# TAMIS: Tailored Membership Inference Attacks on Synthetic Data

Paul Andrey[1] (✉), Batiste Le Bars[1], and Marc Tommasi[1]

Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
paul.andrey@inria.fr

**Abstract.** Membership Inference Attacks (MIA) enable to empirically assess the privacy of a machine learning algorithm. In this paper, we propose TAMIS, a novel MIA against differentially-private synthetic data generation methods that rely on graphical models. This attack builds upon MAMA-MIA, a recently-published state-of-the-art method. It lowers its computational cost and requires less attacker knowledge. Our attack is the product of a two-fold improvement. First, we recover the graphical model having generated a synthetic dataset by using solely that dataset, rather than shadow-modeling over an auxiliary one. This proves less costly and more performant. Second, we introduce a more mathematically-grounded attack score, that provides a natural threshold for binary predictions. In our experiments, TAMIS achieves better or similar performance as MAMA-MIA on replicas of the SNAKE challenge.

**Keywords:** Synthetic Data Generation · Differential Privacy · Membership Inference Attack · Graphical Models.

## 1 Introduction

Synthetic Data Generation (SDG) consists in producing artificial samples that match the specifications and retain distributional properties of actual data from a given domain. Over the past decade, it has received increased focus as a way to enable releasing data that can be used to learn statistics or even train machine learning models without granting access to actual personal data.

However, research has shown that synthetic data is not inherently private. Indeed, a synthetic dataset or a generative model learned from private records can leak private information [17]. To mitigate this risk, most state-of-the-art SDG methods provide differential privacy (DP) guarantees, that are formal properties of the method offering a provable upper bound on the residual privacy risk [5], usually at the cost of decreased utility. Most state-of-the-art SDG methods with DP guarantees rely either on learning a graphical model to approximate the structure of the data distribution [22, 14, 3], or on training a generative neural network, typically adversarially [19, 21].

Privacy risks can also be assessed empirically using privacy attacks, which can be complementary to DP guarantees [4, 6]. Membership Inference Attack (MIAs) are a type of privacy attack where an attacker having access to a trained machine learning model attempts to predict whether certain records were part of its training data [16]. MIAs have been transposed to SDG by a number of authors [17, 10, 9, 12, 2], who introduced a variety of attacks and threat models.

The state-of-the-art method for MIA against SDG methods relying on graphical models is MAMA-MIA [8]. It was developed to win the SNAKE challenge [1], where it achieved great success against the MST [14] and PrivBayes [22] SDG methods, especially in high-$\epsilon$ (that is, low-privacy) settings [7].

In this paper, we introduce TAMIS, a novel attack that achieves better or similar success as MAMA-MIA, has a lower computational cost and requires less attacker knowledge. We focus on attacking MST and PrivBayes, using principles that could be extended to other graphical-model-based methods. This new attack is the product of a two-fold improvement over MAMA-MIA. On the one hand, we propose to recover the structure of the graphical model having generated a synthetic dataset by using solely that dataset, rather than shadow-modeling. This proves less costly and more performant, especially against MST, for which we remove the need for the attacker to know any hyper-parameter used for generation. On the other hand, we introduce a more mathematically-grounded attack score, that achieves similar or better performance as the MAMA-MIA one in experiments over replicas of the SNAKE challenge. It notably achieves high accuracy without requiring the attacker to know the true proportion of training points in the attacked set.

In Section 2 we provide more detailed background on SDG and MIAs. Then, we define the setting for our contributions in Section 3, and come to present them in Section 4. Experiments are reported in Sections 5 and 6.

## 2 Background

### 2.1 Synthetic Data Generation

The aim of SDG is to produce a dataset of synthetic records $\mathcal{D}_{\text{synth}}$ that follow a similar distribution as observed private records in a training dataset $\mathcal{D}_{\text{train}}$. In the remainder of this paper, we note $\mathbb{P}_X$ the underlying distribution of $\mathcal{D}_{\text{train}}$.

Numerous SDG methods have been proposed in the literature. In this work, we focus on those based on graphical models as parametric estimators of $\mathbb{P}_X$, that approximated its dependency structure [14, 22, 3].

Among other methods, we can mention those using generative neural networks as non-parametric estimators of $\mathbb{P}_X$ [20, 19, 21]. While less interpretable [11], these methods cover a wider variety of data types, including time series [13] or multi-relational data [15].

Mechanisms have been introduced in SDG methods in order to provide DP guarantees on the training data. DP was introduced by Dwork & Roth [5], as a way to formalize and quantify the privacy of an algorithm with respect to its input data. Given $\epsilon > 0$ and $\delta \in ]0, 1[$, an algorithm $\mathcal{A} : \mathcal{D} \rightarrow O$ is deemed $(\epsilon, \delta)$-differentially private if, and only if, for any pair of adjacent datasets $\mathcal{D}, \mathcal{D}'$ (that is, datasets that differ by a single record) and for any $S \subseteq O$, $\mathbb{P}(\mathcal{A}(D) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(D') \in S) + \delta$.

An important property of DP is the post-processing theorem. For SDG, it means that the synthetic data inherit DP guarantees of their generative model.

### 2.2 Membership Inference Attacks

MIAs were first defined by Shokri et al. [16]. In a classical MIA, the attacker has access to a trained model, and attempts to predict whether certain records were part of its training dataset. In the context of SDG, the attacker instead has access to a synthetic dataset $\mathcal{D}_{\text{synth}}$, knows some actual records $\mathcal{D}_{\text{target}}$ and tries to assess which of these were part of $\mathcal{D}_{\text{train}}$ from which $\mathcal{D}_{\text{synth}}$ was derived.

A variety of MIAs on SDG have been proposed in the literature [17, 10, 9, 12, 2], that cover distinct threat models. These were notably reviewed by Houssiau et al. [12], that distinguish three main settings: in the white-box one, the attacker has full access to the trained generative model; in the black-box one, they have accurate knowledge of the SDG method; in the no-box one, they only have access to a given $\mathcal{D}_{\text{synth}}$. Finer-grained variants of the black-box setting exist, notably as to the whether the attacker knows the hyper-parameters of the SDG method.

Another common hypothesis is for the attacker to have access to an auxiliary dataset $\mathcal{D}_{\text{aux}}$ that follows the same underlying distribution as $\mathcal{D}_{\text{train}}$. With this, an attacker can notably perform shadow modeling, that is run the SDG method on controlled inputs, resulting in labeled replicas of the MIA setting [17].

**DOMIAS** [2] is a generic framework to conduct MIAs on synthetic data, applicable to either black-box or no-box settings but requiring access to auxiliary data. DOMIAS aims to identify $\mathcal{D}_{\text{train}}$ samples that have been over-fitted by the SDG, resulting in $\mathcal{D}_{\text{synth}}$ concentrating more density around these samples than a perfect estimate of $\mathbb{P}_X$ would have. Note that DP bounds and noises the contribution of samples, hence offering protection against over-fitting and MIAs.

