IEEE Big Data 2023
This is a self-archived pre-print version of this article.
The final publication is available at IEEE via
https://doi.org/10.1109/BigData59044.2023.10386296.

Protecting Multiple Sensitive Attributes in Synthetic Micro-data

Nina Niederhametner
SBA Research
Vienna, Austria
Nina.Niederhametner@gmx.at

Rudolf Mayer
SBA Research
Vienna, Austria
rmayer@sba-research.org

***Abstract*—With the ever-increasing amount of data collected, there is also an increased demand for data analysis and machine learning methods, which are consequently frequently deployed. However, many of the data collected are very sensitive and of a personal nature – thus, data confidentiality and privacy become important considerations. In the wake of this, the use of synthetic data as a privacy-preserving measure for micro-data is gaining more and more popularity, especially due to its ability to maintain a high level of data utility. Synthetic data is artificially generated by a model that has been trained on real data. This means that the observations in the synthetic data do not directly correspond to any individual in the original dataset. While there are many tools for creating synthetic data available, only a little research has focused on specifically treating sensitive attributes and generating synthetic data in a way that concentrates on protecting these selected attributes from inference attacks while keeping the data utility as high as possible. This can be achieved done by setting certain constraints when learning the model from the original data. Earlier work proposed a modification to extend the DataSynthesizer, an approach for synthetic data generation that uses Bayesian Networks to capture the underlying structures in the original data, to protect one sensitive attribute. In this paper, we investigate two different techniques for extending this approach to protect multiple attributes from inference and analyse the subsequent effects on the data utility.**

***Index Terms*—Synthetic Data, Bayesian Networks, Disclosure Risk Reduction**

I. Introduction

The amount of data being collected increases at a steady pace, often including personal or otherwise sensitive details, such as health or financial data. There is an enormous value and potential for analysis of this data, which often requires sharing and gathering data from various sources, e.g. when analysis is performed by knowledgeable third parties, or when parties want to collaboratively investigate distributed data sources. Often, this requires techniques for anonymising or otherwise treating the data, so that e.g. information on individuals can not be inferred anymore. The first step is often in removing identifying attributes from the dataset, e.g. names or attributes such as a social insurance number, in a step often referred to as pseudonymisation. However, even then, several forms of disclosure could happen from such a pseudonymised dataset. To discuss those, we need to first distinguish different types of attributes present in a dataset.

979-8-3503-2445-7/23/$31.00 ©2023 IEEE

Besides *identifying* attributes, a dataset usually also contains *quasi-identifiers* (sometimes also called indirect identifiers), i.e. attributes that cannot themselves uniquely identify a record but can become a unique identifier when combined with other Quasi-Identifiers. Further, *sensitive* attributes normally contain information that should not be disclosed, e.g. a medical diagnosis.

Identity disclosure happens when in a pseudonymised dataset an individual record can be re-identified. This can be achieved e.g. based on the values of quasi-identifying attributes, when linking them to another dataset that contains the same subset of quasi-identifiers (sometimes referred to as *attribute key*), and also identifying attributes; this is commonly referred to as *record linkage*. *Attribute disclosure*, on the other hand, means that for a record, for which the value of an attribute is not known, this value is inferred. Finally, *membership disclosure* reveals if a specific record was part of a given dataset.

To counter these types of disclosure, several techniques have been developed to protect data beyond pseudonymisation. Popular techniques include e.g. k-anonymity [1] or Differential Privacy [2]. In recent years, the usage of synthetic data has emerged as a popular alternative to these techniques. Synthetic data generation normally includes two steps: (i) learning a (statistical) model that *describes* the data, and (ii) a process of generating new data samples based on this model. The aim of synthetic data is to create a new population of records that preserve the overall characteristics of the initial dataset, e.g. the marginal and joint distributions and correlations between variables, without having a direct 1-1 relation to the initial data samples. Inevitably, all kinds of data protection techniques impact the quality and utility of the data treated. Synthetic data has become popular also because it has been shown to maintain high data utility for several tasks [3], such as classification [4], anomaly detection [5], or regression [6].

However, also synthetic data is subject to disclosure attacks. While identity disclosure might not be an issue for synthetic data, due to the lack of a 1:1 relation from original input data, the risk of attribute or membership disclosure remains. Recent work has addressed the reduction of attribute disclosure specifically, by modifying the learned data representation and specifically protecting a single sensitive attribute. In this paper, we are extending this approach, by developing two techniques

that modify the structure of a Bayesian Network that is used for describing the data. We evaluate our approach by measuring the success rate of an attribute disclosure attack on the original as well as on the modified network structure, and comparing how much the additional modification can reduce the attribute disclosure. Further, we measure the change in the utility of the synthetic data, by evaluating its effectiveness on predictive tasks associated with each of the datasets.

The remainder of this paper is organised as follows. Section II introduces related work on data synthetisation, as well as on inference attacks on those. Section III then describes in detail our enhanced approach for synthetisation of data using Bayesian networks before we evaluate the method on well-known benchmark datasets in Section IV. We provide conclusions and an outlook on future work in Section V.

II. Related Work

Various techniques for treating microdata before release or data analysis have been investigated. One of the most prominent approaches is K-anonymity [1], which sanitises data by e.g. generalisation or suppression of values, to ensure that at least k records have the same values for their quasi-identifiers. This approach primarily detects against identity disclosure, though later extensions such as l-diversity [7] also mitigate other forms of disclosure, such as attribute disclosure. Differential Privacy [2] is a mathematical definition and describes a property of an algorithm that publishes aggregate statistics of a database to limit the disclosure of individual records in the database.

Data synthetisation has recently emerged as an alternative approach to earlier data protection methods. In data synthetisation, there are generally two steps: (i) the original data is used to learn a model to describe the data, and then (ii) the model is used to generate new samples. While these new samples are based on the model, they do not have a 1:1 correspondence as being derived from one of the individuals. The data description step tries to find a model that best represents the complex relations between the attributes (sometimes also called predictors, input variables, or features) of the original dataset, as the base for then generating samples that retain the key properties of the original dataset.

One of the earliest usages of synthetic data was in the partial synthetic data approach by [8], where certain columns are generated synthetically. Multiple approaches have been proposed, mainly differing in the models that are used to describe the data [3], including e.g. decision trees [9], Gaussian copulas [10], GANs [11], or Bayesian Networks [12].

