Differentially Private Copulas, DAG and Hybrid Methods: a Comprehensive Data Utility Study

Andrea Galloni[1] [0000−0002−4163−7193] and Imre Lendák[1,2] [0000−0001−6188−49364]

[1] ELTE – Eötvös Loránd University, Faculty of Informatics, Department of Data Science and Engineering, Budapest, Hungary
andrea.galloni@inf.elte.hu
[2] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
lendak@uns.ac.rs

Abstract. Differentially Private (DP) synthetic data generation (SDG) algorithms take as input a dataset containing private, confidential information and produce synthetic data with comparable statistical characteristics. The significance of such techniques is rising due to the growing awareness of the extent of data collection and usage in organizational contexts, as well as the implementation of new stricter data privacy regulations. Given the growing academic interest in DP SDG techniques, our study intends to perform a comparative evaluation of the statistical similarities and utility (in terms of machine learning performances) of a specific set of related algorithms in the realistic context of credit-risk and banking. The study compares PrivBayes, Copula-Shirley, and DPCopula algorithms and their variants using a proposed evaluation framework across three different datasets. The purpose of this study is to perform a thorough assessment of the score and to investigate the impact of different values of the privacy budget (ϵ) on the quality and usability of synthetic data generated by each method. As a result, we highlight and examine the deficiencies and capabilities of each algorithm in relation to the features' properties of the original data.

Keywords: Synthetic Data Generation · Differential Privacy · Evaluation Metrics · Copula Functions · Bayesian Networks.

1 Introduction

The overall Differentially Private (DP) Synthetic Data Generation (SDG) task given a private dataset D_p composed by n_p records and m features: $D_p \subset X_1 \times X_2 \times \cdots \times X_m$ of n_p records and m features is to generate a synthetic dataset D_s which keeps attributes types and their number $X_1, X_2, \cdots, X_m$ as D_p. The goal is to generate a new synthetic dataset D_s such that it resembles the statistics and subsequently the utility of D_p while guaranteeing privacy of its records implementing Differential Privacy as introduced in [10] and refined in [11]. In this context, the generation of synthetic data serves as a common practice to address privacy concerns and adhere to regulations, enabling researchers to employ a substitute for sensitive data. Consequently, synthetic data must meet qualitative and quantitative criteria to sufficiently support scientists in conducting Exploratory Data Analysis (EDA) and subsequently training Machine Learning (ML) models, while ensuring comparable outcomes to those obtained using the original dataset. A quality algorithm should provide as output data that is qualitatively good enough to keep similar performances when used to train a machine learning models if compared to the original dataset.

This study focuses on the comparison of a specific set of interrelated generative methods that adhere to the principles and characteristics of Differential Privacy. The selection of these algorithms is guided by specific criteria to ensure both a fair and rigorous evaluation of individual methods as well as a comprehensive comparison across the group. To facilitate this comparison, we employ a novel and comprehensive evaluation metric for synthetic data generation, as recently introduced in [12] as to date it represents the most advanced and complete evaluation methodology accepted by the scientific community.

2 Related Work

Differential Privacy is a mathematical framework widely used and well recognized by the academic community [9, 11]. A differentially private algorithm - given its sensitivity to the input - is mathematically guaranteed to inject a specific amount of noise [10] making sure to quantify the privacy of each of the records provided as input given a predefined privacy budget ϵ.

For what concerns classical machine learning generative techniques a conspicuous number of differentially private SDG algorithms were developed. Privelet+ [21] making use of wavelet transforms. In PSD (Private Spatial Decomposition) [6] and [24] authors use Tree-based models to model the distributions of a spacial datasets. Filter Priority [7]; P-HP method [2]; PrivateERM [22]; PrivGene [25]; PrivBayes [23] utilizing Bayesian Networks DPCopula [16] focusing on Gaussian Copula functions as learning model or an improved and parallelized approach using Copula Functions as well in [3] and only recently in [13] authors utilize DP Vine Copula models.

The main issue we do aim to tackle is that in literature authors used different evaluation metrics for estimating the quality of synthetic data and consequentially compare an algorithm to another and often these results are not considering the type of usage of generated data if not queries. In [23] authors use *α-way marginals* namely evaluating counting queries on subsets of attributes. In [5] authors evaluate their work by measuring the distance of each marginal for each attribute between the learned noisy marginals and the original marginals. In [20] authors used multiple classifiers fed with synthetic data, where each classifier predicts a specific attribute, here the metric used is Accuracy, Recall and F1 Score compared with different DP SDG algorithms: [22] and [25] which also uses clustering. While [16] authors evaluate the quality of the synthetic data generated by DPCopula answering random range-count queries and compare results against other methods such as [7] and [2].

