Empirical Evaluation of Synthetic Data Created by Generative Models via Attribute Inference Attack

Saloni Kwatra(✉) and Vicenç Torra

Department of Computing Science, Umeå University, Umeå, Sweden
{salonik,vtorra}@cs.umu.se

Abstract. The disclosure risk of synthetic/artificial data is still being determined. Studies show that synthetic data generation techniques generate similar data to the original data and sometimes even the exact original data. Therefore, publishing synthetic datasets can endanger the privacy of users. In our work, we study the synthetic data generated from different synthetic data generation techniques, including the most recent diffusion models. We perform a disclosure risk assessment of synthetic datasets via an attribute inference attack, in which an attacker has access to a subset of publicly available features and at least one synthesized dataset, and the aim is to infer the sensitive features unknown to the attacker. We also compute the predictive accuracy and F1 score of the random forest classifier trained on several synthetic datasets. For sensitive categorical features, we show that Attribute Inference Attack is not highly feasible or successful. In contrast, for continuous attributes, we can have an approximate inference. This holds true for the synthetic datasets derived from Diffusion models, GANs, and DPGANs, which shows that we can only have approximated Attribute Inference, not the exact Attribute Inference.

Keywords: Generative Adversarial Networks · Privacy · Attribute Inference Attack · Differentially Private Generative Adversarial Networks · Diffusion Models

1 Introduction

Synthetic datasets do not have actual user data. Therefore, they are in compliance with the General Data Protection Regulation (GDPR) or similar regulations. Hence, synthetic data can be made publicly available. However, not containing the actual data does not make synthetic datasets private. Therefore, not all the synthetic data is publishable. Today, there are many techniques available for creating synthetic datasets, from the Generative Adversarial Networks (GANs) to the most recent Diffusion Models. In a recent work, the authors

This study was partially funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

F. Bieker et al. (Eds.): Privacy and Identity 2023, IFIP AICT 695, pp. 282–291, 2024.
https://doi.org/10.1007/978-3-031-57978-3_18

showed a data reconstruction attack against Principal Component Analysis (PCA) [1]. The success of their proposed attack entirely relies on the quality of the synthetic data obtained using generative models. Hence, it is essential to measure the privacy of synthetic datasets to know the answer to the question, "Can synthetic data be made publicly available?" Stadler *et al.* [2] quantified the privacy gain of synthetic data publishing by evaluating if the synthetic data and the differentially private synthetic data mitigate the risk of linkability by formalizing linkability as membership inference. It was concluded in [2] that the synthetic data does not provide a better tradeoff between privacy and utility in comparison with the anonymization techniques like k-anonymity [3–6]. A better tradeoff is the one when the increase in privacy gain does not cause much deterioration in the utility. But, in the case of synthetic datasets, especially when Differential Privacy (DP) [7] is introduced in the sythetic datasets, the utility degrades.

One of the major drawbacks of these synthetic data generation techniques is from the context of usable privacy [15]. It is difficult to predict what information a synthetic dataset will preserve and what information will be perturbed and to which extent. In contrast, for privacy models like k anonymity, we know exactly which attributes are being perturbed and to which extent (higher k implies higher perturbation). In our opinion, this is a major roadblock in the case of synthetic datasets for their acceptability into practical use cases, where it is important to quantify privacy gain. In this work, we make an Attribute Inference Attack against synthetic datasets to quantify the privacy of synthetic datasets. We show that inferring an attribute is not so feasible in the case of synthetic datasets, which means that exact inference is difficult, but the attacker can infer values in the vicinity of the true values using the synthetic data. It has been shown that the synthetic datasets, especially the DP-synthetic ones, lower the membership attack accuracy, as shown in [2]. In our work, we show that an approximate attribute inference is possible in the case of continuous attributes like age and estimated salary. But, it is harder to infer the exact value of categorical attributes like sex, disease, or geographical location of a person. Our main contributions of this work are as follows:

– We assess the disclosure risk and utility of the data generated generative models. To the best of our knowledge, this is the first work that analyses the tradeoff between the utility and the privacy for diffusion models and compares with other generative models.
– We did the disclosure risk assessment of synthetic datasets via an attribute inference attack.
– We show that it is difficult to infer the exact value of the attribute. But, the approximate inference is possible for continuous attributes.