Attack scores are defined as the ratio of density functions estimated on either $\mathcal{D}_{\text{synth}}$ or $\mathcal{D}_{\text{aux}}$, which we note $\hat{\mathbb{P}}_X^{\mathcal{D}_{\text{synth}}}$ and $\hat{\mathbb{P}}_X^{\mathcal{D}_{\text{aux}}}$:

$$\Lambda_{\text{DOMIAS}}(x) = \frac{\hat{\mathbb{P}}_X^{\mathcal{D}_{\text{synth}}}(x)}{\hat{\mathbb{P}}_X^{\mathcal{D}_{\text{aux}}}(x)} \tag{1}$$

This ratio is highest for samples that fit $\mathcal{D}_{\text{synth}}$ more than $\mathcal{D}_{\text{aux}}$, denoting possible over-fitting. Any density estimator may be plugged into this equation.

**MAMA-MIA** [8] is explicitly inspired by DOMIAS, but requires black-box knowledge of the SDG method and of all its hyper-parameters. It replaces density estimators with statistics on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{synth}}$ that are likely to have been selected by the SDG method, hence computed on $\mathcal{D}_{\text{train}}$ and perpetuated in $\mathcal{D}_{\text{synth}}$. In other words, attack scores are made to focus on distributional features of $\mathbb{P}_X$ that were explicitly modeled by the SDG method.

To identify statistics on which to focus, MAMA-MIA uses shadow modeling of the SDG method on $\mathcal{D}_{\text{aux}}$, meaning it replicates the SDG method up to the statistics selection step on random subsets of $\mathcal{D}_{\text{aux}}$ that match the size of $\mathcal{D}_{\text{train}}$.

## 3 Setting

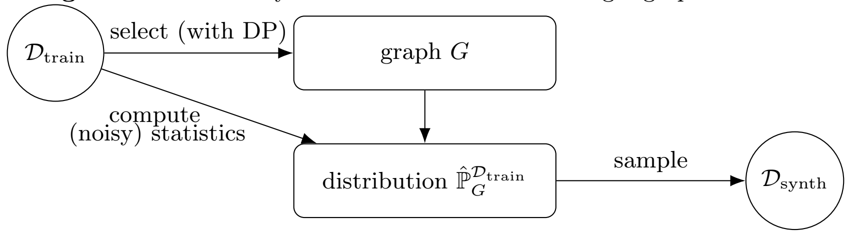
In this paper, we consider MIAs against SDG methods that rely on graphical models. Our threat model is to assume access to a synthetic dataset $\mathcal{D}_{\text{synth}}$ and to an auxiliary dataset $\mathcal{D}_{\text{aux}} \sim \mathbb{P}_X$, as well as black-box knowledge of the nature of the SDG method, and when specified of its hyper-parameters.

We consider data with categorical attributes, hence a multivariate random variable $X = (X_1, \ldots, X_d)$ where $\forall i \in \{1, \ldots, d\}, X_i \in \mathcal{X}_i := \{1, \ldots, n_i\}$ with $n_i \in \mathbb{N}^*$. We note individual records in lower case: $x = (x_1, \ldots, x_d)$. Support for continuous variables can be achieved using quantization in pre- and post-processing [22, 8], which we leave out without loss of generality.

A graphical model over $X$ is a family of probability distributions that can be represented as a graph $G = (V, E)$, with nodes $V = \{1, \ldots, d\}$ and edges $E$ that define the structure of conditional dependencies between attributes of $X$ [18]. Given $G$, a specific distribution is obtained by defining some statistics, based on which its joint density can be factorized.

The SDG methods we consider select a graphical model over $X$ that approximates the structure of $\mathbb{P}_X$, estimate associated statistics over $\mathcal{D}_{\text{train}}$, and generate $\mathcal{D}_{\text{synth}}$ as iid samples from the resulting distribution. To achieve DP, randomness is added to both the graph selection and statistics estimation steps. Figure 1 summarizes this generic approach.

**Fig. 1.** Flowchart of Synthetic Data Generation using a graphical model




### 3.1 MST

MST [14] is a SDG method that relies on a tree graphical model. A tree is a connected undirected graph with a constant number of edges $|E| = |V| - 1$. In such a graph, for any node $i$, we note $N(i) := \{j \in V | (i, j) \in E\}$ its neighbors.

The joint density of a tree graphical model is factorized based on the 1-way marginals over its nodes and 2-way marginals over its edges, as

$$\hat{\mathbb{P}}_G^{\mathcal{D}}(x) = \prod_{i \in V} \mu_i^{\mathcal{D}}(x)^{1-|N(i)|} \prod_{(i,j) \in E} \mu_{ij}^{\mathcal{D}}(x) \tag{2}$$

where $\mu_i^{\mathcal{D}}(x) = \hat{\mathbb{P}}_{\mathcal{D}}(X_i = x_i) := \frac{1}{|\mathcal{D}|} \sum_{\tilde{x} \in \mathcal{D}} \mathbb{1}\{\tilde{x}_i = x_i\}$
and $\mu_{ij}^{\mathcal{D}}(x) = \hat{\mathbb{P}}_{\mathcal{D}}(X_i = x_i, X_j = x_j) := \frac{1}{|\mathcal{D}|} \sum_{\tilde{x} \in \mathcal{D}} \mathbb{1}\{\tilde{x}_i = x_i, \tilde{x}_j = x_j\}$.

**Graph selection in MST** To select a tree graph from $\mathcal{D}_{\text{train}}$, MST uses a differentially-private maximum-spanning tree algorithm. It assigns a score $s_{ij} = \sum_{l_i,l_j \in \mathcal{X}_i \times \mathcal{X}_j} |\hat{\mathbb{P}}_{\mathcal{D}_{\text{train}}}(X_i = l_i, X_j = l_j) - \hat{\mathbb{P}}_{\mathcal{D}_{\text{train}}}(X_i = l_i)\hat{\mathbb{P}}_{\mathcal{D}_{\text{train}}}(X_j = l_j)|$ to each and every possible edge, with random noise added to 1-way marginals. Then, at each of $|V| - 1$ steps, an edge is randomly selected among valid candidates with probabilities proportional to $s_{ij}$ scores.

### 3.2 PrivBayes

PrivBayes [22] is a SDG method that relies on a bayesian network. A bayesian network is a directed acyclic graph. In such a graph, for any node $i$, we note $\Pi_i := \{j \in V | (j, i) \in E\}$ its parent set, that is the set of nodes with an edge towards $i$. We note $x_{\Pi_i}$ the vector of coordinates of $x$ in $\Pi_i$.

