Assessment of Privacy and Utility in Synthetic Data

Honor Thesis - Fall 2023

Yiwen (Oliver) Wu, Luke Francisco, Ambuj Tewari

University of Michigan
Department of Statistics

1 Introduction

As sensitive data, such as financial information, health records, and biometric data, becomes more prevalent in the era of big data, ensuring personal privacy when handling data has become a top priority. However, this poses a challenge for researchers who need access to such data for their studies, as organizations are legally and ethically obligated to keep it confidential.

Finding the right balance between using data effectively and protecting individual privacy is extremely important. This research project focuses on synthetic data, a growing field that aims to preserve the usefulness of datasets while safeguarding privacy.

The need for synthetic data arises from the requirement to find a middle ground between data accessibility and privacy preservation. In situations where the release of original data could result in privacy breaches, synthetic data offers a valuable alternative that can be shared freely. By creating datasets that mimic the statistical characteristics of the original data without including any personal information, we can support research and development while protecting privacy. This approach enables us to harness the valuable insights contained within sensitive data, unlocking new avenues for innovation while adhering to rigorous privacy standards.

Historically, various methods have been employed for generating synthetic data, evolving from traditional statistical models to advanced machine learning approaches. Traditional methods include Bayesian networks (Zhang et al., 2017), Gaussian and vine copulas (Jeong et al., 2016), Markov models (Cai et al., 2021), and decision trees (Reiter, 2005), which focus on capturing low-dimensional data characteristics. More recently, machine learning techniques, particularly deep learning models, have been utilized to address the complexities inherent in high-dimensional data. These methods include PATEGAN (Yoon et al., 2019), CT-GAN, and variational autoencoders (TVAE) (Xu et al., 2019). These methods collectively represent the progression in synthetic data generation, reflecting a shift from basic statistical modeling to sophisticated algorithm-driven approaches. As synthetic data generation methods have evolved, there have been growing challenges in creating synthetic medical data. One of the key challenges highlighted in a recent comprehensive review is the importance of ensuring privacy when sharing health data for medical informatics research (Murtaza et al., 2023). The combination of past and current methodologies highlights the dynamic nature of synthetic data research and its crucial role in advancing privacy-preserving technologies.

Differential privacy (DP) is a mathematical framework for quantifying the privacy afforded by randomized algorithms that process sensitive information, which can be written as:

$$P(A(D_1) \in B) \leq e^{\varepsilon} P(A(D_2) \in B) + \delta,$$

where the term A refers to a specific algorithm, in this context, one that generates synthetic data. The term B refers to a data entry, and D_1 and D_2 represent two datasets that are identical except for a single data record. As noted by Dwork and others, the strength of privacy protection is inversely proportional to the values of ε and δ, with lower values indicating a stronger privacy guarantee. This privacy measure is typically enforced by introducing noise at a particular stage within the algorithmic process (Dwork and Roth, 2013).

DP provides strong guarantees that the presence or absence of any individual's data in a dataset does not significantly affect the outcome of any analysis, thereby preserving the confidentiality of personal data. As outlined in our paper, differential privacy will serve as the foundational privacy reference throughout our research, making sure that the synthetic data we generate and the techniques we use follow strict privacy guidelines.

This project is a simulation designed to evaluate the privacy and utility of synthetic datasets generated from a real dataset using various data generation methods. These methods should be able to maintain data utility while ensuring privacy. By creating synthetic data from private datasets and comparing it with public datasets, we aim to measure both the usefulness and privacy protection of these synthetic data techniques.

Additionally, this project includes an innovative approach to simulate potential data breaches. We will evaluate the resilience of synthetic data against such attacks. The findings of this research will not only advance the field of data science but also contribute to shaping future policies and practices in data privacy and security.

2 Method

2.1 Datasets

2.1.1 Data Description

The Adult dataset from the UCI Machine Learning Repository was selected because it is relevant and representative when evaluating methods for generating synthetic data. This dataset is derived from the 1994 Census database and is well-known in the field of social science for classification tasks. The dataset includes a range of personal attributes, making it a good choice for assessing techniques that protect privacy when generating synthetic data (Becker and Kohavi, 1996).

This dataset comprises 48,842 instances with 14 features, including demographic and employment-related characteristics. Key attributes include age, work class, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours worked per week, native country, and income. These features contain both categorical and continuous data types, providing a comprehensive basis for analysis.

The Adult dataset contains sensitive personal information, particularly relating to individual income levels, making it highly relevant for studies on privacy preservation. The primary focus on the income attribute as the target variable underscores the need for robust privacy-protecting measures in synthetic data methodologies, given the dataset's potential to reveal financial statuses.

2.1.2 Dataset Cleaning and Manipulation

After conducting initial exploration, we discovered that the dataset contains missing values in 3 columns: `Workclass`, `Occupation`, and `Country`. These columns happen to be categorical variables. To address this, we filled in the missing values by sampling from the original column distributions. This approach allowed us to preserve the original distribution.

