Exploring Privacy-Preserving Techniques on Synthetic Data as a Defense Against Model Inversion Attacks

Manel Slokom[1,2,3](✉), Peter-Paul de Wolf[3], and Martha Larson[4]

[1] Delft University of Technology, Delft, The Netherlands
m.slokom@tudelft.nl
[2] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[3] Statistics Netherlands, The Hague, The Netherlands
pp.dewolf@cbs.nl
[4] Radboud University, Nijmegen, The Netherlands
m.larson@cs.ru.nl

Abstract. In this work, we investigate privacy risks associated with model inversion attribute inference attacks. Specifically, we explore a case in which a governmental institute aims to release a trained machine learning model to the public (i.e., for collaboration or transparency reasons) without threatening privacy. The model predicts change of living place and is important for studying individuals' tendency to relocate. For this reason, it is called a *propensity-to-move model*. Our results first show that there is a potential leak of sensitive information when a propensity-to-move model is trained on the original data, in the form collected from the individuals. To address this privacy risk, we propose a data synthesis + privacy preservation approach: we replace the original training data with synthetic data on top of which we apply privacy preserving techniques. Our approach aims to maintain the prediction performance of the model, while controlling the privacy risk. Related work has studied a one-step synthesis of privacy preserving data. In contrast, here, we first synthesize data and then apply privacy preserving techniques. We carry out experiments involving attacks on individuals included in the training data ("inclusive individuals") as well as attacks on individuals not included in the training data ("exclusive individuals"). In this regard, our work goes beyond conventional model inversion attribute inference attacks, which focus on individuals contained in the training data. Our results show that a propensity-to-move model trained on synthetic training data protected with privacy-preserving techniques achieves performance comparable to a model trained on the original training data. At the same time, we observe a reduction in the efficacy of certain attacks.

Keywords: Synthetic data · privacy-preserving techniques · propensity-to-move · model inversion attack · attribute inference attack · machine learning

P.-P. de Wolf—The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

E. Athanasopoulos and B. Mennink (Eds.): ISC 2023, LNCS 14411, pp. 3–23, 2023.
https://doi.org/10.1007/978-3-031-49187-0_1

1 Introduction

A governmental institute that is responsible for providing reliable statistical information may use machine learning (ML) approaches to estimate values that are missing in their data or to infer attributes whose values are not possible to collect. Ideally, the machine learning model that is used to make the estimates can be made available outside of the institute in order to promote transparency and support collaboration with external parties. Currently, however, an important unsolved problem stands in the way of providing external access to machine learning models: the models may pose a privacy threat because they are susceptible to *model inversion attribute inference attacks*. In other words, they may leak information about sensitive characteristics of individuals whose data they were trained on ("inclusive individuals"). Further, going beyond the strict definition of model inversion, access to models may enable the inference of attributes of individuals whose data is not included in the original training set ("exclusive individuals").

In this paper, we investigate the potential leaks that could occur when external access is provided to machine learning models. We carry out a case study on a model that is trained to predict whether an individual is likely to move or to relocate within the next two years. Such models are helpful for understanding tendencies in the population to change their living location and are, for this reason, called *propensity-to-move models*. We study the case in which an institute would like to provide access to the model by allowing external parties to query the model and receive output predictions and by releasing the marginal distributions of the data the model is trained on. Additionally, the output might include confidence scores. Finally, access might include releasing a confusion matrix of the model calculated on the training data. Attackers wish to target a certain set of target individuals to obtain values of sensitive attributes for these individuals. We assume that for this set of target individuals, attackers possess a set of non-sensitive attributes that they have previously obtained, e.g., by scraping social media, including the correct value for the propensity-to-move attribute.

First, we show the effectiveness of our propensity-to-move prediction model. Then, we evaluate a number of existing model inversion attribute inference attacks [14,28] and demonstrate that, if access would be provided to the model, a privacy threat would occur. Next, we address this threat by proposing a synthesis + privacy preservation approach, which applies privacy preserving techniques designed to inhibit attribute inference attacks on top of synthetic data. This two-step approach is motivated by the fact that within our case study, training models on synthetic data is an already established practice and the goal is to address the threat posed by synthetic data. In our previous work [42], we demonstrated that training on synthetic data has the potential to provide a small measure of protection, and here we build on that result.

Our results show that a propensity-to-move model trained on data created with our synthesis + privacy preservation approach achieves performance comparable to a propensity-to-move model trained on original training data. We also observe that the data created by our synthesis + privacy preservation approach contributes to the reduced success of certain attacks over a certain group of target individuals. Last but not least, we use the Correct Attribution Probability (CAP) metric [27] from Statistical Disclosure Control as a disclosure risk measure to calculate the risk of attribute disclosure for individuals.

We summarize our contributions as follows:

- **Threat Model:** Our attacks consider both target individuals who are included in the data on which the model is trained ("inclusive individuals") and target individuals who are *not* ("exclusive individuals"). Studying exclusive individuals goes beyond the strict definition of model inversion and is not well-studied in the literature.
- **Data synthesis + privacy preservation:** We explore a two-step approach that applies privacy-preserving techniques on top of synthetic data. Our approach aims to maintain model utility, i.e., the prediction performance of the model, while at the same time inhibiting inference of the sensitive attributes of target individuals.
- **Disclosure Risk:** In contrast to measures that rely on machine learning metrics, which often average or aggregate scores, we employ the Correct Attribution Probability (CAP) to quantify the level of disclosure risk for individual cases.

2 Threat Model

We start characterizing the case we study in terms of a threat model [39], a theoretical formulation that describes: the adversary's objective, the resources at the adversary's disposal, the vulnerability that the adversary seeks to exploit, and the types of countermeasures that come into consideration. Table 1 presents our threat model. We cover each of the dimensions, in turn, explaining their specification for our case.

As objective, the attacker seeks to infer sensitive information about a set of target individuals. As resources, we assume that the attacker has collected a set of data for each target individual, i.e., from previous data releases or social media. The set contains non-sensitive attributes of the target individuals and that includes the individual's ID and the corresponding true label for propensity-to-move. The target individuals are either in the training data used to train the released model ("inclusive individuals") or not in the training data ("exclusive individuals"). The vulnerability is related to how the model is released, i.e., the access that has been provided to the model. The attacker can query the model and collect the output of the model, both predictions and confidence scores, for unlimited number of inputs. The attacker also has information about the marginal distribution for each attribute in the training data. The countermeasure that we study is a change in the model that is released, which is accomplished by modifying the training data.