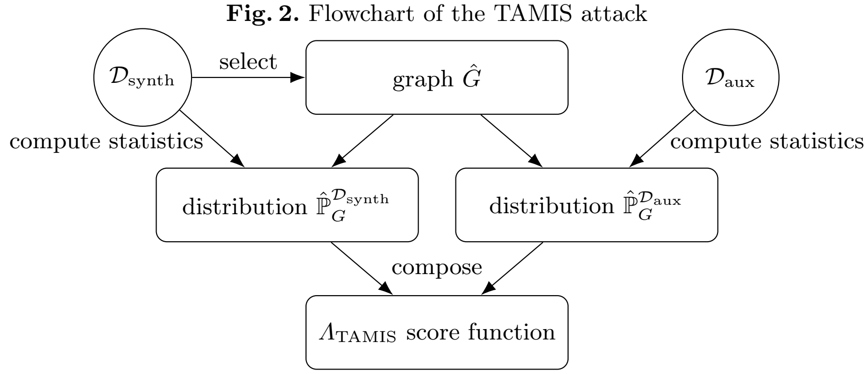
The joint density of a bayesian network is factorized based on the conditionals of nodes with respect to their parent set, as

$$\hat{\mathbb{P}}_G^{\mathcal{D}}(x) = \prod_{i \in V} \mu_{i,\Pi_i}^{\mathcal{D}}(x) \tag{3}$$

where $\mu_{i,\Pi_i}^{\mathcal{D}}(x) = \hat{\mathbb{P}}_{\mathcal{D}}(X_i = x_i | X_{\Pi_i} = x_{\Pi_i}) := \frac{1}{|\mathcal{D}|} \sum_{\tilde{x} \in \mathcal{D}} \mathbb{1}\{\tilde{x}_i = x_i, \tilde{x}_{\Pi_i} = x_{\Pi_i}\}$.

**Graph selection in PrivBayes** To select a bayesian network from $\mathcal{D}_{\text{train}}$, PrivBayes uses a differentially-private greedy algorithm. At each of $|V|$ steps, a (node, parent set) tuple is randomly selected among valid candidates with probabilities proportional to scores $s_{i,\Pi_i} = \frac{1}{2} \sum_{l_i,\pi_i \in \mathcal{X}_i \times \mathcal{X}_{\Pi_i}} |\hat{\mathbb{P}}_{\mathcal{D}_{\text{train}}}(X_i = l_i | \Pi_i = \pi_i) - \hat{\mathbb{P}}_{\mathcal{D}_{\text{train}}}(X_i = l_i)\hat{\mathbb{P}}_{\mathcal{D}_{\text{train}}}(\Pi_i = \pi_i)|$. Nodes and their parent set are constrained to have a total domain size below a threshold that is proportional to the privacy budget $\epsilon$ and to a hyper-parameter $\theta$, introducing a trade-off between structure-induced approximation errors and DP-induced estimation errors.

**Fig. 2.** Flowchart of the TAMIS attack



## 4 TAMIS: Tailored MIA on Synthetic data

In this paper, we introduce TAilored Membership Inference attacks on Synthetic data (TAMIS), that are MIAs against SDG methods that rely on graphical models. Our attack scores are based on the likelihood ratio approach of DOMIAS and MAMA-MIA. The novelty of our approach is to learn the structure of a graphical model matching the SDG method directly from $\mathcal{D}_{\text{synth}}$, and to use the factorized joint density of that model as a density estimator in attack scores. Figure 2 summarizes the TAMIS attack procedure as a flowchart.

By proposing an alternative to shadow modeling, we are able to lower the computational cost of the attack compared with MAMA-MIA. For MST, we are also able to remove the requirement for the attacker to know hyper-parameters. Furthermore, by recovering a graph rather than a set of weights that cover a broader set of edge choices, we are able to resort to a ratio of likelihoods as attack score, which is distinct in nature from the MAMA-MIA attack scores, as detailed below in this section. Table 1 summarizes the key differences between the two attacks.

**Table 1.** Differences in hypotheses, costs and nature of TAMIS and MAMA-MIA

| | | MAMA-MIA | | TAMIS | |
|---|---|---|---|---|---|
| | | MST | PrivBayes | MST | PrivBayes |
| Graph Recovery | H: Known SDG params | yes | yes | no | yes |
| | H: Known $\lvert\mathcal{D}_{\text{train}}\rvert$ | yes | yes | no | no |
| | H: Access to $\mathcal{D}_{\text{aux}}$ | yes | yes | no | no |
| | C: Cost relative to SDG | $K = 50$ | $K = 50$ | $\leq 1$ | 1 |
| | N: Nature of the output | weights $W$ | | graph $G$ | |
| Attack Scores | H: Access to $\mathcal{D}_{\text{aux}}$ | yes | yes | yes | yes |
| | N: Nature of the score | sum of statistics ratios | | ratio of densities | |

### 4.1 Graphical model recovery from the Synthetic Dataset

The first step of our attack consists in selecting a graphical model that matches that learned by the SDG method. Ideally, we would like to recover the exact structure that was used for generating $\mathcal{D}_{\text{synth}}$, in order to tailor attack scores to features of $\mathbb{P}_X$ that were actually (albeit noisily) measured on $\mathcal{D}_{\text{train}}$.

To do so, we introduce SDG-method-specific algorithms that take $\mathcal{D}_{\text{synth}}$ as input, as opposed to the shadow-modeling approach of MAMA-MIA that is run on $\mathcal{D}_{\text{aux}}$. This decreases computational costs and avoids requiring access to $\mathcal{D}_{\text{aux}}$. We also believe this approach to be more rational. Indeed, shadow modeling on $\mathcal{D}_{\text{aux}}$ is bound to provide relatively generic information about the likely structure of the generative model. On the opposite, that structure is bound to be reflected in $\mathcal{D}_{\text{synth}}$, hence easier to identify from it. This point is especially important in DP regimes that introduce a lot of randomness to the SDG graph selection step.

**Graph recovery for MST** To recover the tree underlying $\mathcal{D}_{\text{synth}}$ generated by MST, we introduce a modified version of the graph selection step from MST, that is deprived of DP mechanisms. It consists in measuring all 1- and 2-way marginals of $\mathcal{D}_{\text{synth}}$, assigning the same edge-wise score as the MST selection algorithm does but without noise, and finally using a maximum spanning tree algorithm to find the tree with the highest possible sum of edge scores.

This algorithm is deterministic, involves slightly less computations than a single shadow modeling run of the MST graph selection step, and does not require any attacker knowledge. As such, it could be applied to any $\mathcal{D}_{\text{synth}}$.

**Graph recovery for PrivBayes** To recover the bayesian network underlying $\mathcal{D}_{\text{synth}}$ generated by PrivBayes, we simply apply the model-selection step of PrivBayes on $\mathcal{D}_{\text{synth}}$. This amounts to a single shadow modeling run, which is less costly than the numerous ones run by MAMA-MIA (50 by default) but requires the same attacker knowledge of the PrivBayes hyper-parameters, due to the selection step adjusting the size of considered parent sets in the graph based on these. It is also non-deterministic due to the DP mechanisms.