Several works have studied inference risks in synthetic data. Due to the lack of the 1:1 correspondence of individuals to the original data, *identity* disclosure is mostly considered to be not relevant for synthetic data. **Attribute disclosure** has been studied e.g. in [13], where an attribute inference attack generalising the Correct Attribution Probability (CAP) approach proposed in [14] was proposed. CAP measures the disclosure risk for a target record in the original dataset as the empirical probability of its target value given the value of a set of quasi-identifiers known by the attacker. CAP finds all synthetic records that coincide with the values of the quasi-identifiers of the target record and calculates the sensitive attribute from those. This attack does not yield results when there are no matching records in the synthetic dataset, and CAP thus is overly pessimistic on the disclosure. As an alternative, GCAP proposes to use a fixed-radius nearest neighbour classifier (FR-NN) to always infer a value. Stadler et al. [15] proposed a similar attack, but the attacker splits the synthetic dataset into two parts: a feature matrix containing the set of quasi-identifiers known by the attacker from the partial record and a vector containing the sensitive attribute values. Then, depending on the sensitive attribute type, the attacker trains a classification or regression model. The model then receives as input the partial record and returns a value for the sensitive attribute. Houssiau et al. [16] introduced an enhanced set of feature extraction techniques in comparison to the ones proposed in work [15]. Specifically, they used a feature map based on targeted counting queries as a feature extractor and demonstrated this technique outperforms the previously studied methods. In the context of local neighbourhood attacks, the authors proposed to extend the attack that finds the closest synthetic data record for attribute inference. Annamalai et al. [17] introduced an attack as a privacy game where the adversary's goal is to deduce the randomised secret attribute associated with a specific record. In order to measure individual-level leakage instead of population-level inferences, the secret attribute is randomised. This attack employs the concept of linear reconstruction attacks, with a particular emphasis on exploiting the fact that statistical queries on synthetic data should be as accurate as those on real data.

Regarding *membership disclosure*, [18] proposed a membership inference attack (MIA) based on the intuition that whenever a generative model overfits the data, then a Generative Adversarial Model (GAN) should be able to detect this overfitting. They assume that the attacker has a set of records that they suspect are in the training data, and knows the size of the training set. [19] proposed several different MIAs: one based on Monte Carlo integration that approximates small distance samples from the model, another one exclusively designed for variational auto-encoders as a reconstruction attack and a variation of the traditional MIA scenario which considers set membership. [20] proposed an MIA as a binary classification task motivated by the idea that a target record with a smaller distance to a synthetic record is more likely to be a member of the training set. [21] proposed an MIA based on the over-representation of GAN models. The intuition behind the attack is that there are regions where the proportion of training samples is higher and thus the likelihood that a sample falling in that region is a member is higher. [15] designed an MIA as a privacy game between an adversary and a challenger. The adversary's objective is to determine whether a record belongs to the original dataset, while the challenger acts as both the custodian of the data and the publisher of the synthetic dataset to the adversary. The attacker is assumed to have access to a reference dataset derived

from the same underlying distribution as the original data and performs the attack via training shadow models. [16] in their open source framework TAPAS implement shadow-modelling attacks as well as local neighbourhood attacks based on a distance metric and inference on synthetic data attacks, which rely on overfitting.

Synthetic data generation methods often treat all attributes the same when learning the data description. This means that when generating the data from a learned model, there will be no differentiation between the attributes, and the model will mimic the original correlation between them. The method introduced in [22], however, specifically allows to select **one sensitive attribute**, which will then be protected specifically. To this end, it will be separated from the quasi-identifiers and the other attributes, to eventually create less correlated synthetic data. Meanwhile, the correlation between the target attribute and the rest of the attributes is preserved. This will then ensure that the data utility does not suffer too much from synthetisation while reducing the risk of attribute disclosure for sensitive values. In this paper, we extend the method of [22] and implement the approach for an arbitrary number of sensitive attributes. Subsequently, the influence of the number of sensitive attributes on data utility and privacy was measured. The method focuses on tabular, structured data.

III. Protecting Multiple Sensitive Attributes

We specifically adapt the method of Ping et al. [12] called DataSynthesizer[1], which generates synthetic data with Bayesian networks. Bayesian networks are graphical models that use directed acyclic graphs to model conditional dependence. As a first step of data synthetisation, the original data is described by such a Bayesian network. This is done by constructing a network that learns correlations between the attributes, as well as other data properties. Finding the optimal structure of the Bayesian network is NP-complete [23]. The original DataSynthesizer uses a greedy heuristic algorithm called GreedyBayes, which is based on PrivBayes introduced by [24].

As the algorithm used does not scale to datasets with higher dimensions, a custom genetic algorithm has been introduced in [22] for constructing the network. Genetic algorithms try to imitate natural selection, where the fittest individuals survive and can reproduce. Each individual has a set of properties, the so-called chromosome, which can be altered during the process. All possible individuals form the population. A fitness score is calculated for each individual, and the individuals with the highest score are selected to pass their genes to the next generation. Two individuals will then exchange their genes to create an offspring. The chromosomes might also be mutated with some probability.

In our application of genetic algorithms, the individuals represent a possible Bayesian network. Each individual consists of an ordering chromosome, which represents the order of attributes being added to the network, and a connectivity chromosome, which indicates the parents for each attribute. Our goal is to build a Bayesian network where all the sensitive attributes are separated from the non-sensitive attributes and quasi-identifiers, i.e. sensitive attributes cannot have non-sensitive attributes or quasi-intensifiers as their parents, and vice versa. This will result in a network that generates data where the sensitive attributes are less correlated with any other attribute that might be available to an attacker, thus creating data with a reduced risk for attribute disclosure.

TABLE I: Dataset dimensions

Dataset	# Attributes	# Instances	# Classes
Caesarian	6	80	2
Contraceptive Method Choice	10	1473	3
Adult Census	15	48842	2

In this paper, we introduce and evaluate two techniques for protecting multiple sensitive attributes. In the first approach, sensitive attributes can only have the target variable as a parent. The second approach allows for sensitive attributes to have other sensitive attributes, the target variable or both as their parents. The basic structures of these two approaches can be seen in Figure 1.