The rest of this paper is organized as follows. Section 2 reviews some basic concepts needed in understanding the paper. In particular, we discuss algorithms to generate the synthetic data, which includes Generative Networks and their variants, Diffusion Models, and also Attribute Inference Attack. Our attack methodology is described in Sect. 3. Section 4 describes the datasets and the attack settings used in our experiment. In Sect. 5, we provide and discuss the results. Section 6 gives the conclusion and future directions.

2 Preliminaries

In this section, we explain all the techniques we use to generate the synthetic data to make an Attribute Inference Attack.

2.1 Generative Models for Tabular Forms and Their Variants

Generative Models [8] use a probabilistic model to learn how a dataset is generated and create synthetic datasets by sampling from a learned distribution. In GANs, two neural networks are pitted against each other. The generator tries to befool the discriminator by producing real-looking data, while the discriminator tries to differentiate between real and synthetic data. They have the capability to generate any amount of data records even when they are trained on fewer data records. We work with tabular data. Therefore, one of the techniques we use to generate the synthetic data is Conditional Tabular Generative Adversarial Network **(CTGAN)** [9]. In CTGAN, two changes are made to solve the issues faced by GANs when applied to generate tabular data. CTGAN proposed using mode-specific normalization to avoid losing information for the continuous data. To handle an imbalance in discrete columns, CTGANs designed a conditional vector *cond*, and a training-by-sampling technique. The conditional generator can generate synthetic rows conditioned on one of the discrete columns. Using training-by-sampling, the *cond* and training data are sampled according to the log frequency of each category. Therefore, CTGAN can explore all possible discrete values.

We also use differentially private versions of GANs in order to evaluate the success of attribute inference attack. In Differentially Private Conditional Tabular Generative Adversarial Network **(DP-CTGAN)** [12], differentially private stochastic gradient descent is applied to CTGAN by adding random noise to the discriminator, and thereafter, the norm is clipped to make it differentially private. The noise addition part is similar to the work of differentially private GANs in [10]. Another differentially private GAN we use is Private Aggregation of Teacher Ensembles Conditional Tabular Generative Adversarial Network (**PATE-CTGAN)**. PATE-CTGAN [12] is similar to the work done in [11] on PATE-GAN, where the original data is partitioned into k subsets and k differentially private teachers are trained to distinguish between the real and the fake data. In PATE-CTGAN, k conditional generators are initialized to generate samples, whereas, in PATE-GAN, there is only one generator.

We also use diffusion Models for tabular data, abbreviated as **TAB DDPM** [13]. It is a variant of Denoising Diffusion Probabilistic Models (DDPM) [14] for tabular data. Diffusion models are likelihood-based generative models that handle the data through forward and reverse Markov processes. The forward process works by adding noise (Gaussian noise) to the available training data, and in the reverse process, a deep neural network learns to invert the diffusion process (by learning what noise was added) with a known Gaussian kernel.

2.2 Attribute Inference Attack

Attribute disclosure happens whenever access to data allows an attacker to learn new knowledge about a particular individual [16]. E.g., the value of a specific attribute like disease, age, gender, race, etc. In our work, the objective is to know how successful the attribute inference attack is when the synthetic data is openly shared. Stadler *et al.* [2] demonstrated that in synthetic datasets, the records with rare values or outliers are highly vulnerable to disclosure, and it is very likely that the strong correlation among the attributes in the real data will be propagated in the synthetic data.