Next, the following 3 columns were removed:

1. `Education`: It contains similar information as Education-num, so there is no need to have duplicate variables.
2. `Relationship`: Its meaning is unclear.
3. `fnlwgt`: It is unnecessary for the purpose of this project.

In the data preprocessing step, all columns were binarized to simplify the dataset and facilitate comparisons across different synthetic data generation methods, which all handled categorical data differently. We used the following criteria:

- `Sex` is binarized to `is_female`, assigning 1 to 'Female' and 0 to others.
- `Workclass` is binarized to `work_for_gov`, with government jobs ('State-gov', 'Federal-gov', 'Local-gov') as 1, others as 0.
- `Marital Status` is binarized to `is_married`, marking 'Married-civ-spouse', 'Married-AF-spouse', and 'Married-spouse-absent' as 1, others as 0.
- `Occupation` is binarized to `physical_labor`, with physically demanding jobs like 'Craft-repair', 'Machine-op-inspct', 'Farming-fishing', and 'Handlers-cleaners' marked as 1, representing physical labor, while less physically intensive occupations are set to 0.
- `Race` is binarized to `is_white`, with 'White' as 1, other races as 0.
- `Country` is binarized to `is_US`, setting 'United-States' as 1, other countries as 0.

Then, `Income` was also binarized. In this case, 1 represents income greater than 50k, and 0 represents income equal to or less than 50k.

In the final step of data preparation, the cleansed dataset was randomly divided into two parts: the `private` and the `public` subsets. The `private` subset mimics the data held by institutions, restricted from external access, while the `public` subset simulates the data openly accessible to everyone.

2.2 Synthetic Data Generation Method

In our exploration of synthetic data generation methods, we evaluated four distinct approaches, each catering to specific aspects of privacy and data utility. These methods include the NonPrivate algorithm (Annamalai et al., 2023), the Synthpop (Nowok et al., 2016), PrivBayes (Zhang et al., 2017), and PATE-GAN (Yoon et al., 2019). Each technique offers unique advantages and constraints, addressing different challenges in the realm of synthetic data generation while maintaining privacy.

2.2.1 Non-Private

Non-Private creates synthetic data by randomly selecting records from the private dataset, treating the private dataset as a source of "synthetic" records. This algorithm was selected as a straightforward starting point and a baseline for our investigation into generating synthetic data. Similar to the procedure conducted by Annamalai and others, NonPrivate serves as a control, enabling us to compare it to more sophisticated methods that prioritize privacy (Annamalai et al., 2023). Additionally, we anticipate it to exhibit the highest level of utility since it samples real records, but at the expense of lower privacy due to the same reason.

2.2.2 Synthpop

`synthpop` method is a technique for generating synthetic datasets that are statistically representative of the original data, developed by Nowok, Raab, and Dibben. It starts by assuming the observed data as a sample from a population with estimable parameters, where the synthetic data are essentially drawn from a distribution fitted to these observed parameters. The process of generating synthetic datasets runs in parallel with the fitting of each conditional distribution, ensuring that each synthetic column is generated conditional on previously synthesized columns. The sequential synthesis preserves the relationships between variables, making the synthetic data a statistically coherent representation of the original dataset, but without revealing any individual's specific data (Nowok et al., 2016).

One reason we selected the Synthpop method is because it is a well-established method that is used in a variety of fields. For example, in a paper by Tayefi and others, Synthpop algorithm is used to generate synthetic electronic health records (Tayefi et al., 2021). Moreover, Quintana and others used Synthpop algorithm to generate biobehavioural synthetic data (Quintana, 2020). In addition, Synthpop has a strong and flexible framework that can handle different types and structures of data. It is especially useful for creating multiple synthetic datasets that balance data usefulness and privacy. Additionally, `synthpop`can handle missing values and restricted data, which are important for maintaining the accuracy and usability of the synthetic data. These features make `synthpop` a great choice for researchers who need access to sensitive data while respecting privacy constraints.

The original `synthpop` package is made for R. However, we have used a Python version of `synthpop`, which can be found at hazy/synthpop on Github. This version lets us take advantage of the method's strengths in Python's wide range of tools, making it easier to generate synthetic data and integrate it with our current Python-based workflows.

2.2.3 PrivBayes

Bayesian Network In a dataset D with a set of attributes A, D is considered a joint probability distribution across the product of A's attribute domains. A Bayesian network succinctly captures this distribution by marking the conditional dependencies among A's attributes. This network is a directed acyclic graph (DAG), where each attribute in A is represented as a node, and the directed edges between these nodes represent the conditional independencies. In the example below, a Bayesian network displays five attributes such as age, education, work class, title, and income. Any two attributes, $X, Y \in A$, can relate in one of three ways: direct dependence, weak conditional independence, or strong conditional independence, defining the unique interactions among them.

