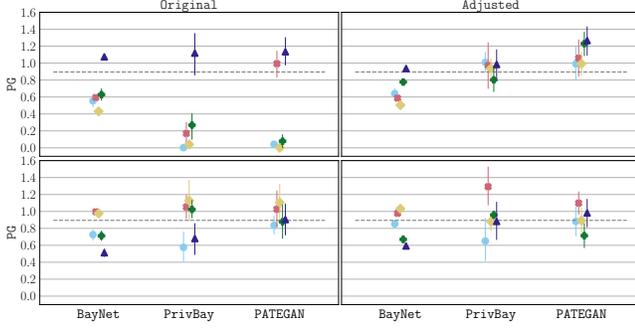


Figure 3: Per-record privacy gain for five outlier targets records from the `Texas` (top row) and `Adult` (bottom row) datasets under an attack using the F_{Hist} feature set.

membership inference adversary to $Adv^{\mathcal{L}} \leq \epsilon^e - 1$ which implies $PG \geq 0.89$ for $\epsilon = 0.1$.

Unexpected leakage. To understand these findings, we conducted an in-depth analysis of the design and implementation of both algorithms. Alongside some minor bugs, our analysis revealed that, while both models on paper fulfil their formal privacy definitions, their available implementations did not. Both `PrivBay` and `PATEGAN` require metadata about a model’s training set to operate. For instance, data holders need to specify *upfront* the range of numerical attributes and the possible values of categorical attributes. To improve usability, the existing implementations of `PrivBay` and `PATEGAN` learn this metadata *directly* from the input dataset. This process, in which algorithmic decisions are based on the raw data, violates important assumptions of the differential privacy model [18]. This discovery explained our previous results: Both models extracted crucial information from their input data in a process *not covered by the formal privacy guarantee*. As a result, targets with rare categorical attributes or whose presence affects the ranges of numerical attributes remained highly vulnerable to our attack. Their presence in a model’s training set became detectable due to the occurrence of new categories or a shift in the ranges of continuous attributes in synthetic datasets sampled from the trained model.

To avoid this leakage, we patched the `PrivBay` and `PATEGAN` implementations so that both models obtain metadata as an *independent* input to their training process. In our experiments with the `Texas` dataset, we used the publicly available data description to define possible categorical values and a disjoint subset of the data (population records from a different year) to obtain an estimate of the expected ranges of numerical attributes. For the `Adult` dataset, where no comparable metadata is available, we used the dataset to estimate categories and ranges and generalised each range to hide the exact value of outlier targets.

Fig. 3 right shows the results of our evaluation under the patched implementations. For most outliers, privacy gain is now bounded by its differential privacy guarantee. For those targets where the expected gain remains below its bound,

the remaining gap can likely be explained either by other aspects of the model’s implementation that violate theoretical assumptions and we were not able to find in our analysis, or due to correlations between the datasets used to derive the necessary metadata and the model’s training set. Further work is needed to fully understand this problem.

Conclusions. Differentially private generative models can provide a significantly higher privacy gain with respect to linkage attacks than traditional data synthesis algorithms. To achieve the desired protection it is necessary that, besides a theoretically sound design, the models’ implementation and operational environment does not break any of the privacy definition’s theoretical assumptions. Our evaluation confirms that otherwise there is no guarantee that outliers will be protected from linkage attacks.

While in our experimental setup we were (mostly) able to avoid undesired privacy leakage through metadata, it is unlikely that in practice data holders will be able to follow our example. Data holders likely do not have access to either a disjoint subset or a public dataset from the same distribution that would allow them to define metadata that fits the raw data they would like to share. Synthetic data sharing is often motivated by the *unique value* of sensitive dataset that are limited in size. This implies that, in practice, data holders might struggle to achieve the desired strict privacy guarantees, or face a large utility loss when either using public data or splitting the available data to derive the necessary metadata. In Section 6.3, we empirically demonstrate this tradeoff.

6 Does synthetic data improve the privacy-utility tradeoff of sanitisation?

Synthetic data is often presented as “a new, better alternative to sanitised data release” [5]. Our results in previous sections confirm Bellovin et al.’s hypothesis that data synthesis algorithms without any formal privacy guarantees leave outliers vulnerable to linkage attacks [5]. Differentially private generative models, although hard to implement, reduce these risks. We now investigate whether “synthetic data may be combined with differential privacy to achieve a best-of-both-worlds [privacy and utility] scenario.” [5]. In other words, whether synthetic data can achieve a higher gain in privacy *at a lower cost in utility* compared to traditional sanitisation.

In this section, we assess the privacy-utility tradeoff of synthetic data publishing and compare it to that of traditional sanitisation. For our comparison, we implement a sanitisation procedure described by NHS England [46] and assess its privacy gain with respect to the risk of linkability (see Section 4.1) and the risk of inference formalised in Section 6.2.

NHS Sanitisation procedure. A sanitisation procedure $S \leftarrow \text{San}(R)$ is a deterministic function that applies a set of predefined row-level transformations to the input data R to produce a sanitised dataset S that fulfils a heuristic privacy defi-

dition [58]. Common transformations are generalisation, perturbation, or deletion of single rows [3]. Following the details given in [46], we implemented a simple sanitisation procedure San that reduces the granularity of categorical attributes through grouping, generalises any granular timing or geographical information, removes any rows with rare categorical values, caps numerical values to the attribute’s 95% quantile, and enforces k -anonymity for a pre-defined set of demographic attributes.

6.1 Privacy gain with respect to linkability

To compare the privacy gain of sanitised and synthetic data publishing, we repeat the experimental procedure from Section 4.3. We adapt the game so that the challenger C , instead of generating a synthetic dataset S from a trained model $g(R)$ (lines 10 to 11 in Fig. 1), produces a sanitised version of R through a pre-defined sanitisation procedure $S \leftarrow \text{San}(R)$. When the adversary receives a sanitised dataset ($b = 1$), she first attempts literal record linkage. Only if the adversary can not uniquely identify a record that matches the target, she attempts classification. As in previous sections, we first train the adversary on a reference dataset $R_{\mathcal{A}}$ and then instantiate the game multiple times for each of the selected targets.

Fig. 4 compares the results of this experiment for the five outlier targets from the Texas dataset for three different data sharing mechanisms: traditional sanitisation San with $k = 10$, synthetic data produced by BayNet-trained models, and differentially private synthetic data sampled from PrivBay models with varying ϵ values. The same experiment on the Adult dataset yields similar results.