Finally the aim of our work is to cover the lack of comprehensiveness of two proposed evaluation frameworks available to date: DPBench [14] which uses only counting queries and [1] using attributes ranges, counts and macro-statistics such as Pearson correlation. Furthermore wo do aim to compare a specific family of algorithms Copula Based, *DAG* (Directed Acyclic Graph) Based and Hybrid version of them.

Our evaluation framework G_ϵ introduced in [12] aims to cover the gap considering three main factors used to compute the quality of output data as a composition of several indicators:

1. *privacy guarantee* (ϵ);
2. the *macro-statistics* between attributes: significant correlation among attributes X_i should be preserved:
3. *data utility* in terms of machine learning performances: similar classification performances.

3 Generative Algorithms Selection

For a matter of focus and in order to perform reasonable comparison between algorithms we've picked a specific set of methods which share properties making them belong to the same set. The selection of the privacy algorithms have been carried based on the following principles.

1. **Differential Privacy:** the generative algorithm must include end-to-end differential privacy with mathematical proof of it.
2. **Tabular Data:** the generative algorithm must be designed to ingest and generate heterogeneous tabular data.
3. **Publication Relevance:** the algorithm must be published in a top conference or journal specialized in data generation and/or privacy.
4. **Code:** authors must have published at least pseudo-code of their implementation or the source-code must be publicly available.
5. **Model:** the algorithm must make use of marginal probabilities and correlation matrix or represent attributes dependence as Directed Acyclic Graphs (DAG).

Algorithm	**Marginal**	**Corr. Matrix**	**DAG Dependence**
NPGauss	✓	✓	X
Gauss	✓	✓	X
Copula-Shirley	✓	✓	✓
DPBayes	✓	X	✓

Table 1. Properties of each algorithm.

We select algorithms that represent data distributions using marginal histograms and/or correlation matrices (which form the basis for copula functions) and/or Directed Acyclic Graphs (DAG) of marginal histograms to model attributes dependence. Specifically, the algorithms considered in our evaluation include *NPGauss* and *Gauss*, which utilize marginal distributions and correlation matrices (Gaussian Copulae). Copula Shirley incorporates marginal histograms, correlation matrices and a tree structure (DAG). Lastly, PrivBayes employs marginal histograms and a DAG to model attributes dependence. Table 1 provides a summary of the key characteristics of each algorithm.

3.1 PrivBayes

PrivBayes [23] is a differential privacy method for disclosing high-dimensional data. It creates a Bayesian Network (namely a DAG) N from a dataset D, which serves as a model of the correlations between attributes in D, and an approximation of the distributions in D using a set P of low-dimensional marginals. Then, PrivBayes introduces noise into each marginal in P to ensure differential privacy and uses the noisy marginals and the Bayesian network to construct an approximation of the data distribution in D. Finally, PrivBayes takes samples from the approximate distribution to create a synthetic dataset. By injecting noise into the low-dimensional marginals in P instead of the high-dimensional dataset D, PrivBayes overcomes the well-known curse of dimensionality issue. PrivBayes uses both low-dimensional marginal probabilities and DAG dependence by nature.

3.2 DPCopula and Gaussian

DPCopula [16] is a collection of techniques for generating differentially private synthetic data using Copula functions for multi-dimensional data. The method works by computing a differentially private copula function from which synthetic data can be sampled. Copula functions are used to describe the dependence between multivariate random vectors and enable the construction of the multivariate joint distribution using one-dimensional marginal distributions. The authors propose two methods for estimating the parameters of the copula functions with differential privacy: maximum likelihood estimation and Kendall's τ correlation estimation (**NPGauss**). Additionally, the authors provide an improved version of the algorithm that aggregates low-cardinality attributes to overcome the degradation performances on those (**Gauss**) through dataset partitioning.