### 4.2 Likelihood-based Attack Scores

The second step of our attack consists in defining attack scores. To do so, we use the graphical model resulting from the first step to estimate the respective joint densities of $\mathcal{D}_{\text{synth}}$ and $\mathcal{D}_{\text{aux}}$, which requires computing some statistics over both datasets. We then plug these densities into the DOMIAS framework, meaning we use their ratio as attack scores.

Our scores are thus obtained by plugging the joint density formulas of graphical models into the generic DOMIAS equation (1), that are equation (2) for MST and (3) for PrivBayes. The resulting formulas are:

$$\Lambda_{\text{TAMIS-MST}}(x;G) = \prod_{i \in V} \left( \frac{\mu_i^{\mathcal{D}_{\text{synth}}}(x)}{\mu_i^{\mathcal{D}_{\text{aux}}}(x)} \right)^{1-|N(i)|} \prod_{(i,j) \in E} \frac{\mu_{ij}^{\mathcal{D}_{\text{synth}}}(x)}{\mu_{ij}^{\mathcal{D}_{\text{aux}}}(x)} \tag{4}$$

$$\Lambda_{\text{TAMIS-PB}}(x;G) = \prod_{i \in V} \frac{\mu_{i,\Pi_i}^{\mathcal{D}_{\text{synth}}}(x)}{\mu_{i,\Pi_i}^{\mathcal{D}_{\text{aux}}}(x)} \tag{5}$$

TAMIS can therefore be thought of as an instantiation of the DOMIAS framework that uses a graphical model as density estimator, and picks that graphical model to mirror the one that was used in generating the synthetic data. This follows the same intuition as MAMA-MIA, but enacts it in a different way.

**Comparison with MAMA-MIA** In MAMA-MIA, the use of shadow modeling results in weights associated with possible edge (for MST) or (node, parent set) (for PrivBayes) choices, rather than in a valid graphical model. After $K$ shadow runs, these weights are defined as $\forall i \in \{1, \dots, d-1\}, \forall j \in \{i+1, \dots, d\}$,

$w_{ij} = \sum_{k=1}^{K} \mathbb{1}\{(i,j) \in E^{(k)}\}$ for MST, where $E^{(k)}$ is the set of edges selected in the $k$-th shadow run. For PrivBayes, they are defined as $\forall i \in V, \forall \Pi_i \subset V \setminus \{i\}$, $w_{i,\Pi_i} = \sum_{k=1}^{K} \mathbb{1}\{\forall j \in \Pi_i,\ (j,i) \in E^{(k)}\}$. We note $W$ a collection of such weights.

From there, the raw attack scores are defined as [1]

$$\Lambda_{\text{MAMAMIA-MST}}(x; W) = \frac{1}{\sum_{w \in W} w} \sum_{w_{ij} \in W} w_{ij} \frac{\mu_{ij}^{\mathcal{D}_{\text{synth}}}(x)}{\mu_{ij}^{\mathcal{D}_{\text{aux}}}(x)} \tag{6}$$

$$\Lambda_{\text{MAMAMIA-PB}}(x; W) = \frac{1}{\sum_{w \in W} w} \sum_{w_{i,\Pi_i} \in W} w_{i,\Pi_i} \frac{\mu_{i,\Pi_i}^{\mathcal{D}_{\text{synth}}}(x)}{\mu_{i,\Pi_i}^{\mathcal{D}_{\text{aux}}}(x)} \tag{7}$$

We remark that these scores correspond to a weighted average of DOMIAS-like scores attached to specific terms characterizing the density, whereas TAMIS uses a DOMIAS-like score over the entire joint density.

We also note that when attacking MST, MAMA-MIA leaves apart information from 1-way marginals, without justification. This omission is probably due to the fact that 1-way marginals are always part of the graph, regardless of selected edges.

**Hybrid scores** We introduce hybrids of our graph-recovery approach with the formulas of MAMA-MIA scores. This is useful to make more visible the difference between our scores and the MAMA-MIA ones, and to disambiguate experimentally the impact of the two folds of our contribution. These scores come from replacing weights resulting from shadow modeling with uniform weights that reflect the graph structure $G$ learned from $\mathcal{D}_{\text{synth}}$ in Equations (6) and (7).

$$\Lambda_{\text{Hybrid-MST}}(x; G) = \frac{1}{|E|} \sum_{(i,j) \in E} \frac{\mu_{ij}^{\mathcal{D}_{\text{synth}}}(x)}{\mu_{ij}^{\mathcal{D}_{\text{aux}}}(x)} \tag{8}$$

$$\Lambda_{\text{Hybrid-PB}}(x; G) = \frac{1}{|V|} \sum_{i \in V} \frac{\mu_{i,\Pi_i}^{\mathcal{D}_{\text{synth}}}(x)}{\mu_{i,\Pi_i}^{\mathcal{D}_{\text{aux}}}(x)} \tag{9}$$

## 5 Experiments

To assess the soundness and performance of our attacks, we conduct experiments that replicate the SNAKE challenge [1].

We make our source code, data and random seeds available, enabling to fully reproduce our experiments and results with minimal effort. They may be found online[2], together with dedicated documentation on implementation details.

---

[1] We note that the division of scores by $\sum_{w \in W} w$, amenable to normalizing weights, is an addition to the original MAMA-MIA formulas. We adopted it to avoid $K$ impacting the magnitude of raw scores. This improved predictions in our experiments.

[2] https://gitlab.inria.fr/magnet/thesepaulandrey/tamis

Due to paper length constraints, this version only details main experimental results. Throughout this section, we make multiple references to appendices that are part of a fuller version of the paper, a pre-print of which is available on arXiv[3].

### 5.1 Dataset

We used the publicly-available base dataset from the SNAKE challenge, that was derived from US socio-demographic data published by the Economic Policy Institute. It consists of about 201k samples with 15 variables, 12 of which are categorical (with 2 to 51 possible labels each) and 3 of which have integer values but can be treated as categorical nonetheless (age, number of children and mean weekly number of worked hours). Additionally, individual samples each belong to a given household, each of which groups 1 to 10 individuals.

We applied the same kind of preparation as was done in SNAKE. We take the base dataset as $\mathcal{D}_{\text{aux}}$, that is made available to the attacker. We randomly pick 100 households out of the 812 ones that group at least 5 individuals, and designate their records as composing $\mathcal{D}_{\text{target}}$. Then, all individuals within half of $\mathcal{D}_{\text{target}}$ households are made part of $\mathcal{D}_{\text{train}}$, and $\mathcal{D}_{\text{train}}$ is completed with random individuals from $\mathcal{D}_{\text{aux}} \setminus \mathcal{D}_{\text{target}}$ until $|\mathcal{D}_{\text{train}}| = 10000$.