Table 1. Model inversion attribute inference threat model, defined for our case.

Component	Description
Adversary: Objective	Specific sensitive attributes of the target individuals
Adversary: Resources	A set of non-sensitive attributes of the target individuals, including the correct value for the propensity-to-move attribute, for "inclusive individuals" (in the training set) or "exclusive individuals" (not in the training set)
Vulnerability:Opportunity	Ability to query the model to obtain output plus the marginal distributions of the data that the model was trained on. Additionally, the output might include confidence scores and a confusion matrix calculated on the training data might be available
Countermeasure	Modify the data on which the model is trained

3 Background and Related Work

In this section, we provide a brief overview of existing literature on data synthesis, privacy-preserving techniques, and model inversion attribute inference attacks.

3.1 Synthetic Data Generation

Synthetic data generation methods involve constructing a model of the data and generating synthetic data from this model. These methods are designed to preserve specific statistical properties and relationships between attributes in the original data [9,16,47]. Synthetic data generation techniques fall into two categories [20]: partially synthetic data and fully synthetic data. Partially synthetic data contain a mix of original and synthetic records [10]. Techniques to achieve partial synthesis replace only observed values for attributes that bear a high risk of disclosure (i.e., sensitive attributes) [11]. Fully synthetic data, which we use in our experiments, creates an entirely synthetic data set based on the original data [10,11]. Next, we discuss existing work on fully synthetic data generation from Statistical Disclosure Control [9,47] and deep learning [48,51].

Data Synthesis in Statistical Disclosure Control. Several approaches have been proposed in the literature for generating synthetic data, such as data distortion by probability distribution [23], synthetic data by multiple imputation [38], and synthetic data by Latin Hypercube Sampling [8]. In [12], the authors proposed an empirical evaluation of different machine learning algorithms, e.g., classification and regression trees (CART), bagging, random forests, and Support Vector Machines for generating synthetic data. The authors showed that data synthesis using CART results in synthetic data that provides reliable predictions and low disclosure risks. CART, being a non-parametric method, helps in handling mixed data types and effectively captures complex relationships between attributes [12].

Data Synthesis Using Generative Models. A lot of research has been carried out lately focusing on tabular data synthesis [7,31,51]. In [7], the authors proposed *MedGAN*, one of the earliest tabular GAN-based data synthesis used to generate synthetic Health Records. MedGAN transformed binary and categorical attributes into a continuous space by combining an auto-encoder with GAN. In [31], the authors proposed *TableGAN*, a GAN-based method to synthesize fake data that are statistically similar to the original data while protecting against information leakage, e.g., re-identification attack and membership attack. TableGAN uses a convolutional neural network that optimizes the label column's quality such that the generated data can be used to train classifiers. In [51], the authors pointed out different shortcomings that were not addressed in previous GAN models, e.g., a mixture of data types, non-Gaussian and multimodal distribution, learning from sparse one-hot encoded vectors and the problem of highly imbalanced categorical attributes. In [51], a new GAN model called *CTGAN* is introduced, which uses a conditional generator to properly model continuous and categorical columns.

3.2 Privacy-Preserving Techniques

In this section, we provide an overview of existing work on privacy-preserving techniques. Privacy-preserving techniques can be categorized as perturbative or non perturbative methods. Perturbative methods involve introducing slight modifications or noise to the original data to protect privacy, while non perturbative methods achieve privacy through data transformation techniques without altering the data itself [47]. These techniques, which have been studied for many years, include randomization, data shuffling, data swapping [29,33], obfuscation [4], post-randomization [50]. We discuss the privacy-preserving techniques that we use in our experiments in more depth:

Data swapping is a non-perturbative method that is based on randomly interchanging values of an attribute across records. Swapping maintains the marginal distributions in the shuffled data. By shuffling values of sensitive attributes, data swapping provides a high level of utility while minimizing risk of disclosure [29].

Post-randomization (PRAM) is a perturbative method. Applying PRAM to a specific attribute (or a number of attributes) means that the values of the record in the PRAMmed attribute will be changed according to a specific probability. Following notations used in [50], let ξ denote the categorical attribute in the original data to which PRAM will be applied. X denotes the same categorical attribute in the PRAMmed data. We suppose that ξ and X have K categories $1, \ldots, K$. $p_{kl} = \mathbb{P}(X = l|\xi = k)$ denotes the transition probabilities that define PRAM. This means the probability that an original value $\xi = k$ is changed to value $X = l$ for $k, l = 1, \ldots, K$. Using the transition probabilities as entries of a $K \times K$ matrix, we obtain $\boldsymbol{P}$ (called the PRAM-matrix).

Differential privacy has gained a lot of attention in recent years [1,22]. Differential privacy (DP) uses a mathematical formulation to measure privacy. DP creates differentially private protected data by injecting noise expressed by ϵ into the original data. In [52] a differentially private Bayesian Network, PrivBayes is proposed to make possible the release of high-dimensional data. PrivBayes first constructs a Bayesian network that captures the correlations among the attributes and learns the distribution of data. After that, PrivBayes injects noise to ensure differential privacy and it uses the noisy marginals and the Bayesian network to construct an approximation of the data distribution. In [34], the authors introduced two methods for creating differentially private synthetic data. The first method adds noise to a cross-tabulation of all the attributes and creates synthetic data by a multinomial sampling from the resulting probabilities. The second method uses an iterative proportional fitting algorithm to obtain a fit to the probabilities computed from noisy marginals. Then, it generates synthetic data from the resulting probability distributions. A more recent work, Differentially Private CTGAN (DPCTGAN) [13] adds a differentially private noise to CTGAN. Specifically, DPCTGAN adds $\epsilon - \delta$ noise to the discriminator $\mathcal{D}$ and clips the norm to make it differentially private. We consider DPCTGAN to be a one-step synthesis approach, as it combines the application of noise and the synthesis process. Here, we test DPCTGAN, alongside our two-step synthesis + privacy preservation approaches.

3.3 Model Inversion Attribute Inference Attacks

Privacy attacks on data [25] include identification (or identity disclosure) attacks [2,3,51], membership inference attacks [41], and attribute inference attacks (or attribute disclosure) [3,19,44]. A lot of attention has been given to identification attacks on synthetic data [26,40,43]. However, less attention has been given to attribute inference attacks on synthetic data [40]. Attacks on data include attacks on models aimed at acquiring information about the training data. Here we investigate a model inversion attribute inference attack.