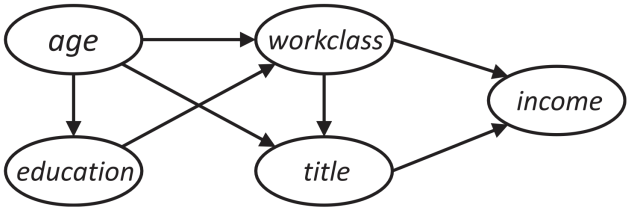

Figure 1: *An Example of a Bayesian Network (Zhang et al., 2017)*

PrivBayes PrivBayes is an advanced method for releasing high-dimensional data while ensuring differential privacy. It constructs a Bayesian network, a probabilistic graphical model, to represent the conditional dependencies among different attributes in the dataset. This network breaks down the joint distribution of attributes into lower-dimensional conditional distributions. PrivBayes achieves privacy protection by introducing differentially private Laplace noise to these distributions, thereby complying with ε-Differential Privacy. This method effectively preserves the utility of the data for analytical purposes while rigorously safeguarding the privacy of individual data points (Zhang et al., 2017).

In our project, we utilized Ping and others' implementation within the DataSynthesizer library (DataSynthesizer) to execute our code. This library provided a robust framework for data synthesis, particularly with its '`correlated_attribute_mode`'. This mode was especially relevant to our analysis as it allowed us to efficiently simulate the complex inter-attribute correlations present in our dataset, thereby ensuring a more realistic and representative synthetic data generation process.

In our application of the PrivBayes method, we made two specific adjustments: the Degree of the Bayesian Network (k) and the Privacy Budget (ε). The Degree of the Bayesian Network determines the maximum number of attributes in the conditional distributions, which is important for accurately capturing the relationships between data attributes. A high k more accurately captures relationships between attributes, but it comes at the expense of much greater computation time.The Privacy Budget, ε, a key parameter in differential privacy, controls how much noise is added to the data, striking a balance between privacy protection and data usefulness.

We adjusted these two parameters to simulate varying levels of privacy. This allowed us to explore different scenarios that balanced data usefulness and privacy protection. In the next phases of our study, we will carefully evaluate the impact of these settings on the ability of synthetic data to accurately represent the original dataset while maintaining privacy standards.

2.2.4 PATE-GAN

Generative Adversarial Networks (GANs) GANs consist of two neural network models that are trained simultaneously through adversarial processes. The first model, the generator, learns to generate new data similar to the training set. The second model, the discriminator, learns to distinguish between the generator's fake data and the real data from the training set. As training progresses, the generator improves its ability to produce data that are indistinguishable from real data, while the discriminator becomes better at telling the difference. This competition drives both networks to improve until the discriminator can no longer easily tell real from fake data, at which point the generator has learned to produce very realistic synthetic data (Goodfellow et al., 2014).

PATE-GAN In the PATE-GAN framework described in the paper *PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees*, noise is introduced during the training of the discriminator through the Private Aggregation of Teacher Ensembles (PATE) mechanism (Yoon et al., 2019). This mechanism involves a set of teacher models that are trained on separate subsets of the private data. When a new input is classified, each teacher votes on an outcome, and these votes are aggregated with Laplace noise to ensure differential privacy. The noise scale is determined by the privacy budget λ, which represents the amount of noise. Therefore, a larger λ means more noise and, consequently, more privacy in the synthetic data generated. The discriminator's final output is the noisy aggregation of the teacher votes, and the generator is trained to produce samples that the discriminator will classify as real. This process ensures that both the discriminator and the generator are differentially private in relation to the original data.

During our research, we attempted two different implementations of the PATE-

GAN method. However, despite our best efforts, the outcomes from both trials were not viable for practical application. As a result, we have decided not to include these results in the *Results* section of our study. In the future, we see the potential for further exploration and refinement of the PATE-GAN method.

2.3 Utility Measuring

In our study, it is important to measure how useful synthetic data is in order to make sure that it accurately copies the statistical features of the original, private data. This evaluation helps us assess how well the synthetic data can be used instead of real data in different analyses.

- **Neural Network Model Predicting AUROC**: We first train a neural network on the private data, assessing its performance with the test Area Under the Receiver Operating Characteristic (AUROC) on public data. We then replicate this process for each synthetic dataset, comparing their AUROC scores on the public data to gauge their predictive accuracy. When comparing the AUROC scores of models trained on synthetic data with the model trained on private data, we can determine if the synthetic data provides similar prediction results, indicating similar prediction capability, as the private data.
- **Correlation Matrix**: By comparing the heat maps of the correlation matrices of each feature between private and synthetic data, we can visualize and assess the preservation of inter-variable relationships. This comparison helps in understanding how closely the synthetic data mimics the complex interactions present in the original dataset.
- **Histogram/Barplot**: Examining the distribution histograms or barplots of each column for both private and synthetic data allows us to compare their distributions. This visual and statistical analysis ensures that the synthetic data replicates the original data distribution characteristics, an important aspect of data utility.