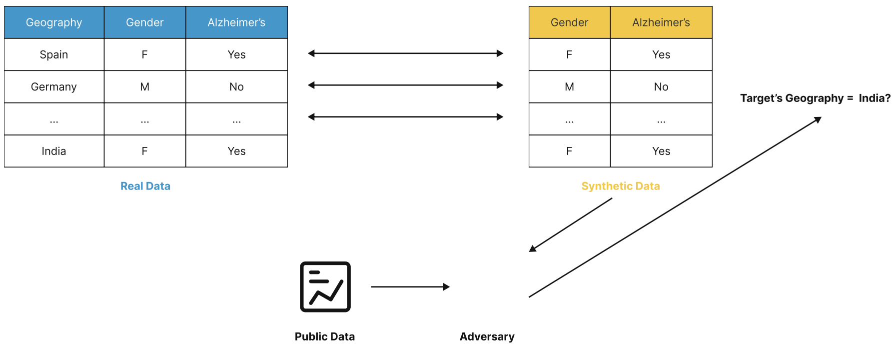

Fig. 1. Attribute Inference Attack

3 Attack Methodology

Our aim is to know how successful the attackers are in inferring the information unknown to them pertaining to a target individual. To successfully execute an attribute inference attack, we assume that the attacker has access to a subset of attributes of a target individual, and a synthesized version of the original data. Using the publicly available synthetic data, and a subset of original attributes, the attacker aims to infer the information of attributes unknown to the attacker, as shown in Fig. 1. For our experimentation, we obtain the synthesized version of the original data using CTGAN, DP-CTGAN, PATE-CTGAN, and TAB DDPM (diffusion models for tabular data).

To infer the value of a sensitive attribute of a target individual, the attacker computes the Euclidean distance between the subset of real attributes known to the attacker and the subset of a synthesized version of those real attributes. For each target record, we have a record at a particular index that is at a minimum Euclidean distance in the synthesized version. Hence, for each target record, we infer the values of unknown attributes from the values at the index (with the minimum distance) in the complete synthetic data. Basically, here we are aligning the synthetic data with the original data by computing the Euclidean distance between the subset of real attributes known to the attacker and the subset of a synthesized version of those real attributes.

After obtaining the aligned synthetic data, we evaluate the success of attribute inference. To measure the success of the attack, we use the following metric called Inference Accuracy.

Table 1. Attack Inference Accuracy on churn-modelling data for Age attribute

Model	$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.5$	$\delta = 0.7$	$\delta = 1$
CTGAN	0.191	0.575	0.871	0.977	0.995
DP-CTGAN	**0.117**	**0.352**	**0.568**	**0.706**	**0.826**
PATE-CTGAN	0.125	0.357	0.585	0.726	0.858
TAB-DDPM	0.232	0.604	0.834	0.919	0.970

Suppose S is the synthetic data obtained after the alignment, and O is the original data. Let n be the total number of samples in the original and the synthetic data, O_j be the value of the sensitive attribute from the original data, which the attacker aims to infer, and S_j is the inferred attribute in the synthetic data corresponding to the sensitive attribute O_j. Let δ be the deviation between the original and the synthetic attribute that can be tolerated to measure the level of inference for a record. The lower the δ, the closer the values of S_j and O_j must be to each other.

For the continuous attributes, the Inference Accuracy, $I.A.$ is defined as follows.

$$I.A. = \frac{\#\left\{\hat{S}_j : |\frac{O_j - S_j}{S_j}| \leq \delta, j = 1 \ldots n\right\}}{n} \tag{1}$$

where # means count. I.A. is the percentage of inferred entries for which the relative errors are within δ.

For the categorical attributes, the above formula is more strict (as we are counting only the **exact matches**) and changes to

$$I.A. = \frac{\#\left\{\hat{S}_j : O_j == S_j, j = 1 \ldots n\right\}}{n} \tag{2}$$

Table 2. Attack Inference Accuracy on churn-modelling data for categorical attributes

Attribute	CTGAN	DP-CTGAN	PATE-CTGAN	TAB DDPM
Gender	0.495	0.505	0.539	0.505
Female	0.629	0.642	0.476	0.628
Male	0.383	0.390	0.505	0.403
Location	0.364	0.327	0.332	0.378
Spain	0.259	0.313	0.345	0.235
Germany	0.263	0.502	0.392	0.238
France	0.468	0.246	0.296	0.522

4 Data and Experimental Settings

We work with churn modelling, and diabetes datasets, as shown in Table 3 from the UCI machine learning repository[1]. The churn modeling data set contains details of a bank's customers, and the target variable is a binary variable that shows whether the customer closed their account or they continues to be a customer.