Model inversion attacks (MIA) aim to reconstruct the data a model is trained on or expose sensitive information inherent in the data [18,49]. Attribute inference attacks use machine learning algorithms to predict, and perform attacks that infer sensitive attributes, i.e., gender, age, income. In a model inversion attribute inference attack, the attacker is interested in inferring sensitive information, e.g., demographic attributes, about an individual [14,25,28].

We distinguish between three categories of model inversion attribute inference attacks [18,25]. An attack is black-box if the attacker only gets access to predictions generated by the model, i.e., can query the model with target individuals to receive the model's output. An attack is gray-box if the structure of the model and or some auxiliary information is further known, e.g., the attacker knows that the prediction is based on decision tree model, or attacker knows about the estimated weights of the model. An attack is white-box if an attacker has the full model, e.g., predictions, estimated weights or structure of model, and other information about training data.

In [14,15], the authors showed that it is possible to use black-box access to prediction models (access to commercial machine learning as a service APIs such as BigML) to learn genomic information about individuals. In [14], the authors developed an attack model that exploits adversarial access to a model to learn information about its training data. To perform the attack, the adversary uses the confidence scores included with the predictions as well as the confusion matrix of the target model and the marginal distributions of the sensitive attributes. In [28], the authors proposed two attack models: confidence score-based MIA (CSMIA) and label-only MIA (LOMIA). CSMIA exploits confidence scores returned by the target model. Different from Fredrikson et al. [14], in CSMIA an attacker is assumed to not have access to the marginal distributions or confusion matrix. LOMIA uses only the model's predicted labels. CSMIA, LOMIA, and Fredrikson et al., [14] are the attacks we study in our work. The three attacks aim to achieve the adversary's objective of inferring sensitive attributes about target individuals, while assuming different resources and opportunities available to the attacker. (Further details are in Sect. 4.4). Other model inversion attacks use variational inference [49] or imputation [21] to infer sensitive attributes.

3.4 Attribute Disclosure Risk

Previous work on identity and attribute disclosure risk has looked either at matching probability by comparing perceived, expected, and true match risk [36], or at a Bayesian estimation approach, assuming that an attacker seeks a Bayesian posterior distribution [37]. Similar to [36], other work [19,27,46] has looked at the concept of Correct Attribution Probability (CAP).

CAP assumes that the attacker knows the values of a set of key attributes for an individual in the original data set, and aims to learn the respective value of a target attribute. The key attributes encompass all attributes within the data, excluding the sensitive attribute that is the target attribute. Correct Attribution Probability (CAP) measures the disclosure risk of the individual's real value in the case where an adversary has access to protected data, and was originally proposed for synthetic data [19,46]. The basic idea of CAP is that an attacker is supposed to search for all records in the synthetic data that match records in the original data for given key attributes. The CAP score is the proportion of matches leading to correct attribution out of the total matches for a given individual [46]. In [46], the authors extended their previous preliminary work [27]. They proposed a new CAP measure called differential correct attribution probability (DCAP). DCAP captures the effect of multiple imputations on the disclosure risk of synthetic data. The authors of [46] stated that DCAP is well-suited for fully synthetic data. In [24], the authors introduced TCAP, for targeted correct attribution probability. TCAP calculates CAP value for targeted individuals that the attacker knows their existence in the original data. In our experiments, we use the CAP measure introduced in [27].

4 Experimental Setup

In this section, we describe our experimental setup. First, we provide an overview of our data set. Second, we describe how we synthesize data and the privacy protection techniques that we use. Next, we discuss target machine learning algorithms that we will use to predict propensity-to-move. Then, we describe the model inversion attribute inference attacks we study in our experiments.

4.1 Data Set

For our experiments, we use a data set from a governmental institute. The data set was previously collected and first used in [5]. It combines different registers from the System of Social Statistical Data sets (SSD). In our experiments, we use the same version of the data set used in [42]. Our data contains 150K individuals' records between 2013 and 2015. We have 40 attributes (categorical and numerical) containing information about individual demographic attributes such as gender and age, and time-dependent personal, household, and housing attributes. The target attribute "$y01$" is binary, indicating whether (=1) or not (=0) a person moved in year j where $j = 2013, 2015$. The target attribute is imbalanced with 129428 0 s (majority class) and 24971 1 s (minority class).

We have three distinct groups of individuals within the data. The difference between the three groups resides in the fact that there are some individuals who are in the data in the year 2013 (called Inclusive individuals 2013). The same individuals appear again in the year 2015 (called Inclusive individuals 2015), where they may have different values for the time-dependent attributes than they did in 2013. The last group (called Exclusive individuals 2015) contains individuals who are "new in the country". We have a total of: 76904 Inclusive individuals 2013, 74591 Inclusive individuals 2015, and 2904 Exclusive individuals 2015.

Our propensity-to-move classifier (i.e., the target model) is trained on all 2013 data (76904 records). The classifier is tested on the 2015 data (77495 records) as in [42]. For the target model trained on (privacy-preserving) synthetic data, we use $TSTR$ evaluation strategy such that we train classifiers on 2013 (privacy-preserving) synthetically generated data and we test on 2015 original data [17, 42].

As adversary resources, we assume that the attacker has access to a set of non-sensitive attributes of the target individuals (see our threat model in Sect. 2). As in [42], we consider three cases:

- `Inclusive individuals (2013)`: the attacker has access to data from the year 2013, which aligns with the data used to train the target model.
- `Inclusive individuals (2015)`: Here, the attacker possesses more recent data from 2015, but it corresponds to the same set of individuals used in training the target model. The data being more recent implies that some of the (time-sensitive) attributes for particular individuals may have changed somewhat.

- `Exclusive individuals (2015)`: In this case, the attacker's data is from 2015, but it pertains to a distinct group of individuals who were not part of the training set for the target model.

We create data sets for each of the three cases. As in [42], for Exclusive individuals (2015) we use all 2904 individuals and for the other two cases we randomly sample to create data sets of the same size (2904 individuals each). The attributes of the target individuals that are in the possession of the attacker include the correct value of the propensity-to-move attribute but do not include the sensitive attributes gender, age, and income, which are targeted by the attack.

4.2 Privacy-Preserving Techniques on Synthetic Training Data

In this section, we describe how we synthesized data, and how we then applied privacy preserving approaches to it. The synthesis and privacy-preserving techniques are applied to the training data of the target model (the 76904 Inclusive individuals 2013), which is intended for release.