We generated 50 distinct replicas of the SNAKE $(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{target}})$ generation. We then ran both MST and PrivBayes on each and every replica with various privacy budgets. For each setting, we generated $|\mathcal{D}_{\text{synth}}| = |\mathcal{D}_{\text{train}}| = 10000$ synthetic samples, and recorded the structure of the graphical model that generated them. We considered $\epsilon \in [0.1, 1, 10, 100, 1000]$. For MST, we always set $\delta = 10^{-9}$. For PrivBayes, the DP mechanisms always achieve $\delta = 0$. For each value of $\epsilon$, we also ran shadow modeling of the graphical model selection step of MST and PrivBayes on 50 random subsets of $\mathcal{D}_{\text{aux}}$, as was done in the MAMA-MIA paper.

**MIA evaluation settings** We evaluate attacks on three distinct settings. The first is $\mathcal{D}_{\text{aux}}^{\text{ind}}$, where we attack each and every $\mathcal{D}_{\text{aux}}$ sample, to assess the average-case MIA risk.

The second is $\mathcal{D}_{\text{target}}^{\text{house}}$, where we conduct set membership inference on households of $\mathcal{D}_{\text{target}}$. This task is balanced by construction and matches original evaluation setting of SNAKE. The raw score for a household is taken to be the average of raw scores for samples in that household. These households are expected to be easier to attack than any $\mathcal{D}_{\text{aux}}$ sample. Indeed, there is some correlation among samples in a household, hence when they are jointly included in $\mathcal{D}_{\text{train}}$ they add weight to a given density region, which can cause the SDG to over-fit that region.

The third is $\mathcal{D}_{\text{target}}^{\text{ind}}$, where we attack $\mathcal{D}_{\text{target}}$ samples. These are expected to be somewhat easier targets than average samples for the reasons exposed before.

[3] https://arxiv.org/abs/2504.00758

### 5.2 Attack methods

**Against MST** we compared three main attacks. First, our TAMIS-MST attack (4), using a tree graph learned from $\mathcal{D}_{\text{synth}}$. Second, the MAMA-MIA attack targeted at MST (6), using weights resulting from shadow modeling over $\mathcal{D}_{\text{aux}}$. Finally, the Hybrid-MST attack (8), that uses MAMA-MIA-like scores over the same tree graph as TAMIS-MST. Some additional baselines were considered to complement our comparison of scores variants, that are reported in Appendix B.

**Against PrivBayes** we compared three main attacks. First, our TAMIS-PB attack (5), using a bayesian network learned from $\mathcal{D}_{\text{synth}}$. Second, the MAMA-MIA attack targeted at PrivBayes (7), using weights resulting from shadow modeling over $\mathcal{D}_{\text{aux}}$. Finally, the Hybrid-PB attack (9), that uses MAMA-MIA-like scores over the same bayesian network as TAMIS-PB.

In addition, we introduced TAMIS-PB* and Hybrid-PB*, that use the same attack scores as TAMIS-PB and Hybrid-PB respectively, but benefit from access to the bayesian network that was truly selected by PrivBayes to generate $\mathcal{D}_{\text{synth}}$. This falls outside of our threat model, but is useful to assess the extent to which recovering that graph can improve the success of MIAs.

### 5.3 Evaluation

**Graph Recovery** We systematically compare the graphical model selected by the SDG methods to generate $\mathcal{D}_{\text{synth}}$ with that inferred from $\mathcal{D}_{\text{synth}}$ as part of our attack. We compute the accuracy of choices between the true and estimated graphs, comparing edges for MST and (node, parent set) tuples for PrivBayes.

We also compare the edges or (node, parent set) tuples that are selected at least once across shadow modeling runs, hence included in MAMA-MIA scores, with true graphs. We compute precision, recall and jaccard index over both sets and report their mean, standard deviation and median across replicas.

**Membership Inference Metrics** To assess the success of MIAs, we report two metrics: the Area Under the Receiver-Operator Curve (AUROC) of attack scores, and the balanced accuracy of the resulting binary membership predictions. The AUROC is invariant to the activation of raw attack scores, and enables comparison with experimental results from the MAMA-MIA paper [8]. The balanced accuracy is defined as $0.5 * (\frac{\mathbb{P}(\hat{Y}=1|Y=1)}{\mathbb{P}(Y=1)} + \frac{\mathbb{P}(\hat{Y}=0|Y=0)}{\mathbb{P}(Y=0)})$, where $Y = \mathbb{1}\{x \in \mathcal{D}_{\text{train}}\}$ is the true membership label, and $\hat{Y}$ is a binary prediction resulting from a raw attack score. Mapping a raw attack score $\Lambda(x)$ into such a binary decision requires a monotonous activation function $f$ that outputs predicted probabilities in $[0, 1]$, and a threshold $t$ so that $\hat{Y} = \mathbb{1}\{f(\Lambda(x)) \geq t\}$.

In practice, we consider two distinct activation and thresholding regimes. In the *simple* activation regime, we use a default threshold $t = 0.5$, together with a classic activation function that is applied independently to each score. We choose the sigmoid function, corrected to account for raw scores being in $\mathbb{R}^+$: $f(x) = 2 * \text{Sigmoid}(x) - 1$, where $\text{Sigmoid}(x) = (1 + e^{-x})^{-1}$. In the *calibrated* activation regime, we adopt the approach introduced with MAMA-MIA, that was used to win the SNAKE challenge [7]. We make the additional strong hypothesis that the attacker knows $\mathbb{P}(Y = 1)$ for the samples under attack, and adjusts the activation of these targets so that $\mathbb{P}(\hat{Y} = 1) = \mathbb{P}(Y = 1)$. First, raw attack scores are standardized into z-scores (that is, centered around their mean then divided by their standard deviation), to avoid numerical under- and over-flow issues. Then, z-scores are centered around their $1 - \mathbb{P}(Y = 1)$ quantile, so that only a fraction $\mathbb{P}(Y = 1)$ are above 0. Finally, these scores are passed through the sigmoid function, and mapped into decisions using $t = 0.5$.

## 6 Results

### 6.1 Attacks against MST

**Graph recovery** Our method accurately recovered the generating tree in all settings, achieving a perfect match in edges selection for all replicas and $\epsilon$ value. In comparison, as detailed in Appendix A, shadow modeling missed a rarely-selected edge, and selected an increasing number of edges as $\epsilon$ lowered due to DP, thus resulting in more un-tailored terms in MAMA-MIA attack scores.

**Membership Inference** Metrics of MST-targeted attacks on $\mathcal{D}^{\text{ind}}_{\text{aux}}$ and $\mathcal{D}^{\text{house}}_{\text{target}}$ are reported in Table 2, excluding high-privacy regimes $\epsilon \in [0.1, 1]$ for which no attack achieves significative success due to the strong DP guarantees[4]. Reported values are the average and standard deviation of metrics across our 50 replicas. We highlight the highest mean value per setting between attacks in bold. Exhaustive results are placed in Appendix B.