Table 3. Description of datasets

Data	#Train	#Validation	#Test	#Num	#Cat	Task Type
Churn-Modelling	6400	1600	2000	7	4	Binary Classification
Diabetes	491	123	154	8	0	Binary Classification

We use the publicly available code for CTGAN[2], DP-CTGAN, PATE-CTGAN[3], and TAB DDPM[4]. We mention the value of the privacy parameter ϵ in our results section. We obtain the anticipated trend, which states that the lower the value of ϵ, the higher the privacy (shown by the results of attribute inference attack), and the lower the utility of the synthetic dataset (shown by the accuracy, precision, recall and the F1 score obtained by the random forest classifier). We provide our code[5] for executing the attribute inference attack on the synthetic datasets.

[1] https://archive.ics.uci.edu/.
[2] https://pypi.org/project/ctgan/.
[3] https://github.com/opendp/smartnoise-sdk.
[4] https://github.com/yandex-research/tab-ddpm.
[5] https://github.com/SaloniKwatra0802/Evaluation-of-Synthetic-Dataset-via-an-Attribute-Inference-Attack.

5 Experimental Results and Analysis

We observe that combining differential privacy with the synthetic datasets degrades the utility. We experimented with $\epsilon = \{0.2, 0.3, 0.5\}$. Our experiments favored DP-CTGAN over PATE-CTGAN, which is similar to [12]. For categorical attributes like gender, geographical location, etc., attribute inference accuracy is low; we can say that it is better than random guessing in the case of binary categorical attributes like gender, as shown in Table 2, as the churn-modelling dataset was not perfectly balanced in terms of the gender and the location. We used $\epsilon = 0.2$ for DP-CTGAN and PATE-CTGAN for the results in Table 2.

Table 4. Attack Inference Accuracy on diabetes data for categorical attributes

Attribute	CTGAN	DP-CTGAN, eps = 0.3	PATE-CTGAN, eps = 0.3	TAB DDPM
No. of Pregnancies (exact match)	0.057	**0.051**	0.053	0.101
$\delta = 1$	0.177	**0.163**	0.181	0.283
$\delta = 2$	0.279	**0.257**	0.287	0.446
Disease	0.505	**0.486**	0.527	0.648
Diabetic	0.678	0.475	0.663	0.462
Non-Diabetic	**0.412**	0.495	**0.370**	0.746

Figure 2, and Fig. 3 shows the comparison of accuracies and F1-scores on the data created by different generative models and on the original data.

In Table 2, the values corresponding to the gender and the location show the accuracy for the correct inference for the gender and the location, and the values corresponding to females, and males show the recall (how many females/males were correctly predicted as females/males). Similarly, the values corresponding to Spain, Germany, and France attribute shows the recall. For continuous attributes like age, we can infer an approximated value for the target record. Hence, we can have an approximated attribute inference, not an exact attribute inference, as shown in Table 1, and 5 for churn-modelling and diabetes data respectively. From our results, DP-CTGAN is better than CTGAN, PATE-CTGAN, and TAB-DDPM from the perspective of privacy, as the attack results are the lowest for DP-CTGAN.

In Table 4, we show the attack inference accuracy for the disease attribute, and the recall of diabetic and non-diabetic records. We also show the attack accuracy for the attribute number of pregnancies (exact match, δ=1, and δ=2). Overall, DP-CTGAN was the best in terms of having lower attack accuracy. Diffusion models are better than other generative models, and their variants, which includes CTGAN, PATE-CTGAN, and DP-CTGAN from the perspective of classification performance, as shown in our results and also in [13].