As expected, the privacy gain of row-level sanitisation tends to be binary: Target records that are likely to be removed from the shared dataset receive close to perfect gain (\blacktriangle and \blackstar with $PG \geq 0.8$ under all three feature sets). Others remain highly vulnerable to linkage attacks and receive a substantially lower gain (\blacklozenge , \blackdiamond , \blackplus with $PG \leq 0.3$ for at least one attack).

BayNet improves the privacy gain for the latter group: The three targets that under sanitisation receive close to no protection from linkage attacks using the naive feature set F_{Naive} obtain a higher minimum gain (\blacklozenge , \blackplus , \blackdiamond with $PG \geq 0.48$). Differentially private model training further improves protection and minimum gain increases as ϵ decreases ($PG = 0.77$ for target \blacklozenge under $\epsilon = 10$ and $PG = 0.97$ under $\epsilon = 1.0$). This indicates that synthetic data produced by either model (BayNet and PrivBay) hides changes in the raw data features caused by the target’s presence and prevents the adversary from inferring the target’s secret. This gain in privacy, however, is not constant across the population. One out of the five targets actually loses protection from linkage attacks when sharing a synthetic instead of the sanitised dataset (\blackstar with $PG = 0.62$ for BayNet and $PG = 0.91$ for PrivBay with $\epsilon = 10$ instead of $PG = 1.0$ for San under F_{Corr}).

This variability in privacy gain highlights one of the major

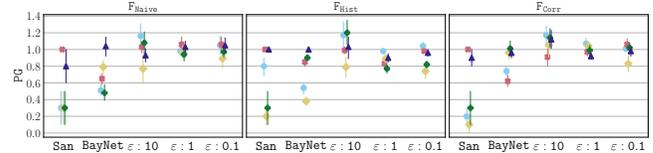


Figure 4: Per-record privacy gain for five outlier target records from the Texas dataset under three different attacks using three distinct feature sets.

drawbacks of synthetic data sharing as a privacy mechanism: unpredictability. Due to the deterministic nature of row-level sanitisation, the privacy gain of traditional anonymisation is largely predictable. The high gain in privacy for \blackstar and \blacktriangle under San is constant across all three feature sets. In contrast, the privacy gain under BayNet and PrivBay is much more variable. Before model fitting and an empirical analysis, it is not possible to predict whether an individual record’s signal will be preserved and what will be its minimum privacy gain.

6.2 Privacy gain with respect to attribute inference

The risk of linkability is not the only concern in the context of privacy-preserving data sharing (see Section 3). Data anonymisation also aims to protect individuals in the raw data from inference attacks. The risk of inference describes the concern that an adversary might “deduce, with significant probability, the value of an attribute from the values of a set of other attributes” [3].

6.2.1 Formalising attribute inference

To evaluate privacy gain with respect to attribute inference, we define a privacy game similar to the attribute inference experiment proposed by Yeom et al. [69].

In the attribute inference game, shown in Fig. 5, the adversary only has access to a *partial* target record $\tilde{r}_t = (r_1, \dots, r_{k-1})$ and aims to infer the value of a sensitive, unknown attribute r_s . At the start of the game, the adversary picks a target from the population $\tilde{\mathcal{R}}$, a set of records from the same domain as \mathcal{R} but with the sensitive attribute removed. The challenger receives the partial target record and assigns it a secret value $r_s \leftarrow \phi(\tilde{r}_t)$ where ϕ represents the projection of a partial record from \mathcal{R} into the domain of the sensitive attribute according to the distribution $\mathcal{D}_{\mathcal{R}}$. C then follows the same procedure as in the linkability game. The adversary obtains the dataset X and the public bit b and outputs a guess about the target’s sensitive attribute value \hat{r}_s . This game can be easily adapted for sanitisation by replacing lines 12 and 13 to produce a sanitised version of R through a pre-defined sanitisation procedure, $S \leftarrow \text{San}(R)$.

Similar to Eq. 2, we define the adversary’s advantage to assess the leakage of publishing dataset X with respect to

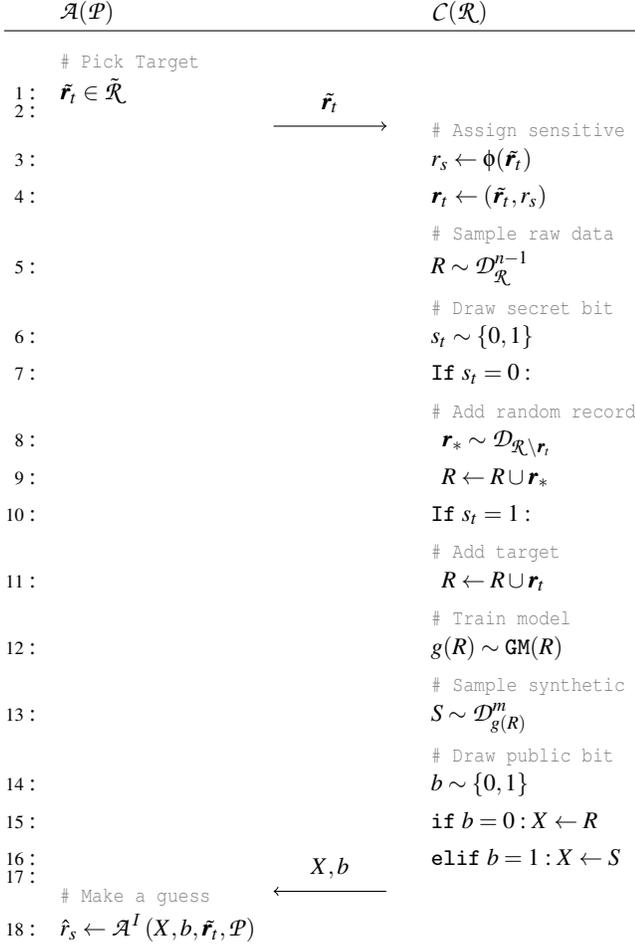


Figure 5: Attribute inference privacy game

attribute inference as:

$$Adv^I(X, \tilde{\mathbf{r}}_t) \triangleq \mathbb{P}[\hat{r}_s = r_s | s_t = 1] - \mathbb{P}[\hat{r}_s = r_s | s_t = 0] \quad (5)$$

where $\hat{r}_s = \mathcal{A}^I(X, b, \tilde{\mathbf{r}}_t, \mathcal{P})$ is the adversary’s guess about the target’s sensitive attribute r_s given dataset X and prior knowledge \mathcal{P} .