2.4 Attack Simulation and Privacy Measuring

In our study, conducting attacks is a critical step to evaluate the privacy robustness of the synthetic data we generate. Property inference attacks pose a sophisticated threat by attempting to deduce sensitive attributes from aggregated data. By simulating potential real-world attacks, including property inference attacks, we can assess how effectively our synthetic datasets protect individual data points from re-identification. We will utilize three different attack methods - Neural Networks (NN), Logistic Regression, and K-Nearest Neighbors (KNN) - each providing a distinct perspective to measure data privacy and resistance to such inference techniques. These methods will help us understand the effectiveness of our privacy-preserving techniques and ensure that the synthetic data offers strong protection against various types of inferential attacks.

However, it is also important to acknowledge the inherent limitations of such exercises. While our methodology provides a measure of the synthetic data's robustness against privacy breaches, we must recognize that it is not exhaustive of all possible attack vectors. The effectiveness of an attack can vary greatly depending on the attacker's method and expertise. Therefore, a simulation that yields poor attack performance should not be seen as a guarantee of security. There is always the potential for other attackers to succeed using alternative strategies. This uncertainty is a fundamental aspect of any simulation-based approach to privacy assessment.

2.4.1 Assumptions

In our attack model, we follow the assumption described in the paper *A Linear Reconstruction Approach for Attribute Inference Attacks against Synthetic Data*. The paper suggests that a hacker has access to all information in a private dataset except for the target column (Annamalai et al., 2023). While we understand that this level of access is rare, our analysis considers this worst-case scenario to test the synthetic data's resilience against severe privacy breaches.

2.4.2 Attack Simulation Steps

1. Train a model on the public dataset and make predictions on the private dataset using AUROC as the performance metric. This result will serve as the baseline measure.

2. Train models on each of the synthetic datasets and make predictions on the private dataset using AUROC as the performance metric as well.

3. Compare the resulting AUROC scores.

In addition to utilizing synthetic data of equivalent size (around 14,000 records) as the ”private” datasets for training attack models, we will also conduct experiments with a synthetic dataset containing a greater number of datapoints (around 50,000 records). The objective is to examine whether the inclusion of more datapoints in the training dataset will enhance the performance of the attack.

It is also important to note that when training the KNN model, the target variable was not included in the computation of nearest neighbors. Furthermore, the Euclidean distance metric was used for identifying the nearest neighbors in the KNN model.

The expected outcome of the attack simulation is that the model trained on synthetic data will have better prediction accuracy compared to the model trained on the public data. Additionally, datasets with less privacy (higher epsilon) are expected to exhibit better performance (higher AUROC).

3 Results

In this section, we will share the results from our analyses of utility measurement and attack simulation. We will thoroughly examine how our synthetic datasets perform against different attack methods and their ability to maintain the statistical accuracy of the original data. Our goal is to provide a complete understanding of how effective our methods are in generating synthetic data in terms of usefulness and privacy.

3.1 Utility Measuring Results

This section presents the results of our utility evaluation, demonstrating how the synthetic data closely aligns with the statistical characteristics of the original private datasets. We assess the predictive performance of the synthetic datasets using AUROC scores, analyze the correlation structures through heatmaps, and examine the distributional congruence using histograms and barplots.

3.1.1 Neural Network Prediction

Table 1 below show the AUROC scores obtained by neural networks trained on private and synthetic datasets in the prediction task. These scores provide insights into how effectively our synthetic data replicates the predictive patterns of the private dataset.

Table 1: Utility Reusult

Dataset	ε	k	AUROC
private	N/A	N/A	0.7593
non_private_1	N/A	N/A	0.7702
synthpop_1	N/A	N/A	0.6090
privbayes_1	0	8	0.7260
privbayes_3	5	5	0.6400
privbayes_4	2	4	0.6104
privbayes_5	1	3	0.5461
privbayes_6	0.1	0	0.5204

Form Table 1, we can see that `non_private_1` achieves a similar AUROC compared to `private`, and that of `synthpop_1` is significantly lower.

In PrivBayes method, we have the flexibility to adjust the values of ε and k to achieve varying levels of privacy. The table below presents the results obtained from various representative datasets with different privacy configurations. A value of $\varepsilon = 0$ signifies no differential privacy. From results in Table 1, as the value of ε decreases, the level of differential privacy increases, and the AUROC of the PrivBayes data decreases correspondingly.

3.1.2 Correlation Heatmaps

The heatmaps below (Figure 2) show the correlation coefficients among different attributes, comparing the private data with each synthetic dataset. These heatmaps are important for illustrating the degree to which our synthetic datasets preserve the statistical relationships present in the original data.