To integrate these two approaches into the genetic algorithm, some constraints for the ordering chromosomes need to be set. For both approaches, the target attribute needs to be fixed as the first component of the ordering chromosome, as done in [22]. Then all sensitive attributes are added before any other attributes can be appended. In the case of the first approach, the order of the sensitive attributes will stay the same. However, for the second approach, the ordering of the sensitive attributes will be altered during the process, to find the best possible network. Note that the order will only be changed within the blocks of sensitive and other attributes, respectively, and not over all attributes. This can be seen in Figure 2, which displays the ordering chromosome. Here t denotes the target variable, s_i are the sensitive attributes, where $i = 1, ..., m$, m denoting the number of sensitive attributes. Lastly, f_j with $j = 1, ..., n$ denotes all remaining attributes.

The method will be demonstrated using three publicly available datasets: Caesarian [2], Contraceptive Method Choice [3] and Adult Census [4]). These datasets were chosen as they represent different sizes in terms of the number of attributes and instances. The characteristics of each dataset can be seen in Table I.

As part of data preparation, some nominal attributes needed to be label encoded. Each dataset was then randomly split into training and test data (75:25% split). After that, the training data was synthesised. All training and test sets were then scaled by removing the mean and scaling to unit variance.

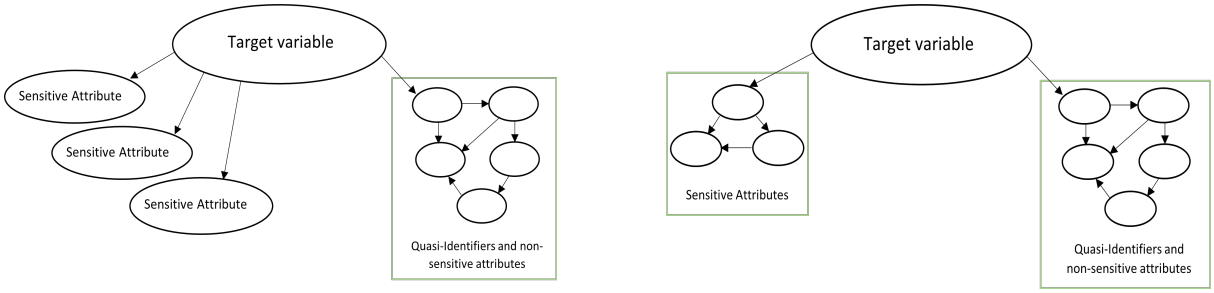
All three datasets have associated classification tasks. On these, we applied five machine learning algorithms: k-Nearest-

[1]Code available at https://github.com/DataResponsibly/DataSynthesizer

[2]https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset

[3]https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

[4]https://archive.ics.uci.edu/ml/datasets/Adult



(a) Method 1 **(b)** Method 2

Fig. 1: Network structures for sensitive and other attributes (quasi-identifiers and non-sensitive attributes

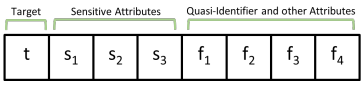


Fig. 2: Ordering Chromosome

TABLE II: Attribute overview

Dataset	target attribute	sensitive attribute	attribute key
Caesarian	caesarian	heart problem, delivery number, blood pressure	delivery time, age
Contraceptive Method	contraceptive method used	number of children ever born, wife's education, wife's age	wife now working?, wife's religion, husband's education, husband's occupation
Adult Census	salary	relationship, education, marital status	work class, race, sex, age, occupation

Neighbors (kNN), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machine (SVC), using their respective sklearn implementation[5]. For model training, the default parameter settings of the algorithms were used. To measure the effect of the number of sensitive attributes on the utility, one, two or three attributes were set as such sensitive attributes for each dataset. These three sensitive attributes were selected based on the feature importance and sensitivity of the attribute. Table II gives an overview of target and sensitive attributes as well as the attribute keys used for the disclosure risk assessment in Section IV-B.

To measure the effect of the different settings on the data utility, the mean accuracy of ten random splits into training and test data was computed for each dataset-algorithm combination, with $p = \{1, 2, 3, 4\}$, where p is the maximum number of parents per node. Furthermore, the different outcomes of the two methods (with and without sensitive attribute interaction) were evaluated.

The disclosure risk was assessed by setting the sensitive attributes as target variables and the quasi-identifiers as input attributes (*attribute key*). Then the same machine learning algorithms as before were used to predict the sensitive attributes. In addition, the Generalized Correct Attribution Probability (GCAP), introduced by [13], was computed. GCAP was specifically designed to measure the risk of disclosure for sensitive data using distances. For these experiments, all synthetic data was generated using Bayesian networks with a maximum of three parents per node.

[5]https://scikit-learn.org/stable/supervised_learning.html

With the approach presented above, the disclosure risk is assumed to decrease, meaning the accuracies for predicting sensitive attributes are expected to be much lower than for the original data and synthetic data generated by the standard approach of the DataSynthesizer.

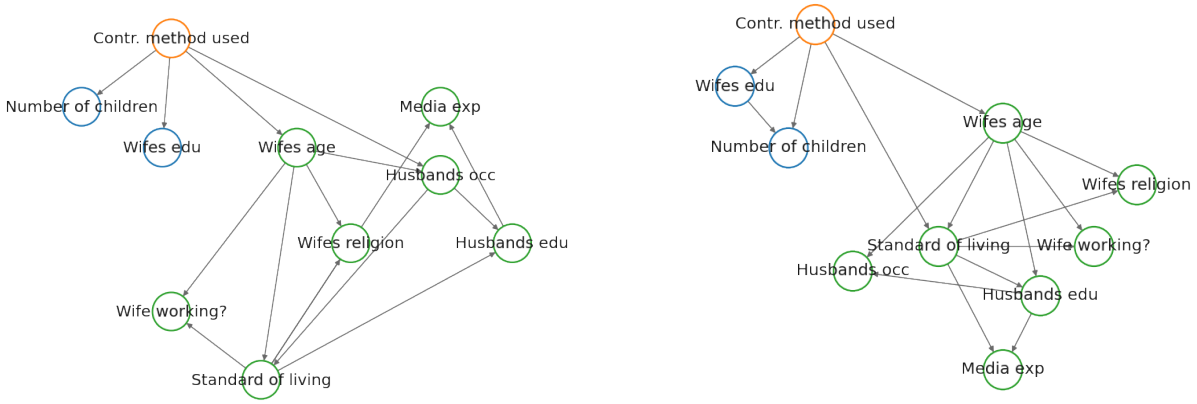
IV. Evaluation

To first demonstrate the process of building the Bayesian network and the resulting synthetic data generated by the network with the new approach, the Contraceptive Method Choice dataset was chosen. Figure 3 shows the different networks learned for both approaches, with $p = 2$. The target variable is coloured in orange, blue nodes show sensitive attributes, and green coloured nodes represent non-sensitive attributes. While for Figure 3a sensitive attributes cannot be parents of other sensitive attributes (method 1), this constraint is relaxed for Figure 3b (method 2). Once the Bayesian Network is learned and the synthetic data is generated, the effect of the custom networks on attribute correlations can be explored.