Adversarial strategy. The procedure to estimate Adv^I and the adversary’s strategy to make a guess about the target’s sensitive value depends on the domain of the sensitive attribute r_s , the value of the public bit b and whether S is a synthetic or sanitised dataset. When \mathcal{C} publishes a raw or sanitised dataset, the adversary first attempts to infer the missing value via record linkage [16, 39, 52]. If the adversary can link the target to a unique record in the dataset X based on its known attributes, she can reconstruct the target’s missing value with probability $\mathbb{P}[\mathcal{A}^I(X, b, \tilde{\mathbf{r}}_t, \mathcal{P}) = r_s | s_t = 1] = 1$.

When linkage fails, i.e., \mathcal{C} publishes a raw dataset without the target, a sanitised dataset that hides the target’s presence or a synthetic dataset, the adversary uses the published data to *train a supervised ML model to predict the target’s sensitive*

value based on the known attributes $\tilde{\mathbf{r}}_t$. To learn a mapping from known to sensitive attributes, the adversary splits the dataset X into two parts: A feature matrix \tilde{X} that contains the values for all attributes known to the adversary and a vector \mathbf{x}_s with the corresponding sensitive attribute values. Depending on the domain of \mathbf{x}_s , the adversary can either train a regression or classification model using \tilde{X} as input features and \mathbf{x}_s as labels. The trained attack model, denoted as $h(\cdot)$, takes as input a partial record containing the set of known attributes and outputs a guess about the label $\hat{x}_s \leftarrow h(\tilde{\mathbf{x}})$.

Implementation. We implement the adversary \mathcal{A}^I as an instantiation of our framework’s `PrivacyAttack` class (see Appendix 8.1). For continuous sensitive attributes with $r_s \in \mathbb{R}$, we implement the attack $h(\cdot)$ using a simple linear regression model provided by the `sklearn` library [49]. We centre all features extracted from the input data and fit a linear model without intercept. The model fits linear coefficients that minimise the root mean squared error between the observed and predicted target values. We analytically calculate the adversary’s probability of success $\mathbb{P}[\hat{r}_s = r_s | s_t]$ as the likelihood of the true value under the learned linear coefficients (see Appendix 8.3 for details). For categorical attributes, we use a simple Random Forests classifier as attack model and estimate the attack’s success via its classification accuracy.

Empirical evaluation. Fig. 6 shows the privacy gain of the three data release mechanisms, sanitisation via `San` and synthetic data produced by `BayNet` and `PrivBay`-models with varying ϵ values, for the five outlier targets from the `Texas` dataset for two distinct sensitive attributes. We chose one continuous (`LengthOfStay`) and one categorical attribute (`Race`) that might be considered sensitive patient information.

For the continuous attribute `LengthOfStay`, synthetic data produced by either `BayNet` or `PrivBay` provides close to perfect gain for all five targets while row-level sanitisation via `San` marginally reduces the adversary’s advantage for three out of the five targets (♣, ♠, ♦ with $PG \leq 0.2$). This implies that, in contrast to the sanitised datasets, synthetic data does not preserve the targets’ signal. Even when the generative model’s training set includes the target record, synthetic data sampled from the trained model does not contain any patterns that allow the adversary to infer the target’s sensitive value and hence $Adv^I(S, \mathbf{r}_t) \ll Adv^I(R, \mathbf{r}_t)$. When the attack targets the categorical attribute `Race`, the privacy gain of synthetic data publishing does not significantly increase over that of sanitised data publishing. Even differentially private generative model training with $\epsilon = 0.1$ does not guarantee a high privacy gain.

To explain this low gain, we plot in Fig. 7 for the target marked as ♣ the adversary’s probability of success on dataset X when the target is in the dataset (■ $s_t = 1$), when it is not (□ $s_t = 0$), and the resulting adversary’s advantage (■ Adv^I). The figure shows that the low privacy gain observed in Fig. 6 results from the fact that the adversary’s advantage

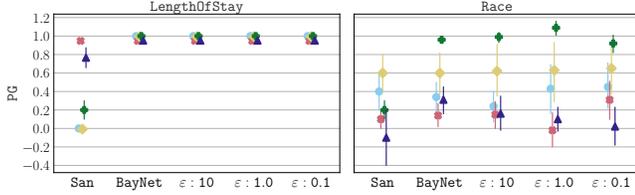


Figure 6: Per-record privacy gain for five outlier target records from the Texas dataset for two distinct sensitive attributes.

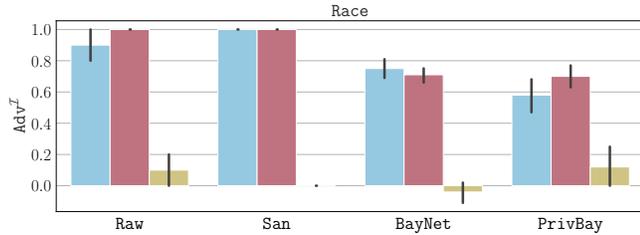


Figure 7: Probability of success and advantage for an attribute inference attack on attribute *Race* on the Texas dataset. \blacksquare $P[\hat{r}_s = r_s | s_t = 0]$, \blacksquare $P[\hat{r}_s = r_s | s_t = 1]$, and \blacksquare Adv^I . Data shown for *PrivBay* with $\epsilon = 1$.

is already small when the adversary receives the raw data R . Thus, publishing a sanitised or synthetic data instead of the raw data does not lead to any substantial gain in privacy for this target.

Conclusions. Depending on the attribute targeted by the attack, the privacy gain of synthetic and sanitised data publishing varies substantially. A low gain in privacy either indicates that the anonymised data still contains enough information specific to the target record to give the adversary a significant advantage or it implies that already publishing the raw data did not incur a significant privacy loss. Where necessary, our framework allows data holders to make a distinction between those two cases based on the reported privacy loss. In both cases, however, a low gain signals that publishing the synthetic or sanitised data does not provide any improvement over publishing the raw data R . We also observe that there are records for which privacy gain is high: the dataset published in place of the raw data successfully hides information about these records. While this might be good news for privacy, it comes, as we show in the next sections, at a cost in utility.

6.3 Utility loss comparison

The promise of synthetic data is that its improvement in privacy gain over traditional sanitisation comes at a negligible cost in utility. In this section, we empirically evaluate the utility loss of synthetic data sharing and compare it to that of row-level sanitisation.