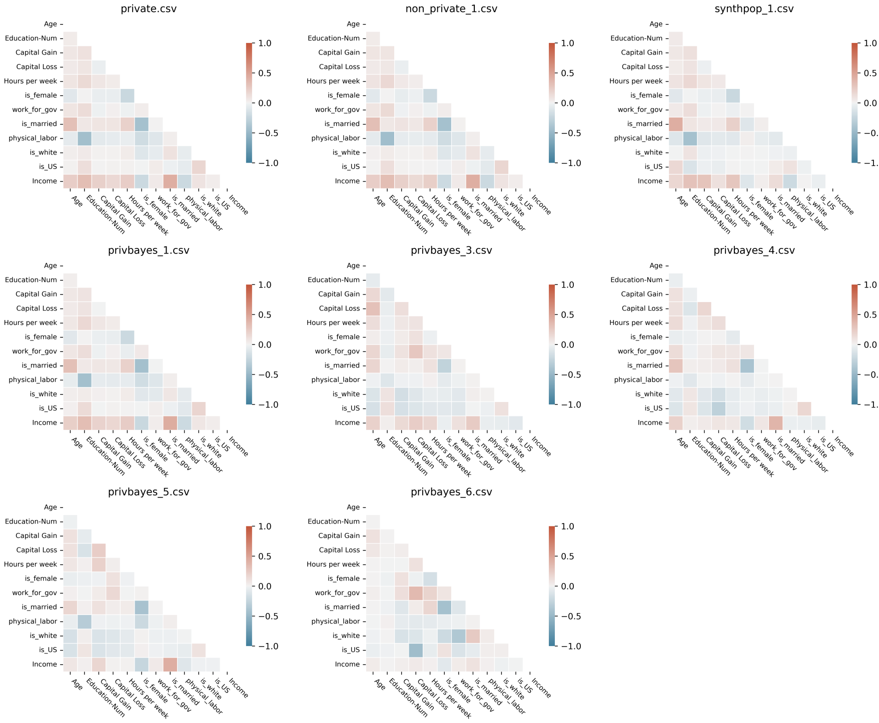

Figure 2: Correlation Heatmaps of Private and Synthetic Datasets

The graphs of `non_private_1` and `privbayes_1` in Figure 2 resemble the `private` dataset the most, considering that these two synthetic datasets have the lowest level of privacy. Conversely, the remaining graphs differ from the private dataset in various ways.

3.1.3 Histogram/Barplot

The histograms and barplots in this section provide a visual comparison of distribution patterns between the private dataset (red) and synthetic datasets (blue). They quantitatively represent the frequency and distribution of data points across different categories and ranges, providing insights into the similarity between the synthetic and the private dataset. Additionally, please note that we have not included graphs for every synthetic dataset. Instead, we only chose a few that are representative.

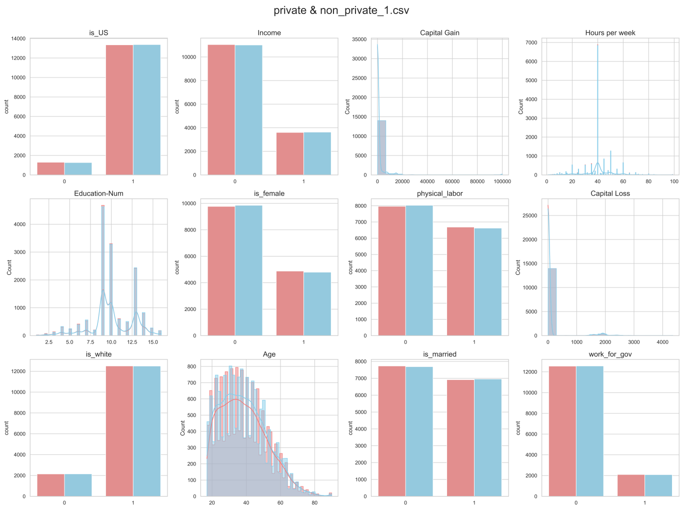

Figure 3: private V.S. non_private_1

Figure 3 shows that by directly sampling from the original dataset, `non_private_1` exhibits a significant similarity to `private`.

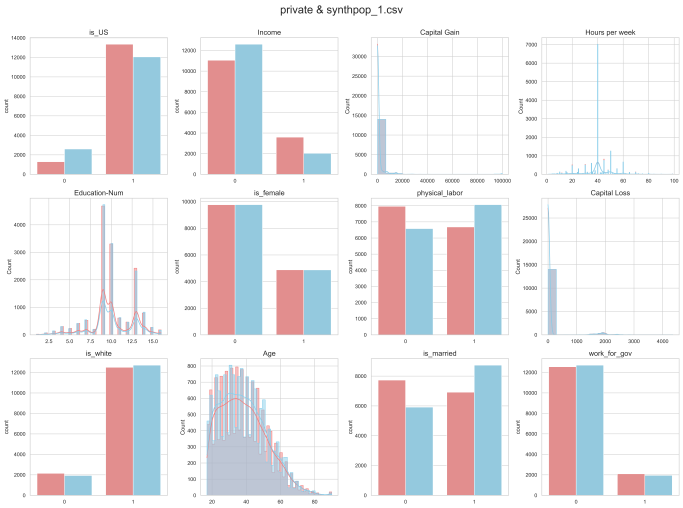

Figure 4: private V.S. synthpop_1

Although there are differences in most of the columns shown by Figure 4, the `synthpop_1` still follows the distribution pattern of the `private`.