Figure 4 shows the correlations between attributes for the original versus synthetic data protecting zero, one, two, or three sensitive attributes. While correlations between sensitive and non-sensitive attributes are still high for the synthetic data without sensitive attributes defined, they are reduced once the attributes are set to be sensitive. For method 1, the only correlations preserved for the sensitive attributes are between each sensitive attribute and the target variable. For method 2, on the other hand, correlations between all the sensitive attributes are maintained as well. This can be seen in Figure 5, where "number of children ever born" and "Wife's education" were chosen as sensitive attributes (each marked with a black frame). Figure 5a shows the correlation for the synthetic data without sensitive attributes, while Figure 5b and Figure 5c display the correlation for the synthetic data created under the conditions of method 1 and 2, respectively. In addition to preserving the correlations with the target attribute ("Contraceptive method used"), the correlation between the two sensitive attributes, albeit it is a low correlation, was also kept at the original level.

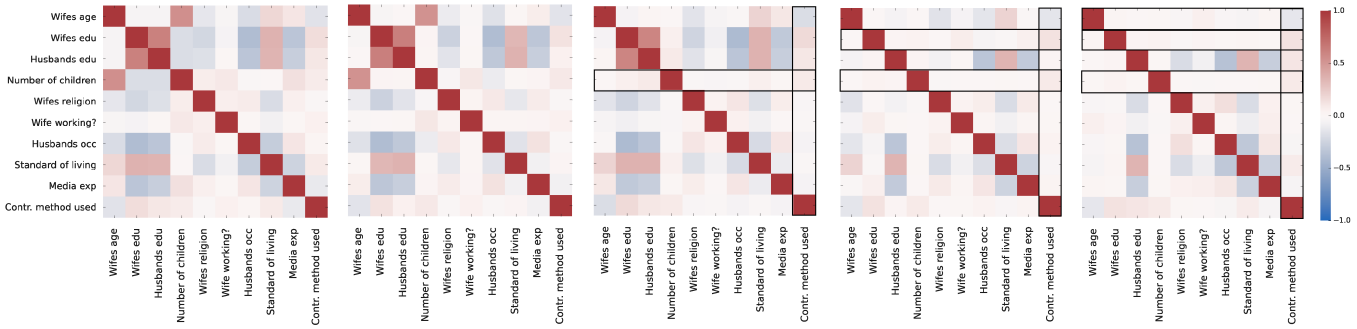
A. Data Utility Assessment

One of the main goals for synthetic data generation is to obtain datasets that can be published without privacy concerns



(a) Method 1 **(b)** Method 2

Fig. 3: Bayesian Networks for Contraceptive Method Choice data. Orange: target attribute; blue: sensitive attributes; green: non-sensitive attributes.



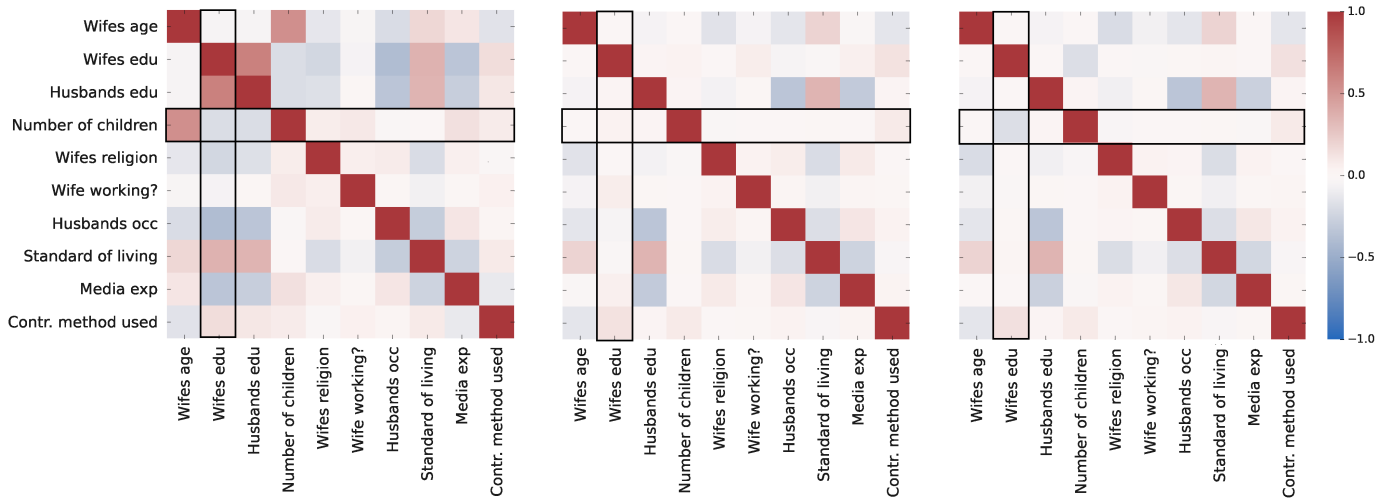
(a) Original data **(b)** No sensitive attribute selected **(c)** 1 sensitive attribute selected **(d)** 2 sensitive attributes selected **(e)** 3 sensitive attributes selected

Fig. 4: Correlations for different numbers of sensitive attributes (method 1)

while keeping the utility loss at a minimum. One option to measure utility is to measure the effectiveness that can be achieved when using the dataset to learn a downstream machine learning task, e.g. to measure the classification accuracy. The utility loss can be estimated by calculating and comparing the accuracies achieved when training the respective models on the original and synthetic datasets. Table III as well as Figures 6 to 8 show the results for each dataset as mean accuracy over all machine learning models. In addition, Table III also shows the rounded differences between the accuracies of the synthetic data to the original data in parenthesis. The highest score per parameter setting and dataset is coloured red. Note that for zero and one sensitive attributes, the results for methods 1 and 2 will be the same, since there are not enough sensitive attributes to have any interaction between them, and the resulting Bayesian Networks are thus identical.