3.3 Copula-Schirley (Vine)

A vine copula is a family of copulas used to model dependencies between variables in high-dimensional data. The term "vine" is used to describe the tree-like structure used to represent the dependence structure between the variables in the copula. This structure is typically represented as a directed acyclic graph (DAG), with nodes representing variables and edges representing the direction of dependence between them. COPULA-SHIRLEY, presented in [13], is a differentially private approach for synthesizing data using vine copulas with differential privacy training. COPULA-SHIRLEY is an interpretable model that can be applied to heterogeneous types of data while maintaining utility. To overcome the curse of dimensionality, COPULA-SHIRLEY uses a set of bi-variate copulas interconnected by a tree-like structure (DAG) to model dependencies. Each node in the DAG represents a bi-variate copula, and the edges between the nodes represent the direction of the dependence between the variables similarly to Bayesian Networks.

4 Evaluation Framework

Within this context, the evaluation framework is similar to the one introduced in [12], but it includes more data utility metrics and machine learning models. The evaluation framework is a combined metric that considers privacy guarantee, macro-statistics, and data utility:

$$G_\epsilon = \alpha\mu(D_s, D_p) + \beta\delta(D_s, D_p) \tag{1}$$

where α and β are weights that the scientist defines to determine the importance of the two metrics, and ϵ represents the privacy budget value fed into the algorithms as any DP algorithm requires.

4.1 Macro Statistics

The proposed macro-statistics measure μ is computed using ϕ_k introduced in [4], as it represents a practical correlation coefficient for heterogeneous datasets where m denotes the number of attributes of the dataset.

$$\mu(D_s, D_p) = \frac{\|\phi_k(D_s) - \phi_k(D_p)\|_2}{m(m-1)/2} \tag{2}$$

In equation 2 we do compute the L_2 norm of the difference of the correlation matrices ϕ_k computed on both private and synthetic datasets divided by the number of elements of an upper triangular matrix (having dimension $m \times m$) due to the symmetric nature of ϕ_k.

4.2 Data Utility

The data utility measure δ is calculated as

$$\delta(D_s, D_p) = \frac{1}{mKL} \sum_{i=1}^{m} \sum_{k=1}^{K} \sum_{l=1}^{L} \|acc^l(M^k_{X_i,D_s}) - acc^l(M^k_{X_i,D_p})\|_2 \tag{3}$$

Where as described more in depth in [12] m is the number of Machine Learning Tasks (one per attribute in [12]), K is the number of different Machine Learning Models and L is the total number of different Accuracy Scores which can be any metric used to evaluate machine learning tasks. It is important to note that M can be any machine learning task which is compatible with the nature of the target attribute in question.

5 Experimental Setup and Results

We consider four datasets with different characteristics, record sizes, and attribute types, all related to credit and financial status. The *Default of Credit Card Clients* (Default Credit) [15] dataset mostly consists of numerical attributes. The *Adults Census* (Adults) [19] and *Credit Approval* (CRX) [18] datasets are mostly composed of categorical attributes, but differ in their sizes as shown in table 2. The *Financial Services* (Fin Services) [17] dataset is also related to finance, but is distinct from the other three datasets in terms of its characteristics. Most of these datasets include a classification label. The first three datasets can be found in the UCI machine learning repository [8] while *Financial Services* comes from OTP Bank which is lacking classification labels.

We have used the following settings: $K = 3$ (*SVC, Logistic Regression and Decision Trees*) and $L = 3$ in equation (3), $\alpha = 1$ and $\beta = 1$ in equation (1). In this context in the data utility metric δ the parameter $m = 1$ and it refers to the specific target class of the dataset in question. While the values of the privacy budget ϵ (parameter of the SDG algorithms) the following values have been selected: (0.05, 0.1, 0.2, 0.4, 0.8, 1.6).

All the classifiers' hyper-parameters are the default ones as we used *Scikit-Learn 0.24.2*. All the algorithms have been run using a computer equipped with an Intel *CPU i7-7500U@2.70GHz* and *16GB RAM DDR4*. All the ML tasks have been deployed using Python's *3.5.10Scikit-Learn 0.22.2* and the average score of three runs have been accounted for each accuracy metric: Accuracy, Recall and F1 Score. The values of degree of PrivBayes network has been fixed to 2.

5.1 G Score

Taking into account the overall score G_ϵ it is observable that the overall behavior at varying of ϵ it's consistent to all the methods. This behavior most probably is due to the fact that the tested methods belong to the same set of algorithms - sharing at least partially the same theoretical foundations. This factor not only validates our results but also enforces previous findings regarding this family of algorithms.