Our experiments with our two-step synthesis + privacy protection approach use a *classification and regression tree* (CART) model to synthesize data since it is shown to perform the best in the literature [12,35]. Recall that CART is a non-parametric method that can handle mixed data types and is able to capture complex and non-linear relationships between attributes. We apply CART to the training data of the target model, which includes individuals from 2013. We use the open public R package, Synthpop for our implementation of the CART model [30][1]. Within Synthpop, there are a number of parameters that can be optimized to achieve a good quality of synthesis [30]. *Visiting.sequence* parameter specifies the order in which attributes are synthesized. The order is determined institute-internally by a human expert. *Stopping rules* parameter dictates the number of observations that are assigned to a node in the tree. Stopping rules parameter helps to avoid over-fitting.

Following synthesis using CART, we apply privacy-preserving techniques, data swapping and PRAM (cf. Sect. 3.2), to the synthetic data. We use two data swapping approaches, referred to as *Swapping* and *Conditional swapping.* For Swapping, we perform data swapping separately for each sensitive attribute, which includes gender, age, and income. Specifically, for the age attribute, we interchange numerical age values among individuals and subsequently map these values to their respective age groups. For Conditional swapping, we perform simultaneous data swapping for gender, age, and income conditioned on the propensity-to-move target attribute. Conditional data swapping ensures that sensitive attributes are swapped while preserving the influence of the target attribute. Additionally, we apply Post-randomization (PRAM) independently to the attributes of gender, age, and income within the synthetic data generated using CART. Our transition matrices can be found in supplementary material.[2]

[1] http://www.synthpop.org.uk/.

[2] Supplemental material is at this link in Section.2: PRAM.

We use the sdcMicro toolkit.[3] It is important to note that our evaluation includes separate testing of PRAM and data-swapping techniques.

In addition to experiments with our two-step synthesis + privacy protection approach, we explore a GAN-based one-step approach for generating (privacy preserving) synthetic data generation. We use *CTGAN*, a popular and widely used GAN-based generative model [51]. The data synthesis procedure of CTGAN involves three key elements, namely: the conditional vector, the generator loss, and the training-by-sampling method. CTGAN uses a conditional generator to deal with the class imbalance problem. The conditional generator generates synthetic rows conditioned on one of the discrete columns. With training-by-sampling, the conditional and training data are sampled according to the log frequency of each category. We used open public toolkit Synthetic Data Vault (SDV)[4] implemented in Python [32]. In our implementation, hyperparameter tuning is applied to batch size, number of epochs, generator dimension, and discriminator dimension. We left other parameters set to default. We generate differentially private CTGAN data using DPCTGAN, which takes the state-of-the-art CTGAN and incorporates differential privacy. We chose to make a comparison with CTGAN and DPCTGAN because of the success of the two techniques reported in the literature [51].

4.3 Target Machine Learning Model

In this section, we discuss the target machine learning algorithm used to predict the propensity to move. We trained and tested a number of machine learning algorithms, including decision tree, random forest, naïve Bayes, and extra trees. We found that all classifiers outperform the majority-class classifier, with classifiers using trees generally being the best performers. For simplicity, in the rest of the paper, we will use random forest classifier as it is shown to perform the best on the original data and on the synthetic data. We report the results of random classifier using the most frequent (majority-class) strategy as a naïve baseline.

Recall that we must ensure that the prediction performance of the model is maintained when it is trained on synthetic + privacy-preservation data. To this end, we use the following metrics: F1-Macro, Matthews Correlation Coefficient (MCC), geometric mean (G-mean), True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). Our choice is motivated by the imbalance of the target attribute.

The macro-averaged F1 score (F1-Macro) is computed using the arithmetic mean (i.e., unweighted mean) of all the per-class F1 scores. This method treats all classes equally regardless of their support values.

The Geometric mean (G-mean) is the geometric mean of sensitivity and specificity [45]. G-mean takes all of the TP, TN, FP, and FN into account.

$$\text{G-mean} = \sqrt{\frac{TP}{TP+FN} * \frac{TN}{TN+FP}} \qquad (1)$$

[3] https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf.

[4] https://github.com/sdv-dev/SDV.

Matthews Correlation Coefficient (MCC) metric is a balanced measure that can be used especially if the classes of the target attribute are of different sizes [6]. It returns a value between -1 and 1.

$$\text{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{2}$$

4.4 Model Inversion Attribute Inference Attacks

In this section, we describe three model inversion attacks that we use in our paper: confidence-score MIA (CSMIA) [28], label-only MIA (LOMIA + Marginals), and the Fredrikson et al. MIA (FMIA) [14].

Confidence-Score MIA (CSMIA) [28] uses the output and confidence scores returned when the attacker queries the target propensity-to-move model. The attacker also has knowledge of the possible values for the sensitive attribute. For each target individual, the attacker creates different versions of the individual's records by substituting in for the missing sensitive attribute all values that would be possible for that attribute. The attacker then queries the model with each version and obtains the predicted class labels and the corresponding model confidence scores. Then, the attacker uses the predicted labels and confidence scores as follows [28]:

`Case (1)`: when the target model's prediction is *correct for only a single* sensitive attribute value, then, the attacker selects the sensitive attribute value to be the one for which the prediction is correct.

`Case (2)`: when target model's prediction is *correct for multiple* sensitive attribute values, then the attacker selects the sensitive value to be the one for which prediction confidence score is maximum.

`Case (3)`: when target model's prediction is *incorrect for all* sensitive attribute values, then the attacker selects the sensitive value to be the one for which prediction confidence score is minimum.

Label-Only MIA with Marginals (LOMIA + Marginals) is based on the LOMIA attack proposed by [28]. LOMIA + Marginals uses the output returned when the attacker queries the target propensity-to-move model and the marginal distributions of the training data (which includes the information about the possible values of sensitive attributes).

As with CSMIA, for each target individual, the attacker queries the target model multiple times, varying the value of the sensitive attribute. To determine the value of the sensitive attribute, the attacker follows `Case (1)` of CSMIA, as described in [28]. Specifically, if the target model's prediction is correct for a single sensitive attribute value, the attacker selects that value as the sensitive attribute. Differently from [28], for cases where the attacker cannot infer the sensitive attribute, we do not run an auxiliary machine learning model. Instead, the attacker uses the released marginal distribution to predict the most probable value of the sensitive attribute.

The Fredrikson et al. MIA (FMIA) [14] uses the output returned when the attacker queries the target propensity-to-move model and the marginal distributions of the training data. Following the threat model of [14], the attacker also has access to a confusion matrix of the target model's predictions on its training data. As with CSMIA and LOMIA + Marginals, the attacker queries the target model multiple times for each target individual, changing the sensitive attribute to take on all possible values and obtaining the predicted labels. Next, the attacker calculates the product of the probability that the target model's prediction aligns with the true label and the marginal distribution for each potential sensitive attribute value across all possibilities. Then, the attacker predicts the sensitive attribute value for which this product is maximized.