Utility metrics. Besides its potential privacy benefits, sharing a synthetic in-place of the original dataset incurs certain risks, such as the risk of false conclusions [2] or the risk of exacerbating existing biases in the data [9]. The goal of our utility

evaluation is to assess to which extent the privacy gain of synthetic data observed in previous sections (see Section 6.1 and 6.2) increases these risks and reduces data utility.

The concepts of utility and utility loss are highly dependent on the data use case and different utility metrics might yield vastly different results [55, 66]. Therefore, in practice data holders should conduct their own evaluation based on appropriate utility definitions when weighing off the risks and benefits of (anonymised) data sharing.

In this work, we chose a set of simple utility function that aim to cover a wide range of synthetic data use cases suggested in the literature and reported to us by practitioners. First, we evaluate in Section 6.3.1 the utility of synthetic data for use cases that rely on aggregate population metrics, such as reporting of summary statistics or training machine learning models [55, 73]. Second, in Section 6.3.2, we turn to one of the main selling points brought forward by proponents of synthetic data as a privacy technology: That it enables the analysis of more fine-grained statistical patterns than aggregate query release mechanisms, including the analysis of outliers. Financial fraud and medical anomaly detection are two of the most commonly suggested synthetic data use cases largely based on the analysis of outliers [40, 63].

The latter class of use cases further motivates us to focus on the privacy gain and utility loss of outlier records. A simple way to improve privacy gain for these most vulnerable records would be to remove them from the dataset, as the high privacy gain for targets \blacktriangle and \blacktriangleleft under *San* in Fig. 4 demonstrates. Directly removing vulnerable records from the raw data increases privacy gain but severely impacts the data’s utility, e.g. in healthcare “this can sometimes mean missing out on important data that could be used to help future patients” [63]. For this reason, we assess in Section 6.3.2 whether synthetic data leads to a high privacy gain for outlier records while preserving their utility benefits for similar records.

6.3.1 Average utility loss

Machine learning utility. We first measure the utility loss of publishing a synthetic or sanitised dataset S in place of the raw data R as the decrease in average prediction accuracy over a hold out set for a prediction model trained on S instead of R . Due to the limited size of the *Adult* dataset, we focus our evaluation on the Texas dataset. We created a pre-processed dataset that contained all publicly available inpatient records for the years of 2013 and 2014. We denote records from 2013 as the train population \mathcal{R}_{Train} , and records from 2014 as the test set \mathcal{R}_{Test} . In each experiment, we sample a raw dataset of size n from \mathcal{R}_{Train} and use it as training set for a prediction model $h_R(\cdot) \sim PM(R_{Train})$. We then train a generative model $g(R_{Train})$ on the same raw dataset and sample multiple copies of synthetic data S_i from this generative model. We use these synthetic datasets to train classifiers $h_S(\cdot)$. We follow an analogous procedure for sanitisation, where we sanitise R_{Train}

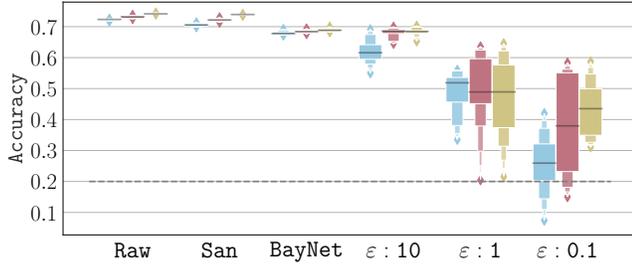


Figure 8: Test accuracy of prediction models for attribute RiskMortality in the Texas dataset for three dataset sizes ($n = 1000$, $n = 2000$ and $n = 5000$). The grey dotted line shows the random guess baseline.

according to the NHS sanitisation procedure before training the classifier. Finally, we evaluate the accuracy of each of the trained classifiers $h_X(\cdot)$ on a test set sampled from \mathcal{R}_{Test} .

Fig. 8 shows the results of this experiment for a multi-classification task on attribute RiskMortality with 5 classes for three different training set sizes ($n = 1000$, $n = 2000$, and $n = 5000$). Fig. 15 in the Appendix 8.4 shows the corresponding privacy gain for various dataset sizes. The random baseline guess rate is indicated with a dotted grey line. With $n = 1000$, classifiers trained on the raw data $h_R(\cdot)$ achieve an average test accuracy of 72.2% which is comparable to models trained on the sanitised dataset with 70.5%. Using a synthetic version of the data as training set leads to a significant utility loss: The classifier’s average test accuracy drops to 68.0% when trained on synthetic data produced by BayNet. Differentially private model training further widens this gap. Even with privacy parameter values as high as $\epsilon = 10$ the classifier’s mean accuracy with 62.0% remains 10 points below that of a model trained on the raw data. Increasing the model’s training set size has a slight effect on utility. While classifiers trained on the raw data achieve an average accuracy of 74.2% for $n = 5000$, average accuracy under synthetic data produced by BayNet and PrivBay with $\epsilon = 10.0$ increases to 68.9% and 68.4%, respectively.

Summary statistics. We use the discrepancy of the mean (Fig. 9 left) and median (right) of three continuous attributes between raw, sanitised, and synthetic datasets as additional utility loss measures. These simple summary statistics are common metrics used for reporting. While data sanitisation largely preserves the raw data’s statistics, synthetic datasets sampled from all models significantly differ from the raw data. For instance, the empirical mean of attribute TotalChargesAccomm in raw datasets R sampled from the population \mathcal{D}_R ranges between 9K and 11K. In contrast, synthetic data sampled from BayNet models trained on the raw data produces values ranging from 218K to 232K. The deviation between the raw and synthetic datasets’ characteristics further increases with decreasing privacy parameter ϵ and grows to an error of *multiple orders of magnitude*.

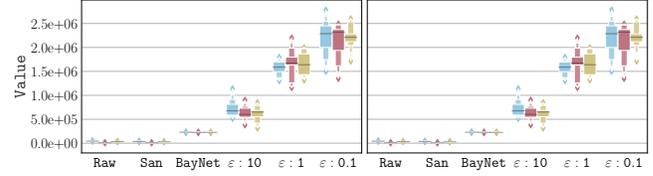


Figure 9: Mean (left) and median (right) for attributes TotalCharges, TotalChargesAccomm, and TotalChargesAncil from the Texas dataset.