Results show that both of our contributions improve attack success. First, Hybrid-MST achieves better results than MAMAMIA-MST, meaning that replacing shadow modeling weights with the tree graph recovered from $\mathcal{D}_{\text{synth}}$ improves MAMA-MIA-like attack scores. Next, TAMIS-MST achieves even better success in nearly all settings, meaning that the ratio of graphical model densities constitutes a better attack score than the average of 2-way marginals ratios.

We remark that against $\mathcal{D}^{\text{ind}}_{\text{aux}}$, TAMIS-MST and Hybrid-MST success metrics have a markedly lower standard deviation than MAMAMIA-MST. We hypothesize that this is due to shadow modeling weights causing MAMAMIA-MST scores to include terms that variably match the actual generative graph of $\mathcal{D}_{\text{synth}}$, and to exclude information on a relevant edge for some replicas. This validates the rationale of both TAMIS and MAMA-MIA to focus on aspects of the distribution that were actively modeled during SDG.

Comparing balanced accuracy across activation regimes, we observe on $\mathcal{D}^{\text{house}}_{\text{target}}$ that TAMIS-MST, which is more accurate than others in both settings, receives

[4] The fact that attacks fail for lower $\epsilon$ values is expected given the theoretical guarantees. It is worth noting that these privacy guarantees come at a high utlty cost.

**Table 2.** Main attack results against MST

| | | AUROC | | |
|---|---|---|---|---|
| | | $\epsilon$=1000 | $\epsilon$=100 | $\epsilon$=10 |
| $\mathcal{D}_{\text{aux}}^{\text{ind}}$ | TAMIS-MST | **66.25** (± 0.4) | **65.53** (± 0.4) | 59.25 (± 0.4) |
| | MAMAMIA-MST | 64.60 (± 1.0) | 64.13 (± 1.0) | 59.05 (± 0.6) |
| | Hybrid-MST | 65.46 (± 0.3) | 64.97 (± 0.3) | **59.59** (± 0.3) |
| $\mathcal{D}_{\text{target}}^{\text{house}}$ | TAMIS-MST | **77.76** (± 4.8) | **77.44** (± 3.6) | **69.87** (± 5.0) |
| | MAMAMIA-MST | 74.74 (± 5.4) | 74.62 (± 4.5) | 68.51 (± 5.4) |
| | Hybrid-MST | 75.97 (± 4.6) | 75.57 (± 4.3) | 69.02 (± 5.5) |
| | | Balanced Accuracy (Simple) | | |
| | | $\epsilon$=1000 | $\epsilon$=100 | $\epsilon$=10 |
| $\mathcal{D}_{\text{aux}}^{\text{ind}}$ | TAMIS-MST | **60.84** (± 0.3) | **60.39** (± 0.3) | **55.66** (± 0.3) |
| | MAMAMIA-MST | 56.66 (± 0.7) | 56.44 (± 0.7) | 54.40 (± 0.4) |
| | Hybrid-MST | 57.27 (± 0.4) | 57.06 (± 0.3) | 54.74 (± 0.3) |
| $\mathcal{D}_{\text{target}}^{\text{house}}$ | TAMIS-MST | **69.86** (± 4.7) | **69.52** (± 4.0) | **64.26** (± 4.6) |
| | MAMAMIA-MST | 61.06 (± 4.2) | 60.66 (± 3.4) | 58.16 (± 3.3) |
| | Hybrid-MST | 61.76 (± 3.6) | 61.88 (± 3.7) | 58.08 (± 3.3) |
| | | Balanced Accuracy (Calibrated) | | |
| | | $\epsilon$=1000 | $\epsilon$=100 | $\epsilon$=10 |
| $\mathcal{D}_{\text{aux}}^{\text{ind}}$ | TAMIS-MST | **55.02** (± 0.2) | **54.75** (± 0.2) | 52.63 (± 0.2) |
| | MAMAMIA-MST | 54.38 (± 0.6) | 54.15 (± 0.6) | 52.50 (± 0.4) |
| | Hybrid-MST | 54.76 (± 0.2) | 54.54 (± 0.2) | **52.69** (± 0.1) |
| $\mathcal{D}_{\text{target}}^{\text{house}}$ | TAMIS-MST | **70.24** (± 5.2) | **69.88** (± 3.4) | **64.26** (± 4.7) |
| | MAMAMIA-MST | 67.74 (± 4.8) | 67.24 (± 4.8) | 63.32 (± 5.1) |
| | Hybrid-MST | 68.32 (± 4.8) | 68.06 (± 4.2) | 63.56 (± 5.4) |

less improvement from calibration. Hence TAMIS-MST attack scores appear to be naturally suitable for sigmoid activation with a basic threshold. Oppositely, MAMA-MIA-like scores appear to rely on calibration, hence on an additional piece of attacker knowledge, to be turned into accurate predictions against $\mathcal{D}_{\text{target}}^{\text{house}}$. We also observe that calibration on the unbalanced $\mathcal{D}_{\text{aux}}^{\text{ind}}$ results in a decrease in balanced accuracy for all attack scores. This hints that the calibration proposed by MAMA-MIA authors may be over-fitted to the balanced $\mathcal{D}_{\text{target}}^{\text{house}}$ setting, that was the target of the SNAKE competition.

### 6.2 Attacks against PrivBayes

**Graph recovery** Our method almost never perfectly recovered the generating bayesian network from $\mathcal{D}_{\text{synth}}$. As detailed in Appendix A, for higher $\epsilon$ values, about half (node, parentset) choices match, while mismatches arise from marginal differences in including this or that edge, which can have a strong impact on the modeled density. In comparison, shadow modeling achieves above 90 % recall, but less than 25 % precision for $\epsilon \geq 1$, meaning MAMA-MIA attack scores contain most conditionals attached to the generative density of $\mathcal{D}_{\text{synth}}$,