When looking at the results for the Caesarian data in Table III and Figure 6, the accuracy scores seem to be quite unstable with no noticeable trends or patterns, which is likely caused by its low number of input attributes and instances. The accuracies decrease with a higher number of sensitive attributes for method 2. For method 1 however, the accuracies increase with a higher number of sensitive attributes, when the number of parents $p = \{2, 3, 4\}$. Figure 6 also shows that accuracies for method 1 are higher than for method 2. Overall, the results imply that the method's effectiveness suffers from data with a small number of input variables. Note that Table III does not have a value for three sensitive attributes for method 1 when the number of parents $p = 3$, since for that case, there cannot be a network with one node having more than two parents, as the dataset has only six attributes in total. The same holds for two and three sensitive attributes when the number of parents $p = 4$.

The results for the contraceptive choice data show that with an increasing number of parents p, the accuracy for synthetic data increases as well. This can be seen in Figure 7. While the values for $p = \{2, 3, 4\}$ and method 1 seem to be very similar, the scores for method 2 increase with an increasing number of sensitive attributes and a maximum number of parents (p). This, however, is not the case for the networks



(a) original **(b)** Method 1 **(c)** Method 2

Fig. 5: Correlations for two sensitive attributes. Under method 1, (b), the correlation between "number of children" and "wifes edu" is removed. However, using method 2 (c), the correlation from the original data (a) is preserved

TABLE III: Data Utility shown as Accuracy Scores (mean over all ML methods); p: number of parents

			#sensitive Attr.	**Method 1**			**Method 2**	
Dataset	p	original	0	1	2	3	2	3
Caesarian	1	64.10	63.90 (-0.20)	61.40 (-2.70)	62.70 (-1.40)	63.70 (-0.40)	62.65 (-1.45)	58.70 (-5.40)
	2	64.10	63.45 (-0.65)	63.60 (-0.50)	63.65 (-0.45)	64.05 (-0.05)	63.60 (-0.50)	63.50 (-0.60)
	3	64.10	62.05 (-2.05)	63.65 (-0.45)	64.00 (-0.10)	-	63.40 (-0.70)	61.75 (-2.35)
	4	64.10	61.40 (-2.70)	65.05 (0.95)	-	-	-	-
Contraceptive Choice	1	50.78	42.15 (-8.62)	44.70 (-6.08)	47.64 (-3.14)	47.47 (-3.30)	46.60 (-4.17)	44.15 (-6.63)
	2	50.78	49.20 (-1.58)	46.94 (-3.83)	47.25 (-3.52)	47.72 (-3.06)	47.64 (-3.14)	49.42 (-1.36)
	3	50.78	50.29 (-0.48)	47.82 (-2.95)	47.57 (-3.20)	48.15 (-2.62)	48.54 (-2.24)	50.25 (-0.52)
	4	50.78	49.80 (-0.98)	47.85 (-2.93)	47.33 (-3.44)	48.23 (-2.55)	48.72 (-2.05)	50.88 (0.11)
Adult Census	1	83.08	79.92 (-3.16)	78.64 (-4.45)	81.04 (-2.04)	80.49 (-2.59)	80.94 (-2.14)	82.28 (-0.80)
	2	83.08	80.45 (-2.63)	80.20 (-2.88)	81.37 (-1.71)	80.74 (-2.34)	81.48 (-1.60)	82.41 (-0.68)
	3	83.08	81.82 (-1.26)	81.47 (-1.61)	81.89 (-1.19)	80.85 (-2.23)	81.73 (-1.35)	82.45 (-0.63)
	4	83.08	82.89 (-0.19)	81.43 (-1.65)	82.02 (-1.06)	80.91 (-2.17)	82.06 (-1.02)	82.64 (-0.44)

with $p = 1$. Although it is expected that preserving the correlations between attributes increases the accuracies for a higher number of sensitive attributes in method 2, the results show the opposite. This is caused by the network structures learned: For networks with three sensitive attributes, there will only ever be one sensitive attribute directly connected to the target variable. The remaining sensitive attributes do not have a direct edge to the target variable, and will therefore fail to preserve most of the information needed to predict it. This showcases that for a large enough dataset and adequate parameter p, the two methods yield the expected results. By preserving more correlations with method 2, the accuracies score higher than when disregarding the correlations between sensitive attributes in method 1.

When looking at the results for the Adult Census data (Figure 8), we can observe a pattern that with a growing number of parents p, the accuracy scores seem to become more stable. Additionally, for all values for p, there is always quite a big difference between methods 1 and 2 when using three sensitive attributes. The difference between two sensitive attributes, however, is minimal. For $p = 4$, the synthetic data without sensitive attributes scores higher than all four data sets with sensitive attributes. For the other values of p, this is not the case.

Generally, experiments show that the higher the maximum number of parents (parameter p), the higher the accuracy and, therefore, the lower the utility loss. Accuracy is especially low for synthetic data generated with a network of degree one. For the Contraceptive Method dataset, method 2 performs worse with three sensitive attributes than with two. The dimension of the dataset also seems to affect the outcome. While the results for the Contraceptive Method data and the Adult Census data seemed to be reasonable, the small size of the Caesarian data is assumed to cause subpar and unstable results.