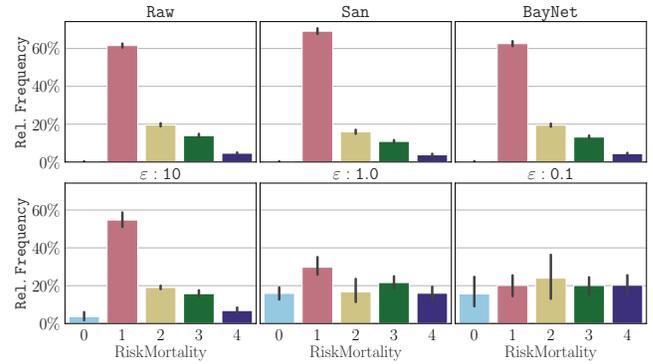


Figure 10: Marginal frequency counts for attribute RiskMortality from the Texas dataset.

We observe the same trend on the marginal frequency counts over the categorical attribute RiskMortality, another common reporting task (see Fig. 10). For instance, publishing a synthetic dataset sampled from a PrivBay-trained model in place of the raw data would lead an analyst to overestimate the relative frequency of category 0 by 3% percentage points when $\epsilon = 10$ and up to 15.3% percentage points for $\epsilon = 0.1$ with a growing variance of the mean error as ϵ decreases.

These results highlight the risks of synthetic data sharing outlined in Section 2. Generative models represent a *lower-dimensional approximation* of the raw data’s distribution and only capture a subset of the dataset’s high-dimensional feature space. Synthetic data sampled from the trained model hence does not preserve all of the raw data’s statistics and can lead to a large error on the derived insights. Synthetic data that accurately reflects the desired statistics would improve utility but not provide any privacy gain over directly publishing those aggregates.

No free lunch. In summary, the promising privacy gain of (differentially private) synthetic data publishing over traditional sanitisation shown in Fig. 4 comes at a significant cost in utility. Unsurprisingly, the higher a model’s privacy gain, the higher its loss in utility. Even models with low theoretical privacy guarantees ($\epsilon > 1$) may not provide the expected utility benefits and their utility loss may be prohibitive for many use cases.

The high utility loss of differentially private models can partially be explained by their reliance on metadata that needs to be derived independent of the model’s training set (see

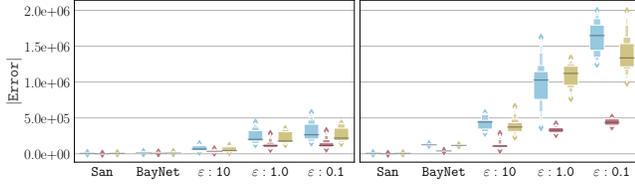


Figure 11: Absolute distance between the mean of attributes ■ TotalCharges, ■ TotalChargesAccomm, and ■ TotalChargesAncil in raw and synthetic datasets with metadata extracted from the raw data (*left*) and metadata derived independently from the training set (*right*)

Section 5). This can lead to a *significant difference* between the ranges in the raw data and the ones the data holder defines based on her background knowledge. To optimise the utility of synthetic datasets, the metadata given as input to the generative model training must reflect the raw data characteristics as closely as possible. However, this increases the privacy leakage of the model and, as we show in Section 5, in the most extreme case undermines its formal privacy guarantees.

We demonstrate this tradeoff in Fig. 11 and Fig. 12. Fig. 11 compares the absolute distance between the mean of three attributes in the raw data to the corresponding mean value in synthetic datasets sampled from models trained given different metadata as input. On the *left*, all models were given the exact ranges of attributes in the raw data sample. On the *right*, models used metadata learned from an independent population sample (see Section 5 for more details). In the case where the model directly learns all metadata from its raw input data (*left*), the utility loss of (differentially private) synthetic data is substantially reduced. For instance, the average distance between the mean of attribute ■ TotalChargesAccomm in the raw and synthetic data shrinks from 215K for a BayNet model trained with fixed ranges to 9K for a model based on metadata tailored to its training set. The error of differentially private models similarly decreases (from 621K to 34K for PrivBay $\epsilon : 10$) and with increasing ϵ values converges to its non-private version. Despite this improvement the utility loss of all differentially private models we tested still remains far above that of non-private synthetic data generation or traditional sanitisation. On the downside, as we show in Fig. 12 (*left* metadata tailored to training set, *right* metadata derived from an independent population sample), the utility gain achieved through better metadata leads to a large loss in privacy. When the model’s metadata is derived from the raw data, even models with low data utility provide close to no protection against linkage attacks ($PG \leq 0.35$ for three out of five targets tested across all ϵ values). Models that due to the use of metadata derived from an independent population sample result in a high privacy gain unfortunately do not provide sufficient data utility (see Fig. 11).

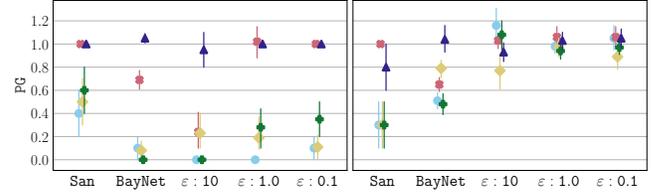


Figure 12: Per-record privacy gain for five outlier target records from the Texas dataset under an attack using the F_{Naive} feature set with metadata extracted from the raw data (*left*) and metadata derived independently from the training set (*right*)

6.3.2 Per-record utility loss

In the previous section, we show that for use cases that aim to derive aggregate insights from sensitive datasets generative models with a *high privacy gain* suffer from low data utility. However, average population metrics are not the main selling point of synthetic data. Synthetic data is often advertised as a solution for analysis tasks focused on more fine-grained statistical patterns that requires access to row-level data, in particular, the analysis of rare events and minority population subgroups. We hence now study whether synthetic data allows data holders to share datasets that *retain the signal of outlier records* without risking their privacy.

To study this relationship between privacy gain and utility loss, we slightly abuse the notion of advantage and formalise our utility metric as an advantage measure. This formalisation allows us to quantify the positive impact that the presence of a single target record in the training set has on a model’s performance for individual test records. We define a utility game, shown in Fig. 13, played between an analyst \mathcal{A} and a challenger \mathcal{C} . The analyst’s goal is to train a predictive model $h(\cdot)$ on a dataset X published by the challenger that performs well on a chosen test record \mathbf{r}_e from the population $\mathcal{R}_{\text{Test}}$. The analyst chooses this test record and a target record \mathbf{r}_t from the training population $\mathcal{R}_{\text{Train}}$. The analyst sends \mathbf{r}_t to the challenger, who follows the same procedure as in the linkability game and sends back the public bit b and the chosen dataset X . The analyst trains a model on dataset X and uses it to predict the test record’s label $\hat{y}_e \leftarrow h_X(\tilde{\mathbf{r}}_e)$. The analyst wins the game if $y_e = \hat{y}_e$.