Measuring Success of Attribute Inference Attack. We use two ways to measure attribute inference attacks:

(1) From a machine learning perspective, we evaluate the success of the attack by measuring precision (also called the positive predicted value (PPV) [21]). The precision metric measures the ratio of true positive predictions considering all positive predictions. A precision score of 1 indicates that the positive predictions of the attack are always correct.

(2) From statistical disclosure control, we use CAP to measure the disclosure risk of the individuals. Following [46], we define D_{org} as the original data and K_{org} and T_{org} as vectors for the key and target sensitive attributes of the original data: $D_{org} = \{K_{org}, T_{org}\}$. Similarly, we denote by D_{syn} as the synthetic data and K_{syn} and T_{syn} as the vectors for the key and target sensitive attributes of the synthetic data: $D_{syn} = \{K_{syn}, T_{syn}\}$. Note that when we are calculating CAP, the synthetic data we use is the data reconstructed by the attacker by inferring the missing sensitive value and adding it to the previously-possessed non-sensitive attributes used for the attack. We consider gender, age, and income as target sensitive attributes, evaluating CAP for each sensitive attribute separately. Key attributes are all other attributes for an individual except for the sensitive attribute being measured by CAP. The CAP for a record j is the probability of its target attributes given its key attributes.

$$\mathrm{CAP}_{org,j} = Pr(T_{org,j}|K_{org,j}) = \frac{\sum_{i=1}^{M} [T_{org,i} = T_{org,j}, K_{org,i} = K_{org,j}]}{\sum_{i=1}^{M} (K_{org,i} = K_{org,j})} \tag{3}$$

where M is the number of records. The CAP score for the original data is considered as an approximate upper bound. Then, the CAP for the record j based on a corresponding synthetic data D_{syn} is the same probability but derived from synthetic data D_{syn}.

$$\mathrm{CAP}_{syn,j} = (Pr(T_{org,j}|K_{org,j}))_{syn} = \frac{\sum_{i=1}^{M} [T_{syn,i} = T_{org,j}, K_{syn,i} = K_{org,j}]}{\sum_{i=1}^{M} (K_{syn,i} = K_{org,j})} \tag{4}$$

CAP has a score between 0 and 1: a low score (close to 0) indicates that the synthetic data has a little risk of disclosure and a high score (close to 1) indicates a high risk of disclosure.

5 Performance of the Target Models

In this section, we compare the performance of the target propensity-to-move models. We evaluate whether a random forest classifier trained on protected synthetic data can attain performance comparable to a random forest classifier trained on the original data. Our results are reported in Table 2. Column "privacy-preservation" provides different privacy-preserving techniques that we applied to synthetic training data. "Privacy-preservation" with "None" means that there are no privacy-preserving techniques applied on top of the synthesis.

In Table 2, we see that random forest classifier trained on synthetic data using CART with *None* (i.e., no privacy-preserving technique applied) has quite close and comparable results to random forest classifier trained on original data. As a sanity check, we observe that both outperform the majority-class classifier.

Table 2. Classification performance of the target model. We generate synthetic data using CART and CTGAN. For privacy-preserving techniques, we used swapping, conditional swapping, PRAM, and differential privacy ($\epsilon = 3$). In each case, the test data is used in its original (unprotected) form.

Target MLs to be Released	Data sets	Privacy-preservation	F1-Macro	MCC	G-mean	TN	FP	FN	TP
Majority-class	**Original data**	*None*	0.4924	0.0012	0.4924	46452	9539	17818	3686
Random Forest	**Original Data**	*None*	0.5946	0.2407	0.5779	61907	2363	10677	2548
Random Forest	**Synthetic data using CART**	*None*	0.5946	0.2426	0.5793	61848	2422	10628	2597
		Swapping	0.5881	0.2389	0.5742	62174	2096	10831	2394
		Conditional swapping	0.4654	0.0216	0.5028	63704	566	13034	191
		PRAM	0.5941	0.2415	0.5789	61844	2426	10638	2587
	Synthetic data using CTGAN	*None*	0.4586	0.0392	0.5021	64207	63	13155	70
		Differential privacy	0.4534	0.000	0.5000	64270	0	13225	0

We observe that in two cases the model trained on our synthesis + privacy preservation data retains a level of performance comparable to a model trained on the original data: CART with *Swapping* and CART with *PRAM*. Surprisingly, we find that when the training data is created with CART synthesis and *Conditional swapping* or CTGAN (with or without *Differential privacy*) the performance is comparable to that of a majority-class classifier. This result suggests that the use of conditional swapping and differential privacy may not effectively preserve the utility of the propensity-to-move data. For the rest of the paper, we will assume that we intend to release machine learning models trained on synthetic data using CART with: *None*, *Swapping*, and *PRAM* as privacy-preserving techniques.

6 Results of Model Inversion Attribute Inference Attacks

In this section, we report the performance of different model inversion attribute inference attacks. We evaluate the performance of attacks on the model when it is trained on the original training data. Then, we investigate whether training the model on synthesis + privacy preservation data can protect against model inversion attribute inference attacks.

6.1 Attacks on the Model Trained on Original Data

First, we look at the performance of model inversion attribute inference attacks on the target model trained on original training data. The results are reported in Table 3.

Table 3. Results of model inversion attribute inference attacks measured using precision (positive predictive value) for three different target individual sets. The target propensity-to-move model is trained on **original training data**. Numbers in bold and italic represent the first and second best inference scores across conditions. A high precision indicates that the attack is good at correctly inferring the sensitive attribute values. We run experiments ten times and we report average scores. The standard deviation is below 0.01.