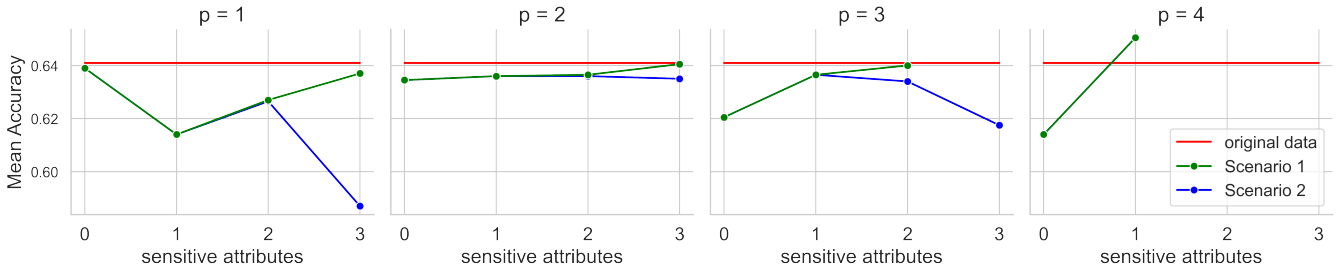


Fig. 6: Accuracy for Caesarian data

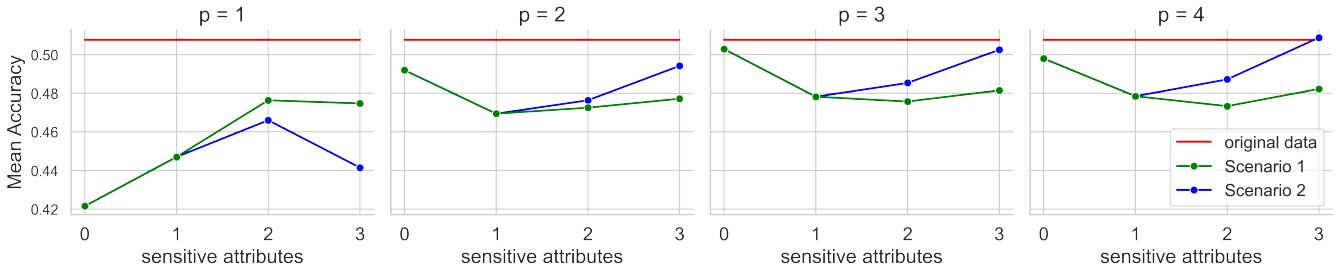


Fig. 7: Accuracy for Contraceptive Method data

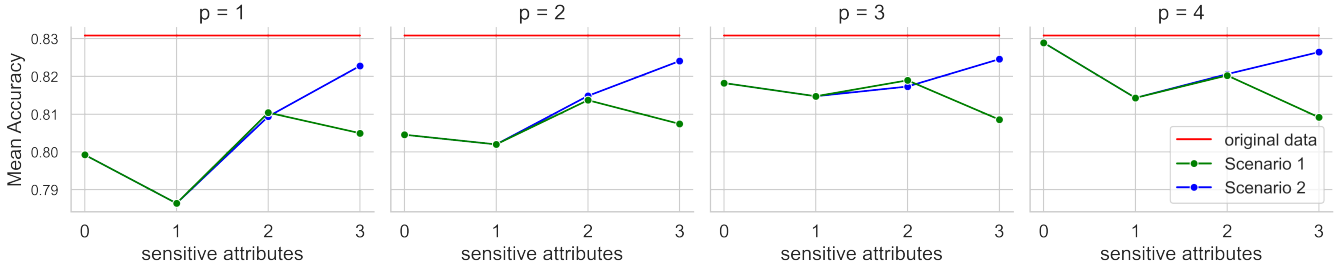


Fig. 8: Accuracy for Adult Census data

B. Disclosure Risk Assessment

The results for disclosure risk experiments can be found in Tables IV to VI. Here, a lower (inference attack) accuracy implies a lower risk for attribute disclosure, which is the preferred outcome. The lowest values per dataset-inference method are marked in red. Note that for these experiments, it is assumed that the adversary does not know any of the sensitive attributes, i.e., the attribute key cannot contain any sensitive attributes.

The results for Caesarian data (Table IV) show that the average inference attack accuracies of the original data are always highest. The accuracy for synthetic data without sensitive attributes is always higher than for the newly introduced approach, which was the intended result. In most cases, the accuracy is lowest for the synthetic data with sensitive attribute interaction (method 2), with the exception of the attribute ”Blood Pressure”, where the score for method 2 is higher than the score for method 1. In any case, the decrease in accuracy from original to synthetic data implies a substantially reduced risk for attribute disclosure, with an even lower risk for data generated using the proposed approach.

Similar to the results of the Caesarian data, the Contraceptive Method data also reaches the highest inference attack success for the original data, and the lowest for synthetic data with sensitive attributes (Table V). Here, the differences between the two methods are minimal. Although for some attributes, like ”Number of children ever born”, the decrease in accuracy might not be as substantial as with the other dataset, there is still a noticeable difference to the synthetic data generated without sensitive attributes.

Table VI displays the results for the Adult Census data. Again, the inference risk on the original data is the highest, followed by the synthetic data without sensitive attributes. Synthetic data generated by the new method has the lowest scores, with similar values for the two methods.

To summarize the above findings, the synthetic data generated with sensitive attributes resulted in the lowest inference success scores, meaning that synthetic data generated by the new method decreases the risk for attribute disclosure even further than synthetic data where no sensitive attributes were selected. Assuming an adversary knew a subject’s values for all quasi-identifiers, the probability of their predictions for the

TABLE IV: Caesarian: Attribute Disclosure Risk

		GCAP	LR	NB	RF	SVC	kNN	Ave
Blood Pressure	original	73.80	53.75	53.75	73.75	57.50	58.75	61.88
	synthetic (no sensitive var.)	59.83	50.58	50.58	62.25	53.75	51.33	54.72
	synthetic (Method 1)	38.80	51.25	50.00	40.00	50.00	51.25	46.88
	synthetic (Method 2)	43.80	50.00	50.00	46.25	50.00	46.25	47.72
Delivery number	original	76.20	51.25	51.25	76.25	61.25	66.25	63.74
	synthetic (no sensitive var.)	61.01	52.17	53.00	64.42	57.17	57.00	57.46
	synthetic (Method 1)	42.50	48.75	51.25	55.00	51.25	50.00	49.79
	synthetic (Method 2)	41.20	48.75	46.25	41.25	48.75	45.00	45.20
Heart Problem	original	87.50	61.25	61.25	87.50	70.00	73.75	73.54
	synthetic (no sensitive var.)	75.91	64.17	64.58	74.08	65.25	68.33	68.72
	synthetic (Method 1)	53.80	63.75	58.75	60.00	65.00	58.75	60.01
	synthetic (Method 2)	60.00	62.50	56.25	57.50	60.00	53.75	58.33