In general it is possible to observe that no algorithm clearly dominates as in figures 1 and 2. But COPULA-SHIRLEY (Vine) and DPCopula Hybrid (Gauss) tend to have similar results both on curve shape/convergence and score values, this is confirmed by both members of equation (1).

The overall G_ϵ score for varying ϵ can be observed in figures 1 and 2. While for the ML performances the data utility term δ in equation 3 over different values of ϵ can be observed in figures 3 and 4. At higher values of ϵ both methods show to perform more reliably than the PrivBayes or NPGauss as they do converge more steadily than the others.

On the other end PrivBayes and NPGauss look similar on their convergence but for the Adults dataset. Further research lead us to the conclusion that Copula

Gauss in general tends to miss-generate low cardinality attribute values and in our experimental setting the Adults dataset is the dataset with the most of those.

It is worth to note that all of the algorithms had their worst performances on the same dataset, namely the Default Credit, this dataset has mostly only numerical attributes except for the target class. This outcome had been imputed to the fact that correlations among attributes are not linear and most of the models used to validate the data utility can't really capture non-linear relations between features, furthermore also the correlation coefficient used for macro-statistics reduces to Pearson's correlation (linear correlation) when evaluating two numerical features.

Data Set	Categorical	Numerical	N. Attributes	N. Records	N. Classes
Default Credit	1	23	24	30000	2
Adults	9	5	14	32561	2
CRX	10	5	15	653	2
Fin Services	3	11	14	4122	0

Table 2. Properties of each dataset.

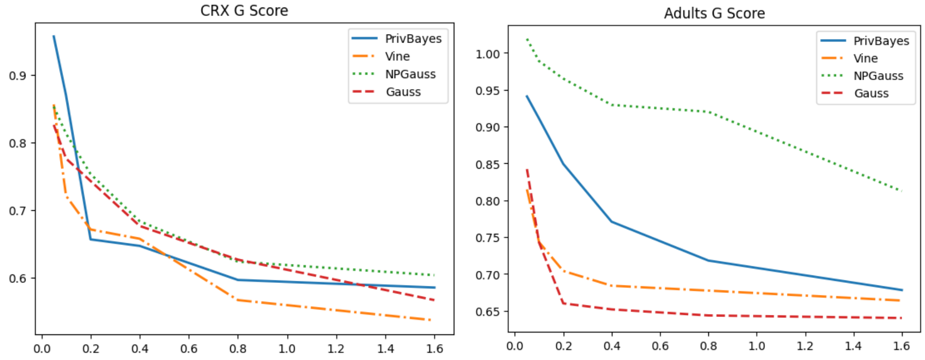

Fig. 1. Values of G_ϵ over the four algorithms deployed on CRX and Adults datasets (a lower value G_ϵ is better as it means that the synthetic dataset is similar to the original private one). The x axes represent values of ϵ while the y axes represent G_ϵ.

5.2 Accuracy Metrics

Regarding δ, it is noticeable that the behavior of all methods is quite similar when varying ϵ. At higher values of ϵ, all methods perform more reliably, as the average of the three accuracy metrics used visibly increases (figures 3 and

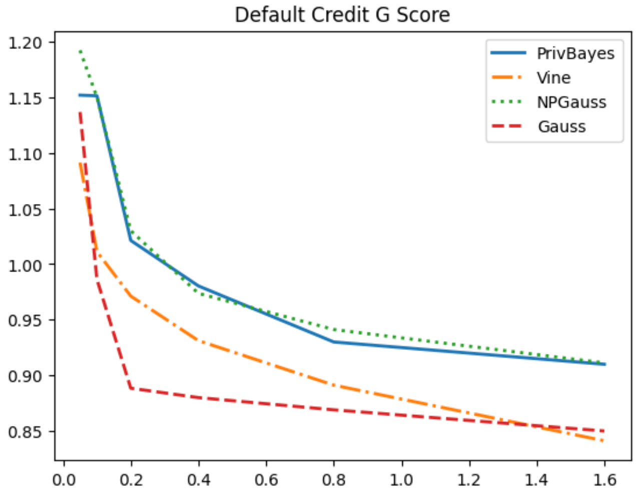

Fig. 2. Values of G_ϵ over Default Credit dataset (a lower value G_ϵ is better as it means that the synthetic dataset is similar to the original one). The x axis represents values of ϵ while the y axis represents G_ϵ.