Adversary Resources	Inclusive individuals (2013)			Inclusive individuals (2015)			Exclusive individuals (2015)		
Attack models	*Gender*	*Age*	*Income*	*Gender*	*Age*	*Income*	*Gender*	*Age*	*Income*
Marginals Only	0.4977	0.1238	0.1982	0.5029	0.1244	0.1991	0.5012	0.1275	0.2001
CSMIA	0.3206	0.0105	0.0514	0.4660	0.0638	0.1581	0.4943	0.0721	0.1602
LOMIA + Marginals	*0.5157*	*0.1336*	*0.2105*	**0.5035**	**0.1291**	0.1983	*0.5014*	0.1234	**0.2005**
FMIA	**0.7563**	**0.6777**	**0.6898**	0.4647	0.0170	**0.2499**	**0.5205**	0.1091	0.1452

The attack models show varying performances compared to the Marginals Only Attack. We observe that attribute inference scores for the attack models "LOMIA + Marginals" and "FMIA" outperform the inference scores of the Marginals Only Attack. In particular, FMIA for Inclusive individuals (2013) achieves the highest precision for all three sensitive attributes gender, age, and income. It outperforms other attack models in terms of correctly predicting positive instances. LOMIA + Marginals shows moderate performance, obtaining precision values higher than Marginals Only Attack. The fact that the attack performance for Inclusive individuals (2013) is highest is not surprising since these individuals are in the training set of the target model. For Inclusive individuals (2015) and Exclusive individuals (2015), we see that the performance for all attack models is relatively low and comparable to the Marginals Only Attack, except for a few cases such as FMIA on age for Inclusive individuals (2015). Recall that for FMIA, the attacker is exploiting a larger opportunity for attack than for the other attacks. Specifically, the attacker can query the model but also possesses the marginal distributions of the training data and a confusion matrix (cf. Sect. 4.4. For this reason, it is not particularly surprising that FMIA is the strongest attack).

6.2 Attacks on the Model Trained on Protected Synthetic Data

Second, we investigate whether we can counter the attack by replacing original data used to train target model by a privacy-preserving synthetic data. The results of the model inversion attribute inference attacks are reported in Table 4.

Table 4. Results of model inversion attribute inference attacks measured using precision for three different target individual sets. The target propensity to move model is trained on *privacy-preserving (PP) + synthetic training* data. Numbers in bold and italic represent the first and second best inference scores across conditions. We run experiments ten times and we report average scores. The standard deviation is below 0.02.

PP+ Synthetic data	Attack Models	Inclusive individuals (2013)			Inclusive individuals (2015)			Exclusive individuals (2015)		
		Gender	*Age*	*Income*	*Gender*	*Age*	*Income*	*Gender*	*Age*	*Income*
Synthesis Only	*Marginals Only*	0.5036	0.1228	0.2021	0.4938	0.1225	0.2033	0.4979	**0.1233**	0.1980
	CSMIA	0.4901	0.0675	0.1423	0.4947	0.0775	0.1544	***0.5018***	0.1012	0.1826
	LOMIA + Marginals	0.4980	*0.1261*	0.1995	*0.5003*	**0.1282**	0.1972	*0.4989*	**0.1252**	0.1985
	FMIA	**0.5153**	0.0498	**0.3453**	**0.5007**	0.0588	**0.2772**	**0.5069**	0.1080	0.1452
Synthesis + Swapping	*Marginals Only*	0.4980	0.1238	0.1974	0.4979	0.1233	0.2060	0.4975	0.1248	**0.1973**
	CSMIA	0.4958	0.1198	**0.2032**	**0.4996**	0.1175	0.1848	***0.5093***	**0.1457**	*0.1986*
	LOMIA + Marginals	**0.5012**	**0.1280**	*0.1984*	0.4972	*0.1265*	0.1984	*0.5032*	**0.1242**	***0.1988***
	FMIA	0.4473	0.0901	0.0792	0.4320	**0.1362**	**0.3098**	**0.5351**	0.1020	0.1452
Synthesis + PRAM	*Marginals Only*	0.5002	0.1259	0.2010	0.5063	0.1239	0.2039	0.5002	0.1255	0.2000
	CSMIA	0.4967	0.1175	0.1701	0.4913	0.1059	0.1827	0.4895	**0.1371**	**0.2070**
	LOMIA + Marginals	**0.5038**	**0.1274**	0.1963	0.5004	0.1238	0.2002	0.5004	**0.1247**	**0.1987**
	FMIA	0.4827	0.0282	0.1635	**0.5286**	0.1129	0.1188	**0.5120**	0.1019	0.1452

Overall we see that the effectiveness of the synthesis + privacy-preserving techniques varies across different attributes, attack models, and adversary resources (target sets). While some attributes have an inference score higher than the inference score of the Marginals Only attack, others only have comparable performance to the Marginals Only attack. We notice a decrease in the performance of attack models specifically for Inclusive individuals (2013) compared to the performance of attack models for the same group of individuals in Table 3. For Inclusive individuals (2015) and Exclusive individuals (2015) which were not part of the training of the synthesis nor the training of the target model, we do not see a clear impact of privacy-preserving techniques on attack models. In most cases, the leak of sensitive information is low and comparable to the performance of the Marginals Only attack.

7 Correct Attribution Probability

Now, we shift our focus to calculate the risk of attribute disclosure for individual target subjects using CAP (Correct Attribution Probability). CAP captures how many specific individuals face a high risk of attribute disclosure and how many a lower risk. We measure CAP using Eq. 4, where D_{org} is the attacker's data with key attributes K_{org} and the original target sensitive attribute T_{org} (gender, age, income). D_{syn} represents the attacker's data where $K_{syn} = K_{org}$ are the key attributes and T_{syn} is the outcome of the model inversion attribute inference attacks.

Figure 1 and Fig. 2 show the frequency of CAP scores for sensitive attributes age and income, respectively. Due to space limitation, we specifically, focus on FMIA attack because it outperformed other attack models in Table 3. The top row of Fig. 1 and Fig. 2 shows the frequency of CAP scores on the original data

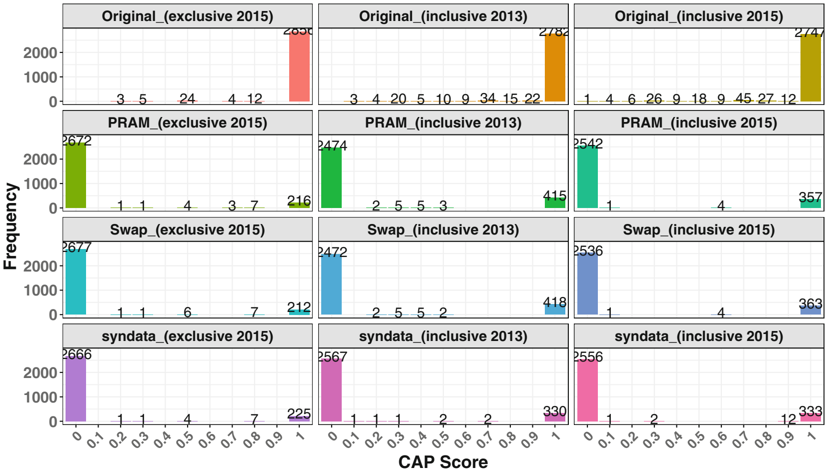

Fig. 1. Frequency of CAP scores for attribute *age*. The total number of queries is 2904. The numbers inside the bars represent the count of individuals with corresponding CAP scores.