TABLE V: Contraceptive Method Choice: Attribute Disclosure Risk

		GCAP	LR	NB	RF	SVC	kNN	Ave
Number of children ever born	original	24.50	20.37	4.62	24.51	22.81	19.89	19.45
	synthetic (no sensitive var.)	19.42	19.24	19.15	19.35	19.57	16.71	18.91
	synthetic (Method 1)	17.83	18.49	17.17	17.71	18.12	15.24	17.43
	synthetic (Method 2)	17.79	18.72	17.95	17.73	18.44	15.79	17.74
Wife's age	original	10.90	7.60	1.49	10.93	9.71	7.26	7.98
	synthetic (no sensitive var.)	6.50	6.50	5.87	6.56	6.68	4.77	6.15
	synthetic (Method 1)	4.19	5.17	4.14	4.33	4.96	3.46	4.38
	synthetic (Method 2)	4.24	5.09	4.11	4.33	4.89	3.52	4.36
Wife's education	original	56.10	55.19	52.21	56.14	55.06	52.14	54.47
	synthetic (no sensitive var.)	54.57	54.79	51.81	54.42	54.79	48.33	53.12
	synthetic (Method 1)	41.64	39.67	41.53	41.19	40.98	30.78	39.30
	synthetic (Method 2)	42.25	40.26	42.00	41.75	42.03	30.83	39.85

TABLE VI: Adult Census: Attribute Disclosure Risk

		GCAP	LR	NB	RF	SVC	kNN	Ave
education	original	51.40	33.67	32.91	51.42	36.95	41.55	41.32
	synthetic (no sensitive var.)	36.92	33.69	32.60	36.69	36.56	32.13	34.77
	synthetic (Method 1)	28.60	32.31	32.31	28.22	32.31	24.34	29.68
	synthetic (Method 2)	28.70	32.31	32.32	28.29	32.31	24.54	29.74
marital-status	original	73.10	67.35	65.71	73.12	67.97	67.23	69.08
	synthetic (no sensitive var.)	64.90	67.36	65.31	65.26	67.62	63.69	65.69
	synthetic (Method 1)	52.10	56.00	57.14	51.61	60.71	44.92	53.74
	synthetic (Method 2)	52.18	56.19	56.80	52.03	59.90	45.45	53.76
relationship	original	67.30	59.77	46.05	67.28	60.60	61.40	60.40
	synthetic (no sensitive var.)	57.78	59.72	36.72	57.95	60.38	56.71	54.87
	synthetic (Method 1)	40.82	41.58	43.37	40.50	41.03	37.84	40.86
	synthetic (Method 2)	40.74	42.15	43.61	40.54	41.39	38.24	41.11

sensitive attributes being correct is therefore much higher for the original and the synthetic data generated in the standard way without specifically protecting sensitive attributes. This shows that even when adding more than one sensitive attribute, the method still produces the desired outcome. Furthermore, the difference between methods 1 and 2 is notably small for all scores. Nevertheless, the effectiveness of the introduced approach is assumed to also depend on the sensitive attributes chosen and their initial correlations with quasi-identifiers.

Until now, it was assumed that a sensitive attribute cannot be part of the attribute key an adversary knows. The Contraceptive Method dataset is used to demonstrate the effect of including the sensitive attribute "Wife's age" in the attribute key. The experiment results are shown in Table VII. Here, Setting 1 uses the original attribute key for the attribute disclosure risk assessment, as seen before and listed in Table II, while the attribute key for Setting 2 includes the sensitive attribute "Wife's age".

The inference attack accuracies for the other two sensitive attributes not included in the attribute key, namely "Number of children ever born" and "Wife's education", are expected to increase for the original dataset, the commonly generated synthetic data and the synthetic data generated with method 2, but should stay the same for method 1, since the sensitive

TABLE VII: Including a Sensitive attribute in the Attribute Key

		Setting 1	Setting 2
Number of children ever born	original	19.45	37.03
	synthetic (no sensitive var.)	18.91	26.63
	synthetic (Method 1)	17.43	16.59
	synthetic (Method 2)	17.74	23.21
Wife's education	original	54.47	62.45
	synthetic (no sensitive var.)	53.12	55.73
	synthetic (Method 1)	39.30	36.53
	synthetic (Method 2)	39.85	38.85

attributes for these networks are uncorrelated. This is exactly what can be observed for the attribute "Number of children born": All accuracies increase for Setting 2, except for the accuracy in method 1, where we can even mark a slight decrease. For the attribute "Wife's education" on the other hand, the above assumption does not hold. While the original and synthetic data without sensitive attributes score higher than before, methods 1 and 2 achieve slightly smaller values in Setting 2. A possible explanation as to why the two attributes yield different results is that their correlation with "Wife's age" differs. When looking at the heat-map of the original data in Figure 4, it can be seen that the correlation between "Wife's age" and "Number of children ever born" is a lot higher (0.54) than the correlation between "Wife's age" and "Wife's education" (-0.048). Thus, when the networks are built, the two attributes "Wife's age" and "Wife's education" are not connected by a direct edge in most cases. Therefore, Setting 2 does not significantly influence either of the two methods.

V. Conclusions and Future Work

In this paper, we presented an approach for protecting multiple sensitive attributes from attribute inference attacks when generating synthetic data. To this end, we extended a previous method that is limited to protecting *one* sensitive attribute to be able to consider multiple attributes. We designed two different methods, which vary on whether they preserve correlations within the sensitive attributes themselves.

Our evaluation shows that it is possible to protect multiple attributes at the same time, with limited loss in data utility, and our new method can thus be successfully applied. We further observed that a learned Bayesian network that allows interaction between sensitive attributes will generate data with higher utility than a network that does not allow sensitive attribute interaction. This means that if it can be assumed that an adversary does not have any information about any of the sensitive attributes, it is advised to opt for a network built under the conditions of method 2, which preserves the correlations by allowing interactions in the network between those attributes. Using networks with a high value for p and more sensitive attributes will cause almost no utility loss. In two cases, the accuracy of the synthetic data was even higher than that of the original data.