Based on this game, we define the advantage that adding a record \mathbf{r}_t to the dataset R gives to record \mathbf{r}_e as

$$\text{Adv}^u(X, \mathbf{r}_e, \mathbf{r}_t) \triangleq \mathbb{P}[\hat{y}_e = y_e | s_t = 1] - \mathbb{P}[\hat{y}_e = y_e | s_t = 0] \quad (6)$$

where $\hat{y}_e = h_X(\tilde{\mathbf{r}}_e)$ is the label produced by a predictive model trained on data X for test record $\mathbf{r}_e = (\tilde{\mathbf{r}}_e, y_e)$ chosen by the analyst with feature set $\tilde{\mathbf{r}}_e$ and label y_e ; and s_t indicates the presence of target record \mathbf{r}_t in the raw dataset R .

To empirically evaluate the utility loss of synthetic data publishing, we use the same split of the Texas dataset as in the previous section. We ensured that the outlier targets

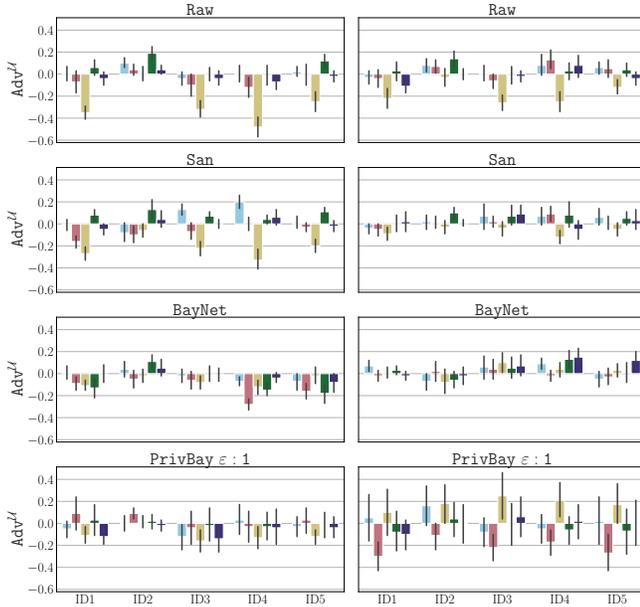


Figure 14: Utility advantage for 5 test (indicated on the x-axis) and 5 target (indicated by the colour of each bar) record pairs for predictions on attribute `RiskMortality` from the Texas dataset for $n = 1000$ (left) and $n = 5000$ (right)

the characteristics of the original data with high accuracy, and hence retains utility for the use cases it is advertised for, it simultaneously enables adversaries to extract sensitive information about individuals. A high gain in privacy through any of the anonymisation mechanisms we evaluated here can only be achieved if the published synthetic or sanitised version of the raw data does not carry through the signal of individual records and in effect suppresses their record.

Our findings not only show that synthetic data is subject to the same tradeoffs as previous anonymisation techniques, but also demonstrate that the privacy gain of synthetic data publishing is *highly unpredictable*. Because it is not possible to predict which data features a generative model will preserve, it is neither possible to anticipate the minimum gain in privacy from synthetic data publishing nor its utility loss. In comparison to deterministic sanitisation techniques, synthetic data does not allow data holders to provide transparency about what information will be omitted in the published dataset and what information will be retained.

We conclude that synthetic data does not provide a better tradeoff between privacy and utility than traditional row-level sanitisation, especially for data use cases that focus on the analysis of outlier signals, such as financial fraud or medical anomaly detection [40, 63]. Generative models with formal privacy guarantees reduce private information leakage with respect to anonymisation but do not preserve the fine-grained statistical patterns needed for outlier analysis. Our evaluation further shows that, even for use cases that focus on aggregate insights, synthetic datasets with a high privacy gain can suffer

from a significant utility loss and can lead to false conclusions. Even synthetic datasets that do preserve the desired statistics still present a noisier summary of the data than traditional privacy-preserving query release mechanisms [17] due to the additional uncertainty introduced by the output sampling process.

Acknowledgments. We would like to thank Jon Ullman, Aloni Cohen, Kobbi Nissim, and Salil Vadhan for their valuable feedback on earlier versions of this work, Laurent Girod for his support in open-sourcing our code, and the anonymous reviewers and our shepherd Takao Murakami whose input helped to further improve this paper. This work was partially funded by the Swiss National Science Foundation with grant 200021-188824.

References

- [1] J. M. Abowd and L. Vilhuber. How protective are synthetic data? In *PSD*, 2008.
- [2] C. Arnold and M. Neunhoffer. Really useful synthetic data – a framework to evaluate the quality of differentially private synthetic data. *arXiv:2004.07740*, 2020.
- [3] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, 2014.
- [4] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. In *NeurIPS*, 2019.
- [5] S. M. Bellovin, P. K. Dutta, and N. Reiter. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22, 2019.
- [6] V. Bindschaedler, R. Shokri, and C. A. Gunter. Plausible deniability for privacy-preserving data synthesis. *arXiv:1708.07975*, 2017.
- [7] N. Carlini, S. Deng, S. Garg, S. Jha, S. Mahloujifar, M. Mahmood, S. Song, A. Thakurta, and F. Tramèr. An attack on instahide: Is private learning possible with instance encoding? *arXiv:2011.05315*, 2020.
- [8] D. Chen, N. Yu, Y. Zhang, and M. Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against GANs. *arXiv:1909.03935*, 2019.
- [9] V. Cheng, V. M. Suriyakumar, N. Dullerud, S. Joshi, and M. Ghassemi. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *ACM FAccT*, 2021.
- [10] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun. Generating multi-label discrete patient records using generative adversarial networks. *arXiv:1703.06490*, 2017.
- [11] C. Culnane, B. I. P. Rubinstein, and V. Teague. Health data in an open world. *CoRR*, 2017.
- [12] DataResponsibly. DataSynthesizer. <https://github.com/DataResponsibly/DataSynthesizer>, 2017.