**Table 3.** Main attack results against PrivBayes

| | | AUROC | | |
|---|---|---|---|---|
| | | $\epsilon$=1000 | $\epsilon$=100 | $\epsilon$=10 |
| $\mathcal{D}_{\text{aux}}^{\text{ind}}$ | TAMIS-PB | 64.65 (± 2.1) | 62.65 (± 1.3) | 53.99 (± 0.3) |
| | MAMAMIA-PB | **79.34** (± 1.5) | 64.36 (± 1.1) | **54.47** (± 0.3) |
| | Hybrid-PB | 79.33 (± 2.7) | **64.48** (± 1.7) | 54.42 (± 0.4) |
| | TAMIS-PB* | 66.72 (± 0.8) | 64.61 (± 0.4) | 54.16 (± 0.3) |
| | Hybrid-PB* | **83.01** (± 0.9) | **67.06** (± 0.3) | **54.72** (± 0.2) |
| $\mathcal{D}_{\text{target}}^{\text{house}}$ | TAMIS-PB | 82.74 (± 5.6) | 76.85 (± 5.2) | **62.00** (± 5.4) |
| | MAMAMIA-PB | **92.16** (± 2.6) | 78.06 (± 5.5) | 61.53 (± 4.7) |
| | Hybrid-PB | 91.87 (± 3.0) | **78.37** (± 5.6) | 61.49 (± 5.5) |
| | TAMIS-PB* | 85.75 (± 4.2) | 80.47 (± 4.8) | **62.64** (± 4.7) |
| | Hybrid-PB* | **94.32** (± 2.6) | **81.98** (± 4.8) | 62.23 (± 4.9) |
| | | Balanced Accuracy (Simple) | | |
| | | $\epsilon$=1000 | $\epsilon$=100 | $\epsilon$=10 |
| $\mathcal{D}_{\text{aux}}^{\text{ind}}$ | TAMIS-PB | 61.67 (± 1.8) | **58.53** (± 1.1) | **52.48** (± 0.3) |
| | MAMAMIA-PB | 69.97 (± 1.4) | 55.49 (± 0.6) | 50.81 (± 0.1) |
| | Hybrid-PB | **70.75** (± 2.4) | 56.21 (± 0.9) | 50.85 (± 0.1) |
| | TAMIS-PB* | 63.38 (± 0.7) | **60.18** (± 0.3) | **52.68** (± 0.2) |
| | Hybrid-PB* | **74.24** (± 1.0) | 57.67 (± 0.4) | 50.90 (± 0.1) |
| $\mathcal{D}_{\text{target}}^{\text{house}}$ | TAMIS-PB | 76.04 (± 5.5) | **70.20** (± 4.8) | **57.04** (± 4.2) |
| | MAMAMIA-PB | 78.54 (± 4.2) | 57.34 (± 3.7) | 51.74 (± 2.5) |
| | Hybrid-PB | **82.14** (± 4.9) | 59.86 (± 3.6) | 51.70 (± 2.4) |
| | TAMIS-PB* | 79.48 (± 4.8) | **73.12** (± 4.5) | **57.64** (± 3.5) |
| | Hybrid-PB* | **86.58** (± 4.0) | 62.36 (± 3.4) | 51.54 (± 2.3) |

but at least 3 times more terms made of other conditionals. Interestingly, graph recovery is easier when $\epsilon$ lowers, due to PrivBayes reducing possible choices.

**Membership Inference** Metrics of PrivBayes-targeted attacks on $\mathcal{D}_{\text{aux}}^{\text{ind}}$ and $\mathcal{D}_{\text{target}}^{\text{house}}$ are reported in Table 3, excluding high-privacy regimes $\epsilon \in [0.1, 1]$ for which no attack achieves significative success due to the strong DP guarantees. We also excluded balanced accuracy under the calibrated activation regime, which is of lesser interest as rapidly discussed below. Exhaustive results are placed in Appendix C. Reported values are the average and standard deviation of metrics across our 50 replicas. For each setting, we highlight two values in bold, that are the highest mean value either between TAMIS-PB, MAMAMIA-PB and Hybrid-PB or between TAMIS-PB* and Hybrid-PB*.

We first compare attacks that are in line with our threat model, namely TAMIS-PB, Hybrid-PB and MAMAMIA-PB. Results are somewhat ambivalent. On the one hand, when considering the balanced accuracy with simple activation on either $\mathcal{D}_{\text{aux}}^{\text{ind}}$ or $\mathcal{D}_{\text{target}}^{\text{house}}$, TAMIS-PB achieves the best results of all methods, save for the lowest-privacy regime $\epsilon = 1000$, where it is the worst. On the other hand, when considering the AUROC, TAMIS-PB is most often the worst method. On $\mathcal{D}_{\text{aux}}^{\text{ind}}$, MAMAMIA-PB achieves the best results, followed by Hybrid-PB that has similar average values but higher variance. On $\mathcal{D}_{\text{target}}^{\text{house}}$, Hybrid-PB and MAMAMIA-PB have similarly good results, while TAMIS-PB is worse or equal, with lower differences towards the other methods than on $\mathcal{D}_{\text{aux}}^{\text{ind}}$.

Next, we compare the TAMIS-PB* and Hybrid-PB* attacks, that are granted knowledge of the true bayesian network. We retrieve the same ambivalence as to which attack score is best: Hybrid-PB* has a higher or similar AUROC as TAMIS-PB*, while it has a markedly lower balanced accuracy using simple activation, save for the $\epsilon = 1000$ where it is markedly better. We also observe that TAMIS-PB* and Hybrid-PB* achieve better performance than the other three attacks. This again validates the hypothesis that focusing on distributional features selected by the SDG is key in crafting efficient attack scores. While is is unlikely that an attacker would be directly provided knowledge of the bayesian network structure that generated $\mathcal{D}_{\text{synth}}$, our results highlight that improving over our proposed graph recovery method could enable out-performing both the MAMA-MIA and current TAMIS-PB attacks.

Regarding calibrated activation, we renew the observations made for MST. On the one hand, this activation is detrimental to the accuracy of predictions on $\mathcal{D}_{\text{aux}}^{\text{ind}}$ for all methods. On the other hand, on the specific $\mathcal{D}_{\text{target}}^{\text{house}}$ setting, it is markedly beneficial to the accuracy of predictions for MAMA-MIA-like scores (MAMAMIA-PB and Hybrid-PB) while having very limited impact on TAMIS-PB, that still achieves better or similar balanced accuracy in most cases.

Factoring all previous results, we conclude that while MAMA-MIA-like scores seemingly extract more information than TAMIS-PB, as showed by their higher AUROC, the TAMIS-PB scores are more suitable for prediction in the realistic setting where the attacker has no additional knowledge to craft decision thresholds beyond the default $t = 0.5$.

We hypothesize that TAMIS scores underperforming for $\epsilon = 1000$ may be due to numerical effects of extreme values in ratios of conditionals. Notably, when attacking a sample that exhibits a combination of attributes unseen in either $\mathcal{D}_{\text{aux}}$ or $\mathcal{D}_{\text{synth}}$, both the MAMA-MIA and TAMIS implementations assign an arbitrarily-low conditional probability to it, which is bound to result in extreme ratio values. This is more likely to appear with high $\epsilon$, as parent sets are authorized to have a larger domain size, possibly containing very-rare combinations. This could probably be addressed by refining the way these cases are handled.

### 6.3 Cross-targeted attacks

In order to further assess how beneficial it is for attacks to be tailored to the SDG method, we experimented running MST-targeted attacks against PrivBayes-generated synthetic data, and conversely. Results are provided in Appendix D.

Overall, cross-targeted attacks under-perform compared with their counterparts. Notably, PrivBayes-targeted attacks have poor results against MST-generated data. This highlights that using a more complex density estimator does not necessarily result in more performant attacks, again validating the value of using a density estimator matching the generating one.

However, TAMIS-MST achieves good results against PrivBayes, with a higher or similar balanced accuracy with simple activation than MAMAMIA-PB and Hybrid-PB for all settings, save for $\epsilon = 1000$. This is remarkable, as TAMIS-MST requires less attacker knowledge and computational power than MAMA-MIA.