The risk for attribute disclosure on sensitive data attributes can be reduced even more by using the introduced approach instead of conventional synthetisation methods. If the possibility of an adversary having access to a sensitive attribute cannot be excluded, using method 1 might be a safer option. Generally, how well the new approach performs also seems to depend on the attribute key, attributes selected as sensitive and the chosen method.

In future, we will investigate how protecting sensitive attributes from disclosure can be performed in other data synthetisation methods, such as the synthetic data vault, which is based on Gaussian Copulas, or methods based on neural networks. Further, we will investigate how to port our findings to other data domains.

Acknowledgement

This work received funding from the European Union's Horizon Europe Research and Innovation programme under grant agreement No 101095530 (SYNTHEMA). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union can not be held responsible for them.

SBA Research (SBA-K1) is a COMET Center within the COMET - Competence Centers for Excellent Technologies Programme and funded by BMK, BMAW, and the federal state of Vienna. The COMET Programme is managed by FFG.

References

[1] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, Dec. 2001. [Online]. Available: http://ieeexplore.ieee.org/document/971193/

[2] C. Dwork, "Differential Privacy," in *Automata, Languages and Programming*, ser. Lecture Notes in Computer Science, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds., vol. 4052. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12.

[3] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, vol. 493, pp. 28–45, Jul. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231222004349

[4] M. Hittmeir, A. Ekelhart, and R. Mayer, "On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks," in *International Conference on Availability, Reliability and Security*, ser. ARES. Canterbury, United Kingdom: ACM, Aug. 2019.

[5] R. Mayer, M. Hittmeir, and A. Ekelhart, "Privacy-preserving Anomaly Detection using Synthetic Data," in *Data and Applications Security and Privacy XXXIV*, ser. DBSec. Regensburg, Germany: Springer International Publishing, 2020, pp. 195–207.

[6] M. Hittmeir, A. Ekelhart, and R. Mayer, "Utility and Privacy Assessments of Synthetic Data for Regression Tasks," in *IEEE International Conference on Big Data (IEEE BigData 2019)*, Los Angeles, CA, United States, Dec. 2019, pp. 5763–5772.

[7] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*L* -diversity: Privacy beyond *k* -anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 3, Mar. 2007. [Online]. Available: https://dl.acm.org/doi/10.1145/1217299.1217302

[8] D. B. Rubin, Ed., *Multiple Imputation for Nonresponse in Surveys*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, Jun. 1987. [Online]. Available: http://doi.wiley.com/10.1002/9780470316696

[9] B. Nowok, G. M. Raab, and C. Dibben, "synthpop: Bespoke Creation of Synthetic Data in R," *Journal of Statistical Software*, vol. 74, no. 11, Oct. 2016.

[10] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," in *IEEE International Conference on Data Science and Advanced Analytics*, ser. DSAA. Montreal, QC, Canada: IEEE, Oct. 2016, pp. 399–410.

[11] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data Using Conditional GAN," in *International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[12] H. Ping, J. Stoyanovich, and B. Howe, "DataSynthesizer: Privacy-Preserving Synthetic Datasets," in *International Conference on Scientific and Statistical Database Management*, ser. SSDBM. Chicago IL USA: ACM, Jun. 2017, pp. 1–5.

[13] M. Hittmeir, R. Mayer, and A. Ekelhart, "A Baseline for Attribute Disclosure Risk in Synthetic Data," in *ACM Conference on Data and Application Security and Privacy*, ser. CODASPY. New Orleans LA USA: ACM, Mar. 2020, pp. 133–143.

[14] J. Taub, M. Elliot, M. Pampaka, and D. Smith, "Differential Correct Attribution Probability for Synthetic Data: An Exploration," in *Privacy in Statistical Databases*, ser. Lecture Notes in Computer Science, J. Domingo-Ferrer and F. Montes, Eds. Valencia, Spain: Springer International Publishing, 2018, pp. 122–137.

[15] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic Data â€“ Anonymisation Groundhog Day," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 1451–1468. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/stadler

[16] F. Houssiau, J. Jordon, S. N. Cohen, O. Daniel, A. Elliott, J. Geddes, C. Mole, C. Rangel-Smith, and L. Szpruch, "TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data," Nov. 2022, arXiv:2211.06550 [cs]. [Online]. Available: http://arxiv.org/abs/2211.06550

[17] M. S. M. S. Annamalai, A. Gadotti, and L. Rocher, "A Linear Reconstruction Approach for Attribute Inference Attacks against Synthetic Data," Jun. 2023, arXiv:2301.10053 [cs]. [Online]. Available: http://arxiv.org/abs/2301.10053

[18] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "LOGAN: Membership Inference Attacks Against Generative Models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 1, pp. 133–152, Jan. 2019. [Online]. Available: https://petsymposium.org/popets/2019/popets-2019-0008.php

[19] B. Hilprecht, M. Härterich, and D. Bernau, "Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 4, pp. 232–249, Oct. 2019. [Online]. Available: https://petsymposium.org/popets/2019/popets-2019-0067.php

[20] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Virtual Event USA: ACM, Oct. 2020, pp. 343–362. [Online]. Available: https://dl.acm.org/doi/10.1145/3372297.3417238

[21] H. Hu and J. Pang, "Membership Inference Attacks against GANs by Leveraging Over-representation Regions," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Virtual Event Republic of Korea: ACM, Nov. 2021, pp. 2387–2389. [Online]. Available: https://dl.acm.org/doi/10.1145/3460120.3485338

[22] M. Hittmeir, R. Mayer, and A. Ekelhart, "Efficient Bayesian Network Construction for Increased Privacy on Synthetic Data," in *2022 IEEE International Conference on Big Data (Big Data)*. Osaka, Japan: IEEE Computer Society, Dec. 2022.

[23] D. M. Chickering, "Learning Bayesian Networks is NP-Complete," in *Learning from Data*. New York, NY: Springer New York, 1996, vol. 112, pp. 121–130, series Title: Lecture Notes in Statistics. [Online]. Available: http://link.springer.com/10.1007/978-1-4612-2404-4_12

[24] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: Private Data Release via Bayesian Networks," *ACM Transactions on Database Systems*, vol. 42, no. 4, pp. 1–41, Dec. 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3134428