- [13] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 2013.
- [14] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 2015.
- [15] Department of Commerce, National Institute of Standards and Technology. Differential private synthetic data challenge. <https://www.challenge.gov/challenge/differential-privacy-synthetic-data-challenge/>, 2019. Accessed 2020-06-01.
- [16] J. Drechsler and J. P. Reiter. Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 2010.
- [17] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.
- [18] C. Dwork and A. et al. Roth. The algorithmic foundations of differential privacy. *FTTCS*, 2014.
- [19] M. Elliot, K. O'hara, C. Raab, C. M. O'Keefe, E. Mackey, C. Dibben, H. Gowans, K. Purdam, and K. McCullagh. Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, 2018.
- [20] European Commission. European data strategy. <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy>, 2020. Accessed 2020-06-01.
- [21] European Commission. White paper on artificial intelligence. <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence>, 2020. Accessed 2020-06-01.
- [22] S. Gambs, F. Ladouceur, A. Laurent, and A. Roy-Gaumond. Growing synthetic data through differentially-private vine copulas. In *PoPETS*, 2021.
- [23] Z. Ghahramani. An introduction to hidden markov models and bayesian networks. In *Hidden Markov models: applications in computer vision*. 2001.
- [24] I. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv:1701.00160*, 2016.
- [25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [26] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models. In *PoPETS*, 2019.
- [27] Hazy. Safe synthetic data: Privacy, utility and control. <https://hazy.com/images/videos/hazy-privacy-explainer.pdf>, 2020. Accessed 2020-11-12.
- [28] B. Hilprecht, M. Härterich, and D. Bernau. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. In *PoPETS*, 2019.
- [29] N. Homer, S. Szelling, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLOS Genetics*, 2008.
- [30] J. Hu, J. P. Reiter, and Q. Wang. Disclosure risk evaluation for fully synthetic categorical data. In *PSD*, 2014.
- [31] B. Jayaraman and D. Evans. Evaluating differentially private machine learning in practice. In *USENIX Security*, 2019.
- [32] J. Jordon, Jinsung Yoon, and Mihaela van der Schaar. PATEGAN: Generating synthetic data with Differential Privacy guarantees. In *ICLR*, 2019.
- [33] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *NeurIPS*, 2014.
- [34] R. Kohavi and B. Becker. UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/Adult>, 2013.
- [35] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [36] B. Kulynych, M. Yaghini, M. Cherubin, G. Veale, and C. Troncoso. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. In *PoPETS*, 2022.
- [37] K. Leino and M. Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. *arXiv:1906.11798*, 2019.
- [38] Y. Long, L. Wang, D. Bu, V. Bindschaedler, X. Wang, H. Tang, C. A. Gunter, and K. Chen. A pragmatic approach to membership inferences on machine learning models. In *IEEE EuroS&P*, 2020.
- [39] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian. L-diversity: privacy beyond k-anonymity. In *ICDE*, 2006.
- [40] MostlyAI. Power up machine learning for fraud detection and aml with synthetic data. <https://mostly.ai/synthetic-data-for-fraud-detection-and-aml/>. Accessed 2021-08-23.
- [41] S. Mukherjee, Y. Xu, A. Trivedi, N. Patowary, and J. L. Ferrer. privGAN: Protecting GANs from membership inference attacks at low cost to utility. In *PoPETS*, 2021.
- [42] A. Narayanan, E. Shi, and B.I. P. Rubinstein. Link prediction by de-anonymization: How we won the Kaggle social network challenge. In *IJCNN*, 2011.
- [43] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE S&P*, 2008.
- [44] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE S&P*, 2009.
- [45] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets: a decade later. <https://www.cs.princeton.edu/~arvindn/publications/de-anonymization-retrospective.pdf>, 2019.

- [46] NHS England. A&E Synthetic Data. <https://data.england.nhs.uk/dataset/a-e-synthetic-data>, 2020. Accessed 2020-11-12.
- [47] P. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, 2009.
- [48] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*, 2017.
- [49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JLMLR*, 12, 2011.
- [50] H. Ping, J. Stoyanovich, and B. Howe. DataSynthesizer: Privacy-preserving synthetic datasets. In *SSDBM*, 2017.
- [51] A. Pyrgelis, C. Troncoso, and E. De Cristofaro. Knock knock, who's there? Membership inference on aggregate location data. In *NDSS*, 2018.
- [52] J. P. Reiter and R. Mitra. Estimating risks of identification disclosure in partially synthetic data. *JPC*, 2009.
- [53] J. P. Reiter, Q. Wang, and B. Zhang. Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *JPC*, 2014.
- [54] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Comms*, 2019.
- [55] L. Rosenblatt, X. Liu, S. Pouyanfar, E. de Leon, A. Desai, and J. Allen. Differentially private synthetic data: Applied evaluations and enhancements. *arXiv:2011.05537*, 2020.
- [56] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Friarxivtz, and M. Backes. ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *NDSS*, 2019.
- [57] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE S&P*, 2017.
- [58] L. Sweeney. k-anonymity: A model for protecting privacy. *IJUFKS*, 2002.
- [59] L. Sweeney and J. S. Yoo. De-anonymizing south korean resident registration numbers shared in prescription data. *Technology Science*, 2015.
- [60] Texas Department of State Health Services, Austin, Texas. Texas Hospital Inpatient Discharge Public Use Data File 2013 Q1-Q4. <https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm>, 2013. Accessed 2020-06-01.
- [61] The Open Data Institute. Diagnosing the NHS: SynAE. <https://www.odileeds.org/events/synae/>. Accessed 2020-06-01.
- [62] O. Theraux. Anonymisation and synthetic data: towards trustworthy data. <https://theodi.org/article/anonymisation-and-synthetic-data-towards-trustworthy-data>, 2019. Accessed 2020-06-01.
- [63] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digital Medicine*, 2020.
- [64] UK Government - Defence Science and Technology Laboratory. Synthetic data - technical report. <https://www.gov.uk/government/publications/synthetic-data>. Accessed 2020-10-27.
- [65] US Federal Department of Commerce US Federal Office of Management & Budget and Science & Technology. Federal data strategy. <https://strategy.data.gov/>, 2020. Accessed 2020-06-01.
- [66] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachani. Modeling tabular data using conditional GAN. In *NeurIPS*, 2019.
- [67] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. Bennett. Assessing privacy and quality of synthetic health data. In *AIDR*, 2019.
- [68] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. Bennett. Privacy preserving synthetic health data. In *ESANN*, 2019.
- [69] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *CSF*, 2018.
- [70] J. Yoon. MLforHealthLab: PATEGAN. <https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/82d7f91d46db54d256ff4fc920d513499ddd2ab8/alg/pategan/>, 2019.
- [71] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.*, 2017.
- [72] Hazy. <https://hazy.com/>.
- [73] SDGym - Benchmarking framework for synthetic data generators. <https://sdv.dev/>.
- [74] Stalice. <https://statice.ai/>.
- [75] Synthetic data privacy evaluation framework. https://github.com/spring-epfl/synthetic_data_release.