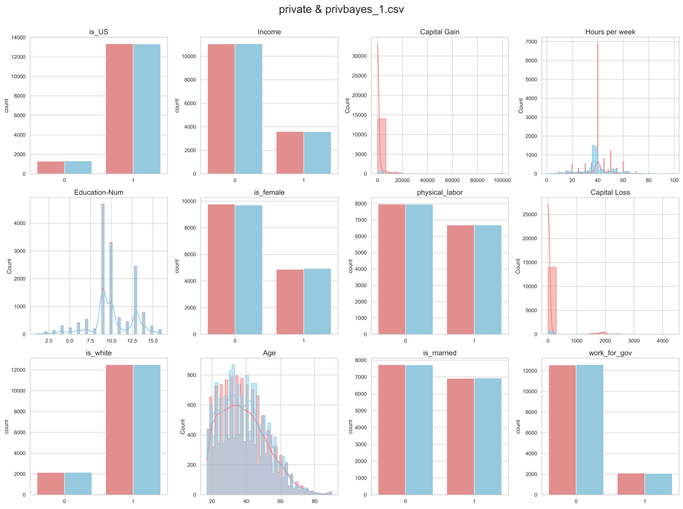

Figure 5: private V.S. privbayes_1

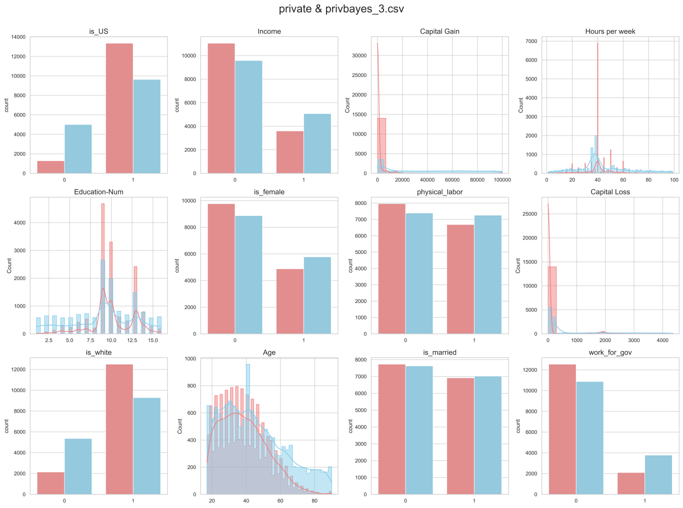

Figure 6: private V.S. privbayes_3

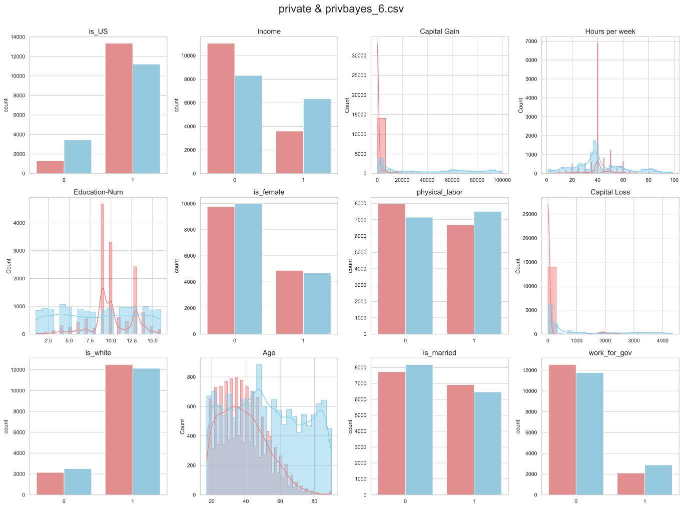

Figure 7: private V.S. privbayes_6

In Figure 5, 6, and 7, `privbayes_1`, `privbayes_3`, and `privbayes_6` represent no privacy, moderate privacy, and high privacy levels, respectively. As more noise is added to the data, we can observe a shift in the distribution of the data from being very similar to becoming increasingly different from `private`.

3.2 Attack Simulation and Privacy Measuring Results

In this part, we will present the results of our comprehensive evaluations on the resilience of synthetic data against privacy attacks. By utilizing attack models such as Neural Networks (NN), Logistic Regression, and K-Nearest Neighbors (KNN), we aim to assess the effectiveness of our synthetic datasets in protecting individual data points from potential re-identification. The tables will provide AUROC scores for each method, reflecting the performance of attacks across different datasets.