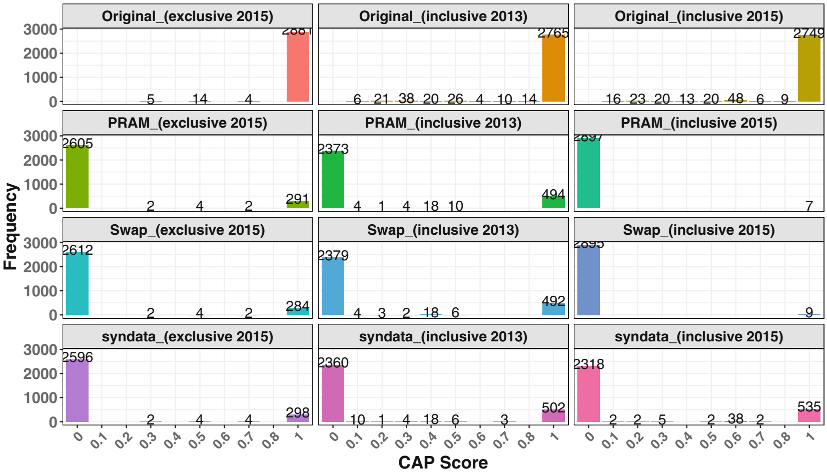

Fig. 2. Frequency of CAP scores for attribute *income*. The attack model is FMIA. The total number of queries is 2904. The numbers inside the bars represent the count of individuals with corresponding CAP scores.

(unprotected data). We see that across all three cases, Inclusive individuals (2013), Inclusive individuals (2015), and Exclusive individuals (2015), there is a high CAP score, signifying a high disclosure risk. However, when we calculate CAP scores based on the outcome of the model inversion attack, we observe that the risk of disclosure is relatively low, with approximately up to 92% of individuals considered protected. Only for the remaining individuals (8% individuals), we observe that an attacker can easily infer sensitive attributes age, and income with high CAP scores. Also, the number of disclosed individuals varies depending on the privacy-preserving technique applied. Comparing different resources, we see that for sensitive attribute age, Inclusive individuals (2013) have the highest number of disclosed individuals, next are Inclusive individuals (2015), and finally, Exclusive individuals (2015) have the lowest number of disclosed individuals. This aligns with the findings in Table 4. Notably, even though we generated privacy-preserving synthetic training data sets, the target model appears to retain some information about the original data, leading to a risk of disclosure for certain individuals.

8 Conclusion and Future Work

We have conducted an investigation aimed at protecting sensitive attributes against model inversion attacks, with a specific focus on a case study for a governmental institute. Our objective was to determine the feasibility of releasing a trained machine learning model predicting propensity-to-move to the public without causing privacy concerns. To accomplish this, we evaluated a number of existing privacy attacks, including CSMIA, LOMIA + Marginals, and FMIA, each distinguished by the resources available to the attacker. Our findings revealed that FMIA presented the highest degree of information leakage, followed by LOMIA + Marginals, while CSMIA exhibited the least leakage.

To mitigate these privacy risks, we employed privacy-preserving techniques on top of synthetic data utilized to train the machine learning model prior to its public release. Our results indicated that, in specific cases, such as with Inclusive individuals (2013), our privacy-preserving techniques successfully reduced information leakage. However, in other cases Inclusive individuals (2015) and Exclusive individuals (2015), the leakage remained comparable to that of a Marginals Only Attack, which uses the marginal distributions of the training data. We found a high disclosure risk, measured with CAP, when the target model is trained on original data. When the target model is trained on data protected with our two step synthesis + privacy preservation approach a lower percentage of individuals risk disclosure.

Furthermore, we think that the performance of the target machine learning model, as well as the correlation between the sensitive attribute and the target attribute, play a key role in the success of model inversion attacks. Future work should explore other case studies, in which this correlation might be different. Also, future work can look at other threat models such as white-box attacks, where the model predictions, model parameters, and explanation of the model's output are made public.

References

1. Abay, N.C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Sweeney, L.: Privacy preserving synthetic data release using deep learning. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11051, pp. 510–526. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10925-7_31
2. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proceedings of the ACM International Conference on Management of Data, vol. 29, pp. 439–450 (2000)
3. Andreou, A., Goga, O., Loiseau, P.: Identity vs. attribute disclosure risks for users with multiple social profiles. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 163–170 (2017)
4. Brunton, F., Nissenbaum, H.: Obfuscation: A User's Guide for Privacy and Protest. MIT Press, Cambridge (2015)
5. Burger, J., Buelens, B., de Jong, T., Gootzen, Y.: Replacing a survey question by predictive modeling using register data. In: ISI World Statistics Congress, pp. 1–6 (2019)
6. Chicco, D., Jurman, G.: The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. **21**(1), 1–13 (2020)
7. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W.F., Sun, J.: Generating multi-label discrete patient records using generative adversarial networks. In: Doshi-Velez, F., Fackler, J., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (eds.) Proceedings of the 2nd Machine Learning for Healthcare Conference, vol. 68, pp. 286–305 (2017)
8. Dandekar, R.A., Cohen, M., Kirkendall, N.: Sensitive micro data protection using Latin hypercube sampling technique. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 117–125. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-47804-3_9
9. Domingo-Ferrer, J.: A survey of inference control methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining. Advances in Database Systems, vol. 34, pp. 53–80. Springer, Boston (2008). https://doi.org/10.1007/978-0-387-70992-5_3
10. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, vol. 201. Springer, New York (2011). https://doi.org/10.1007/978-1-4614-0326-5
11. Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. Trans. Data Priv. **1**(3), 105–130 (2008)
12. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. Comput. Stat. Data Anal. **55**(12), 3232–3243 (2011)
13. Fang, M.L., Dhami, D.S., Kersting, K.: DP-CTGAN: differentially private medical data generation using CTGANs. In: Michalowski, M., Abidi, S.S.R., Abidi, S. (eds.) AIME 2022. LNCS, vol. 13263, pp. 178–188. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-09342-5_17
14. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM Conference on Computer and Communications Security, pp. 1322–1333 (2015)

15. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: 23rd USENIX Security Symposium, pp. 17–32. USENIX Association (2014)
16. Garofalo, G., Slokom, M., Preuveneers, D., Joosen, W., Larson, M.: Machine learning meets data modification. In: Batina, L., Bäck, T., Buhan, I., Picek, S. (eds.) Security and Artificial Intelligence. LNCS, vol. 13049, pp. 130–155. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98795-4_7
17. Heyburn, R., et al.: Machine learning using synthetic and real data: similarity of evaluation metrics for different healthcare datasets and for different algorithms. In: Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference, pp. 1281–1291. World Scientific (2018)
18. Hidano, S., Murakami, T., Katsumata, S., Kiyomoto, S., Hanaoka, G.: Exposing private user behaviors of collaborative filtering via model inversion techniques. In: Proceedings on Privacy Enhancing Technologies, no. 3, pp. 264–283 (2020)
19. Hittmeir, M., Mayer, R., Ekelhart, A.: A baseline for attribute disclosure risk in synthetic data. In: Proceedings of the 10th ACM Conference on Data and Application Security and Privacy, pp. 133–143 (2020)
20. Hundepool, A., et al.: Statistical Disclosure Control. Wiley, Hoboken (2012)
21. Jayaraman, B., Evans, D.: Are attribute inference attacks just imputation? In: Proceedings of the ACM Conference on Computer and Communications Security, pp. 1569–1582 (2022)
22. Li, H., Xiong, L., Zhang, L., Jiang, X.: DPSynthesizer: differentially private data synthesizer for privacy preserving data sharing. Proc. Very Large Data Bases (VLDB Endow.) **7**(13), 1677–1680 (2014)
23. Liew, C.K., Choi, U.J., Liew, C.J.: A data distortion by probability distribution. ACM Trans. Database Syst. **10**(3), 395–411 (1985)
24. Little, C., Elliot, M., Allmendinger, R.: Comparing the utility and disclosure risk of synthetic data with samples of microdata. In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 234–249. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13945-1_17
25. Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., Lin, Z.: When machine learning meets privacy: a survey and outlook. ACM Comput. Surv. **54**(2), 1–36 (2021)
26. Lu, P.H., Wang, P.C., Yu, C.M.: Empirical evaluation on synthetic data generation with generative adversarial network. In: Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, pp. 1–6 (2019)
27. Elliot, M.: Final report on the disclosure risk associated with synthetic data produced by the SYLLS team (2014). http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/. Accessed 13 Oct 2023
28. Mehnaz, S., Dibbo, S.V., Kabir, E., Li, N., Bertino, E.: Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In: Proceedings of the 31st USENIX Security Symposium, pp. 4579–4596. USENIX Association (2022)
29. Muralidhar, K., Sarathy, R.: Data shuffling: a new masking approach for numerical data. Manage. Sci. **52**(5), 658–670 (2006)
30. Nowok, B., Raab, G.M., Dibben, C.: Synthpop: bespoke creation of synthetic data in R. J. Stat. Softw. **74**(11), 1–26 (2016)
31. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on Generative Adversarial Networks. In: Proceedings of the 44th

International Conference on Very Large Data Bases (VLDB Endowment), vol. 11, no. 10, pp. 1071–1083 (2018)

32. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: IEEE International Conference on Data Science and Advanced Analytics, pp. 399–410 (2016)
33. Polat, H., Du, W.: Privacy-preserving collaborative filtering using randomized perturbation techniques. In: Proceedings of the 3rd IEEE International Conference on Data Mining, pp. 625–628 (2003)
34. Raab, G.M.: Utility and disclosure risk for differentially private synthetic categorical data. In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 250–265. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13945-1_18
35. Reiter, J.P.: Using CART to generate partially synthetic public use microdata. J. Off. Stat. **21**(3), 441 (2005)
36. Reiter, J.P., Mitra, R.: Estimating risks of identification disclosure in partially synthetic data. J. Priv. Confidentiality **1**(1) (2009)
37. Reiter, J.P., Wang, Q., Zhang, B.: Bayesian estimation of disclosure risks for multiply imputed, synthetic data. J. Priv. Confidentiality **6**(1) (2014)
38. Rubin, D.B.: Discussion statistical disclosure limitation. J. Off. Stat. **9**(2), 461–468 (1993)
39. Salter, C., Saydjari, O.S., Schneier, B., Wallner, J.: Toward a secure system engineering methodology. In: Proceedings of the Workshop on New Security Paradigms, pp. 2–10 (1998)
40. Shlomo, N.: How to measure disclosure risk in microdata? Surv. Stat. **86**(2), 13–21 (2022)
41. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: IEEE Symposium on Security and Privacy, pp. 3–18 (2017)
42. Slokom, M., de Wolf, P.P., Larson, M.: When machine learning models leak: an exploration of synthetic training data. In: Domingo-Ferrer, J., Laurent, M. (eds.) Proceedings of the International Conference on Privacy in Statistical Databases (2022). Corrected and updated version on arXiv at https://arxiv.org/abs/2310.08775
43. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data-anonymisation groundhog day. In: Proceedings of the 29th USENIX Security Symposium. USENIX Association (2020)
44. Sun, M., Li, C., Zha, H.: Inferring private demographics of new users in recommender systems. In: Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems, pp. 237–244 (2017)
45. Sun, Y., Wong, A.K., Kamel, M.S.: Classification of imbalanced data: a review. Int. J. Pattern Recognit. Artif. Intell. **23**(04), 687–719 (2009)
46. Taub, J., Elliot, M., Pampaka, M., Smith, D.: Differential correct attribution probability for synthetic data: an exploration. In: Domingo-Ferrer, J., Montes, F. (eds.) PSD 2018. LNCS, vol. 11126, pp. 122–137. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99771-1_9
47. Torra, V.: Privacy in data mining. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 687–716. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-09823-4_35

48. Tripathy, A., Wang, Y., Ishwar, P.: Privacy-preserving adversarial networks. In: 57th IEEE Annual Allerton Conference on Communication, Control, and Computing, pp. 495–505 (2019)
49. Wang, K.C., Fu, Y., Li, K., Khisti, A.J., Zemel, R., Makhzani, A.: Variational model inversion attacks. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 9706–9719 (2021)
50. Wolf, P.-P.: Risk, utility and PRAM. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 189–204. Springer, Heidelberg (2006). https://doi.org/10.1007/11930242_17
51. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alche Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 7335–7345 (2019)
52. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: PrivBayes: private data release via Bayesian networks. ACM Trans. Database Syst. **42**(4), 1–41 (2017)