8 Appendix

8.1 Framework implementation

We implemented the evaluation framework as a Python library [75]. The library has two main classes: `GenerativeModels` and `PrivacyAttacks`. For both classes we define a parent class that determines the core functionality that objects of the class need to implement.

`GenerativeModel` provides two main functions. `GM.fit()` is called with a raw dataset R as input and implements the model's training procedure. `GM.sample(m)` generates a synthetic dataset S of size m corresponding to $S \sim \mathcal{D}_{g(R)}^m$. The library enables easy integration of existing model training procedures. `GM.fit()` simply wraps any existing training algorithm and exposes the appropriate API endpoints.

`PrivacyAttack` objects have two functions: `PA.train` and `PA.attack`. `PA.train(r_t , BK)` trains the attack for

Table 1: Generative Model Hyperparameters

	IndHist	BayNet		PrivBay	
	nbins	nbins	degree	nbins	degree
Adult	45	45	1	45	1
Texas	25	25	1	25	1

	CTGAN			
	embeddings	gen_dim	dis_dim	l2scale
Adult	128	(256, 256)	(256, 256)	10^{-6}
Texas	128	(256, 256)	(256, 256)	10^{-6}

a specific target record r_t on background knowledge BK. $\text{PA.attack}(S)$, takes a dataset S and outputs a guess about a secret value. In our implementation, we instantiate `PrivacyAttack` with two attacks, a membership inference adversary and an attribute inference attack. The library also includes procedures to estimate the privacy gain of synthetic and sanitised data publishing.

Generative model and feature set parametrisation. We integrated five existing models into our Python library. Each of the models has a set of model hyper-parameters that can be adjusted to fit the input data. In Table 1 we list the parameter values for each model and dataset used in our experiments.

We implement the adversary’s feature sets as feature extraction objects `FeatureSet`. Each `FeatureSet` takes in a synthetic dataset S of size $m \times k$ and outputs a vector of size $l \times 1$. Our library includes the following feature sets:

F_{Naive} . The naive feature set computes the mean, median, and variance of each numerical attribute and encodes the number of distinct categories plus the most and least frequent category for each categorical attribute.

F_{Hist} . The histogram feature set computes the marginal distribution of each data attribute. Numerical attributes are binned with configurable bin size and frequency counts are computed for categorical attributes. The number of bins per attribute is configured for each dataset independently. In our experiments, we set the number of bins to 45 and 25 for the `Adult` and `Texas` dataset, respectively.

F_{Corr} . The correlations feature set encodes pairwise attribute correlations. Categorical attributes are dummy-encoded before computing the pairwise correlation matrix. The F_{Hist} and F_{Corr} feature sets include a pre-processing step in which continuous columns are binned. The number of bins is a configurable parameter that can be adjusted to fit the input data. In our experiments, we set the number of bins to 45 and 25 for the `Adult` and `Texas` dataset, respectively.

8.2 Datasets

We include two tabular datasets, commonly used in the machine learning (ML) literature, in our experimental evaluation. Tabular datasets are the most relevant data type in the syn-

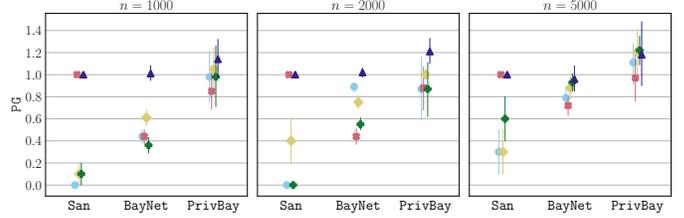


Figure 15: Per-record privacy gain for five outlier records for the Texas dataset under an attack using the F_{Naive} feature set.

thetic data publishing case. One datasets contains financial data the other one health data:

`Adult` [34]. The `Adult` dataset contains information from 45,222 individuals extracted from the 1994 US Census database. Each entry consists of 15 attributes among which 6 are continuous attributes and 9 are categorical attributes.

`Texas` [60]. The `Texas Hospital Discharge` dataset is a large public use data file provided by the Texas Department of State Health Services. The dataset we use consists of 50,000 records uniformly sampled from a pre-processed data file that contains patient records from the year 2013. We retain 18 data attributes of which 11 are categorical and 7 continuous.

8.3 Attribute inference

We formalise the risk of attribute inference as a prediction problem in which an attacker learns to predict the value of an unknown sensitive attribute from a set of known attributes given access to a raw, sanitised, or synthetic dataset which we denote as X . The adversary splits the dataset X into two parts: A feature matrix \tilde{X} that contains the values for all attributes known to the adversary and a vector \mathbf{x}_s with the corresponding sensitive attribute values.

If the attribute targeted by the attack is a continuous, real-valued attribute $r_s \in \mathbb{R}$, we model attribute inference as a linear regression problem. The linear regression attack models the relationship between the sensitive attribute values in \mathbf{x}_s and the attributes in \tilde{X} as a linear relationship with coefficients \mathbf{w}_X , and treats the records in X as i.i.d. samples:

$$\mathbf{x}_s = \tilde{X} \mathbf{w}_X + \epsilon, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad (7)$$

During training, the adversary takes the dataset X , splits it into a feature matrix \tilde{X} and target variable \mathbf{x}_s , and uses maximum likelihood estimation to obtain a set of regression coefficients $\mathbf{w}_X = \max_{\mathbf{w}} \mathbb{P}[\mathbf{x}_s | \tilde{X}, \mathbf{w}]$.

The simplicity of the model enables us to analytically derive the attacker’s posterior distribution over the target’s secret given access to dataset X $\mathbb{P}[\hat{r}_s | X, \tilde{r}_i] = \mathcal{N}[\tilde{r}_i \mathbf{w}_X, \hat{\sigma}_X^2]$, with variance $\hat{\sigma}_X^2 = \frac{1}{n-(k-1)} \sum_{i=1}^n (x_s^i - \tilde{\mathbf{x}}^i \mathbf{w}_X)^2$.

8.4 Privacy-utility tradeoff

Fig. 15 shows the privacy gain for five outlier targets from the `Texas` dataset for varying dataset sizes.