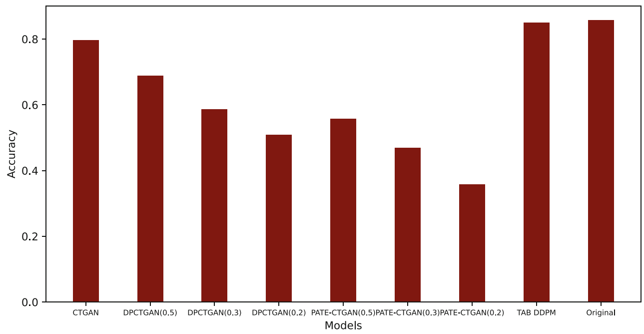

Fig. 2. Comparison of the accuracies of the data generated by GANs and the original data for churn-modeling data

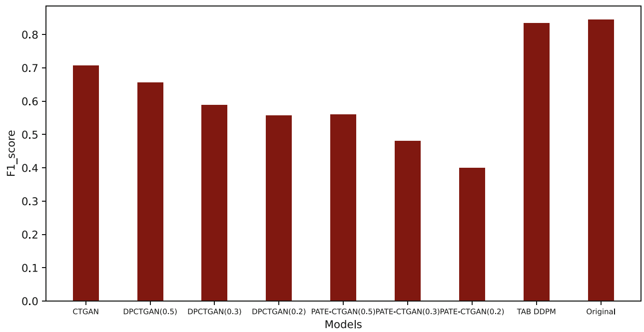

Fig. 3. Comparison of the F1-scores of the data generated by GANs and the original data for churn-modeling data

Table 5. Attack Inference Accuracy on diabetes data for the age attribute

Model	$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.5$	$\delta = 0.7$	$\delta =1$
CTGAN	0.297	0.713	0.881	0.924	0.963
DP-CTGAN	**0.098**	**0.367**	**0.566**	**0.672**	**0.792**
PATE-CTGAN	0.222	0.595	0.864	0.938	0.993
TAB-DDPM	0.226	0.513	0.792	0.931	0.969

6 Conclusion and Future Works

The goal of this work is to evaluate whether the synthetic data generated from different generative models is private enough to be made publicly available. To achieve our objective, we executed an Attribute Inference Attack against synthetic datasets created using different generative algorithms. From our results, we show that the exact attribute inference is not highly feasible for synthetic datasets in the case of categorical attributes. Nevertheless, approximate inference is possible in the case of continuous attributes.

References

1. Kwatra, S., Torra, V.: Data reconstruction attack against principal component analysis. In: Arief, B., Monreale, A., Sirivianos, M., Li, S. (eds.) Security and Privacy in Social Networks and Big Data. SocialSec 2023. LNCS, vol. 14097, pp. 79–92. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-5177-2_5
2. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data-anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22), pp. 1451–1468 (2022)
3. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(05), 557–570 (2002)
4. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression (1998)
5. Samarati, P.: Protecting respondents identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
6. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(05), 571–588 (2002)
7. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79228-4_1
8. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Pocess. Syst. **27** (2014)
9. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. Adv. Neural Inf. Process. Syst. **32** (2019)
10. Zhang, X., Ji, S., Wang, T.: Differentially private releasing via deep generative model (technical report) (2018). arXiv preprint arXiv:1801.01594
11. Jordon, J., Yoon, J., Van Der Schaar, M.: PATE-GAN: generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations, September 2018
12. Rosenblatt, L., Liu, X., Pouyanfar, S., de Leon, E., Desai, A., Allen, J.: Differentially private synthetic data: applied evaluations and enhancements (2020). arXiv preprint arXiv:2011.05537
13. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: Tabddpm: modelling tabular data with diffusion models (2022). arXiv preprint arXiv:2209.15421
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
15. De Cristofaro, E.: What is synthetic data? The good, the bad, and the ugly (2023). arXiv preprint arXiv:2303.01230
16. Hittmeir, M., Mayer, R., Ekelhart, A.: A baseline for attribute disclosure risk in synthetic data. In: Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy, pp. 133–143, March 2020