Table 2: Neural Network Model Attack

Datasets	ε	k	AUROC
public	N/A	N/A	0.7619
nonprivate_1	N/A	N/A	0.7784
synthpop_1	N/A	N/A	0.6518
privbayes_1	0	8	0.7746
privbayes_3	5	5	0.6012
privbayes_4	2	4	0.6871
privbayes_5	1	3	0.7168
privbayes_6	0.1	0	0.5080

In Table 2, `non_private_1` exhibits a higher AUROC than `public`, which is expected. Additionally, `synthpop_1` demonstrates better privacy preservation despite having a lower utility score. Furthermore, the PrivBayes datasets do not demonstrate a monotonic increase in privacy preservation as the privacy level increases.

Table 3: Logistic Regression Model Attack

Datasets	ε	k	AUROC
public	N/A	N/A	0.8938
nonprivate_1	N/A	N/A	0.8944
synthpop_1	N/A	N/A	0.8409
privbayes_1	0	8	0.8926
privbayes_3	5	5	0.8596
privbayes_4	2	4	0.8229
privbayes_5	1	3	0.8045
privbayes_6	0.1	0	0.7456

From Table 3, we can see that the Logistic Regression Model Attack achieves the best attacking results among all three attack models. In contrast to the results from the Neural Network Model Attack, the PrivBayes datasets show a monotonic increase in privacy preservation in this table.

Table 4: KNN Model Attack

Datasets	ε	k	AUROC
public	N/A	N/A	0.8716
nonprivate_1	N/A	N/A	0.9034
synthpop_1	N/A	N/A	0.7500
privbayes_1	0	8	0.8979
privbayes_3	5	5	0.7916
privbayes_4	2	4	0.7793
privbayes_5	1	3	0.7584
privbayes_6	0.1	0	0.5423

Table 4 exhibits that the results from the KNN Model Attack have a similar pattern to the results from the Logistic Regression Model Attack, except with a slightly worse prediction AUROC. This indicates that the KNN model was not as successful compared to the Logistic Regression model.

Table 5: Attack AUROC with Larger Synthetic Datasets

Attack Method	Generation Method	Original Data	Large Data
NN	Synthpop	0.6518	0.6962
NN	PrivBayes ($\varepsilon = 0$)	0.7746	0.7479
NN	PrivBayes ($\varepsilon = 0.1$)	0.5080	0.5181
Logistic Regression	Synthpop	0.8409	0.8459
Logistic Regression	PrivBayes ($\varepsilon = 0$)	0.8926	0.8927
Logistic Regression	PrivBayes ($\varepsilon = 0.1$)	0.7456	0.7613
KNN	Synthpop	0.7500	0.7378
KNN	PrivBayes ($\varepsilon = 0$)	0.8979	0.9312
KNN	PrivBayes ($\varepsilon = 0.1$)	0.5423	0.5067

In Table 5 are the AUROC scores from a total of 9 tests, which include 3 attack methods with 3 AUROC comparisons in each method. We can observe that there is no clear improvement in predicting AUROC with larger datasets.

4 Discussions

Privacy-Utility Tradeoff In the context of synthetic data generation, there is a crucial balance between utility and privacy. This balance highlights the trade-off between the usefulness of the data and the level of privacy protection. Strong privacy measures, such as adding noise to the data, can protect individual data points from re-identification risks. However, these measures may compromise data accuracy and the level of detail required for meaningful analysis. On the other hand, prioritizing high data utility often involves less aggressive privacy interventions, which may increase the risk of privacy breaches (Annamalai et al., 2023).

4.1 NonPrivate

This method involves sampling directly from the private dataset, so it is expected for the utility AUROC to be similar to, or even higher than, that of the *public* dataset (Table 1). Furthermore, when examining the correlation heatmaps and histogram/barplot graphs, this dataset exhibits the closest correlation and column distribution to the original private dataset (Figure 2 & Figure 3).

While this dataset offers high utility, it does not provide any privacy. When comparing the attack AUROC of the model trained on the public dataset to that of the dataset generated by the NonPrivate algorithm, the latter consistently shows a higher value. However, this result is not surprising, as the NonPrivate algorithm essentially draws samples from the private dataset.

4.2 Synthpop

Compared to the private dataset's utility AUROC score of 0.76, the synthetic dataset generated by the synthpop algorithm has a significantly lower score of 0.61 (Table 1). This indicates that the synthetic dataset contains much less usable information. As a result, we expect a better privacy preservation, which is also supported by the data in Table 2, 3, and 4.

Moreover, it is worth noticing that a synthetic data generation method displays different levels of privacy protection against different attacks. Synthpop synthetic data is a great example of this. When attacked by models like Neural Network and KNN, Synthpop shows a significantly lower attack AUROC compared to either the public dataset or the NonPrivate dataset. However, when it comes to Logistic Model attack, it has a much closer attack AUROC, indicating that it is easier to deduce information using the Logistic Model attack, and this dataset has a higher privacy breach risk in this situation (Table 2, 3, 4).