## 7 Conclusion

In this paper, we have investigated and validated the assumption that tailoring MIAs on synthetic data to the SDG method was beneficial to attack performance, and proposed alternatives to refine both steps of the state-of-the-art MAMA-MIA attack. The resulting TAMIS attacks were demonstrated to further improve the state-of-the-art against MST and PrivBayes on replicas of the SNAKE challenge, that MAMA-MIA recently won.

Our experiments have shown that recovering the graphical model underlying a synthetic dataset resulted in more successful attacks than gathering more generic information via shadow modeling. For MST, we were able to propose a straightforward graph recovery method that achieved perfect accuracy in our experiments, and requires both less computational power and less attacker knowledge than shadow modeling. For PrivBayes, our method only improves over shadow modeling in terms of computational power, and should be a focus for improvement in future work.

Our experiments have also shown that our proposed attack scores, which are more mathematically-grounded than their MAMA-MIA counterpart, produce more accurate predictions when using a simple sigmoid activation function and a default decision threshold. We have also shown that methods are not ranked similarly depending on whether their AUROC or balanced accuracy is considered. This highlights that there may be a gap between the information contained in attack scores and that which can instrumented into actual predictions by an attacker. Future research may help close that gap, either by designing clever activation and thresholding mechanisms that do not rely on unrealistic attacker knowledge hypotheses, or by further refining the way how focused statistics are combined into attack scores that both achieve high AUROC and behave nicely with simple activation functions.

Finally, while the MAMA-MIA and TAMIS attacks have been designed for the black-box model threat where the attacker knows which SDG method was used, we have remarked that the TAMIS-MST attack could in fact be run agnostic to the SDG method. Our experiments on PrivBayes-generated data have shown that it may be a competitive if not optimal method, and a future research direction could be to assess the value of that attack in a no-box threat model, using more diverse datasets, SDG methods and relevant baseline attacks.

**Acknowledgments.** This work was supported by the ANR 22-CMAS-0009 CAPS'UL (CAmpus Participatif en Santé numérique du Site Universitaire de Lille) project of the France 2030 AMI-CMA, and the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Allard, T., Béziaud, L., Gambs, S.: Snake challenge: Sanitization algorithms under attack. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. p. 5010–5014. CIKM '23, Association for Computing Machinery (2023). https://doi.org/10.1145/3583780.3614754
2. van Breugel, B., Sun, H., Qian, Z., van der Schaar, M.: Membership inference attacks against synthetic data through overfitting detection. In: Ruiz, F., Dy, J., van de Meent, J.W. (eds.) Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 206, pp. 3493–3514. PMLR (25–27 Apr 2023), https://proceedings.mlr.press/v206/breugel23a.html
3. Cai, K., Lei, X., Wei, J., Xiao, X.: Data synthesis via differentially private markov random fields. Proc. VLDB Endow. **14**(11), 2190–2202 (2021). https://doi.org/10.14778/3476249.3476272
4. Desfontaines, D.: Empirical privacy metrics: The bad, the Ugly... and the good, maybe? In: 2024 USENIX Conference on Privacy Engineering Practice and Respect. USENIX Association, Santa Clara, CA (2024), https://www.usenix.org/conference/pepr24/presentation/desfontaines
5. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014). https://doi.org/10.1561/0400000042
6. Giomi, M., Boenisch, F., Wehmeyer, C., Tasnádi, B.: A unified framework for quantifying privacy risk in synthetic data. Proceedings on Privacy Enhancing Technologies **2023**, 312–328 (2023). https://doi.org/10.56553/popets-2023-0055
7. Golob, S., Pentyala, S., Maratkhan, A., Cock, M.D.: High epsilon synthetic data vulnerabilities in MST and PrivBayes (2024), https://arxiv.org/abs/2402.06699
8. Golob, S., Pentyala, S., Maratkhan, A., Cock, M.D.: Privacy vulnerabilities in marginals-based synthetic data. In: 3rd IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) (2025), https://arxiv.org/abs/2410.05506
9. Hayes, J., Melis, L., Danezis, G., Cristofaro, E.D.: LOGAN: Membership inference attacks against generative models. Proceedings on Privacy Enhancing Technologies (2019), https://petsymposium.org/popets/2019/popets-2019-0008.php
10. Hilprecht, B., Härterich, M., Bernau, D.: Monte Carlo and reconstruction membership inference attacks against generative models. Proceedings on Privacy Enhancing Technologies **2019**(4), 232–249 (2019). https://doi.org/10.2478/popets-2019-0067
11. Houssiau, F., Cohen, S.N., Szpruch, L., Daniel, O., Lawrence, M.G., Mitra, R., Wilde, H., Mole, C.: A framework for auditable synthetic data generation (2022), https://arxiv.org/abs/2211.11540
12. Houssiau, F., Jordon, J., Cohen, S.N., Elliott, A., Geddes, J., Mole, C., Rangel-Smith, C., Szpruch, L.: Prive: Empirical privacy evaluation of synthetic data generators. In: NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research (2022), https://openreview.net/forum?id=9hXskf1K7zQ
13. Lin, Z., Jain, A., Wang, C., Fanti, G., Sekar, V.: Using gans for sharing networked time series data: Challenges, initial promise, and open questions.

In: Proceedings of the ACM Internet Measurement Conference. p. 464–483. IMC '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3419394.3423643

14. McKenna, R., Miklau, G., Sheldon, D.: Winning the NIST contest: A scalable and general approach to differentially private synthetic data. Journal of Privacy and Confidentiality **11**(3) (2021). https://doi.org/10.29012/jpc.778
15. Pang, W., Shafieinejad, M., Liu, L., Hazlewood, S., He, X.: Clavaddpm: Multi-relational data synthesis with cluster-guided diffusion models. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 83521–83547. Curran Associates, Inc. (2024), https://doi.org/10.48550/arXiv.2405.17724
16. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 3–18 (2017). https://doi.org/10.1109/SP.2017.41
17. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data – anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 1451–1468. USENIX Association, Boston, MA (2022), https://www.usenix.org/conference/usenixsecurity22/presentation/stadler
18. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning **1**(1–2), 1–305 (2008). https://doi.org/10.1561/2200000001
19. Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J.: Differentially private generative adversarial network (2018), https://arxiv.org/abs/1802.06739
20. Yoon, J., Drumright, L.N., van der Schaar, M.: Anonymization through data synthesis using generative adversarial networks (ads-gan). IEEE Journal of Biomedical and Health Informatics **24**(8), 2378–2388 (2020). https://doi.org/10.1109/JBHI.2020.2980262
21. Yoon, J., Jordon, J., van der Schaar, M.: PATE-GAN: Generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=S1zk9iRqF7
22. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: PrivBayes: Private data release via bayesian networks. ACM Trans. Database Syst. **42**(4) (2017). https://doi.org/10.1145/3134428