4.3 PrivBayes

PrivBayes is a method that allows us to adjust the level of privacy we want to include in the synthetic dataset. Based on all three utility measurement criteria, we can observe that as we increase the value of ε and adjust the corresponding k to fit the dataset, more noise is introduced to the dataset, leading to a decrease in the utility AUROC (Table 1).

The correlation heatmaps barplots further support this observation. In the `privbayes_1` dataset where no privacy is applied, the dataset exhibits the closest correlation and distribution to the private dataset compared to all other PrivBayes datasets. As we gradually increase the level of privacy, the correlation heatmaps and distribution graphs start to deviate from the baseline. Finally, at a high level of privacy, the resulting graph shows the least similarity to the baseline (Figure 2, 5, 6, 7).

The prediction performance of attack simulation also shows a tradeoff between privacy and utility. When more noise is introduced to the synthetic data, the attack AUROC decreases, indicating a higher level of privacy protection.

It is important to note that, according to the results of the attack simulation, the performance of the Neural Network appears to be inferior compared to the other two attacking models. While the attack AUROC of the other two models decreases monotonically with increasing noise, the attack AUROC scores of the Neural Network exhibit more variability. This could be due to the fact that the NN model we built does not fit this specific dataset (Table 2).

Furthermore, after examining the results, it becomes clear how important it is to carefully choose parameters when generating synthetic datasets. A notable example is the "privbayes_5" dataset. In terms of utility measurement, it only achieves a utility AUROC score of 0.55, which is only slightly better than random guessing (considering that the AUROC for random guessing is

0.5). However, during an attack, this same dataset shows a significantly higher attack AUROC, reaching a maximum score of 0.80 when subjected to the Logistic Regression Model attack (Table 1, 3). This suggests that while the dataset may seem almost useless, it actually contains a significant amount of private information from the original dataset. Consequently, hackers can easily infer sensitive information from this private data, highlighting the extreme danger associated with the leakage of datasets like this.

4.4 Performance of Attack with Larger Datasets

When analyzing the result table, we can see that there is no clear pattern indicating whether having more data points results in a higher or lower attack AUROC score. The variations in attack AUROC are minimal and can be mainly attributed to the random nature of the model training process (Table 5).

5 Conclusion

The findings of this research reveal the complex relationship between the utility of synthetic data and the importance of privacy. Our empirical tests demonstrate that while the NonPrivate method offers significant utility by closely replicating the original data, it lacks any privacy safeguards. This underscores the critical need for synthetic data generation methods that prioritize privacy. Techniques like Synthpop and PrivBayes strike a delicate balance between privacy and utility, as evidenced by their utility and attack AUROC scores.

Our study provides valuable metrics on the utility and privacy of synthetic data derived from a real dataset, offering practical insights for researchers considering publishing synthetic datasets. Findings from this project demonstrate an important aspect of synthetic data: despite sacrificing some utility, there is still a risk of privacy breaches. This complexity emphasizes the need to consider more than just utility measures when assessing privacy safety, highlighting the balance required in generating synthetic data to protect individual privacy. Interestingly, generating larger datasets did not significantly enhance privacy, suggesting that the quantity of data alone does not dictate privacy levels. This reinforces the necessity for strategic parameter calibration in synthetic data generation to achieve optimal results for analysis and privacy.

For future research, it is crucial to explore improvements in synthetic data methodologies, focusing on striking the delicate balance between data utility and privacy. This could involve investigating novel algorithms capable of providing higher accuracy in data synthesis, such as GAN-based approaches (PATE-GAN, CTGAN, etc) or diffusion models. Additionally, expanding the scope of research to encompass diverse datasets and domains would help validate the resilience of synthetic data generation techniques across different contexts, with examples including high-dimensional data and longitudinal data. Another area of promising future research is the development of adaptive privacy mechanisms, potentially leveraging artificial intelligence to dynamically adjust privacy parameters, ensuring an optimal balance tailored to the specific characteristics of each dataset.

References

Annamalai, M. S. M. S., Gadotti, A., and Rocher, L. (2023). A linear reconstruction approach for attribute inference attacks against synthetic data.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Cai, K., Lei, X., Wei, J., and Xiao, X. (2021). Data synthesis via differentially private markov random fields. *Proceedings of the VLDB Endowment*, 14(11):2190–2202.

Dwork, C. and Roth, A. (2013). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., and OTHERS (2014). Generative adversarial networks.

Jeong, B., Lee, W., Kim, D.-S., and Shin, H. (2016). Copula-based approach to synthetic population generation. *PLOS ONE*, 11(8):e0159496.

Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., and Bano, A. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546.

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74(11).

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife*, 9:e53275.

Reiter, J. P. (2005). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21:441–462.

Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., and Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Computational Statistics*, 13(6).

Xu, L., Skoularidou, M., Cuesta-Infante, A., and OTHERS (2019). Modeling tabular data using conditional gan.

Yoon, J., Jordon, J., and van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2017). Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems*, 42(4):1–41.