4). Once again, PrivBayes and NPGauss look similar, but their convergence regarding this metric is different, with PrivBayes resulting in the best performing method and NPGauss being the worst. Further analysis led us to the conclusion that Copula Gauss tends to misgenerate low cardinality attributes in general, and this might occur when binning continuous variables into too small bins.

All of the algorithms had their worst performances on the same dataset, namely the Default Credit. This dataset mostly contains numerical attributes except for the target class. This outcome had been attributed to the fact that correlations among attributes are not linear, and most of the models used to validate the data utility cannot capture non-linear relations between features.

5.3 Macro-Statistics

Along this study we've recognized that the overall preservation of macro-statistics term μ for different values of ϵ it's fundamental both for practicing *Exploratory Data Analysis* (EDA) and eventually for *Machine Learning* (ML) performances and benchmarks. As expected we've observed that in terms of macro-statistics defined in equation (2) as the value of ϵ decreases correlations get weaker and weaker.

At the same time given a fixed value of ϵ (which defines a lower-bound of privacy thus a lower-bound of noise injection but not necessarily an upper-bound) each algorithm can behave differently depending on its design and thus to its internal representation of the attributes distributions/relations possibly leading to a different magnitude of the sampling error figure 5 gives an illustration of

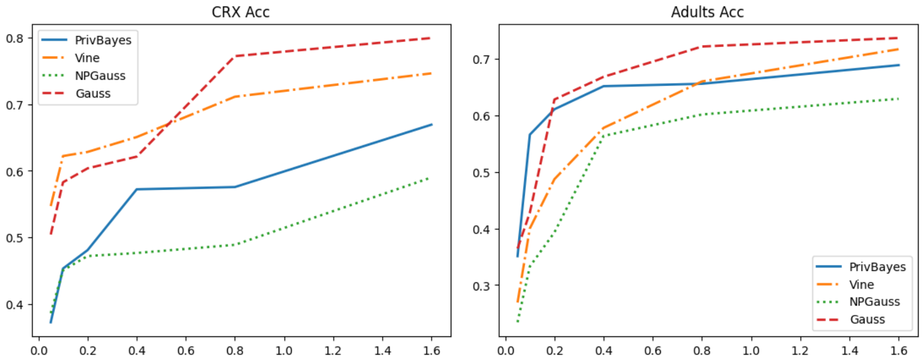

Fig. 3. Average values of Acc_ϵ over the four algorithms deployed on CRX and Adults datasets (higher Acc is better as its values get closer to the values achieved on the original private datasets). The x axes represent values of ϵ while the y axes represent the average Acc_ϵ.

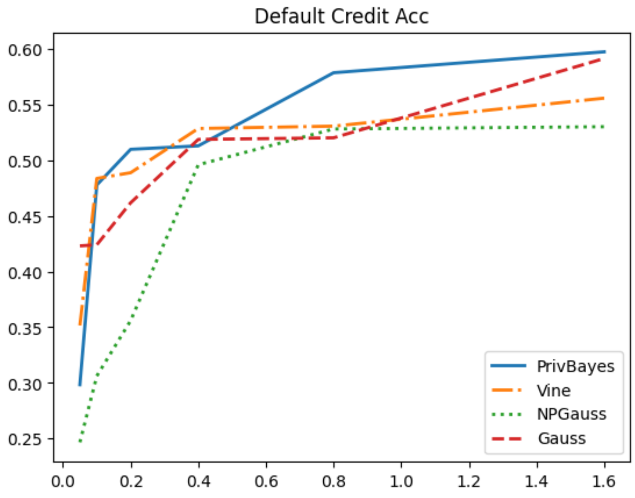

Fig. 4. Average values of Acc_ϵ over Default Credit dataset (higher Acc is better as its values get closer to the values achieved on the original private datasets). The x axes represent values of ϵ while the y axes represent the average Acc_ϵ.

this outcome. In Fig. 5 it is possible to observe the four correlation matrices (NPGauss has been omitted because the results are very similar to PrivBayes). It is possible to note that for the same dataset (Default Credit) the pairwise values of ϕ_k appear to be more unstable as we look clockwise from the ground truth onward Vine tends to slightly weaken all the main correlations but it's still possible to observe the main structure of the matrix (though some "new stronger" correlations seem to be created). Gauss Copula maintain an overall weaker structure with several new correlations. While PrivBayes preserves mainly the correlations of the categorical attributes. This latter behavior might be due to the splitting of continuous values performed by PrivBayes as per its design.

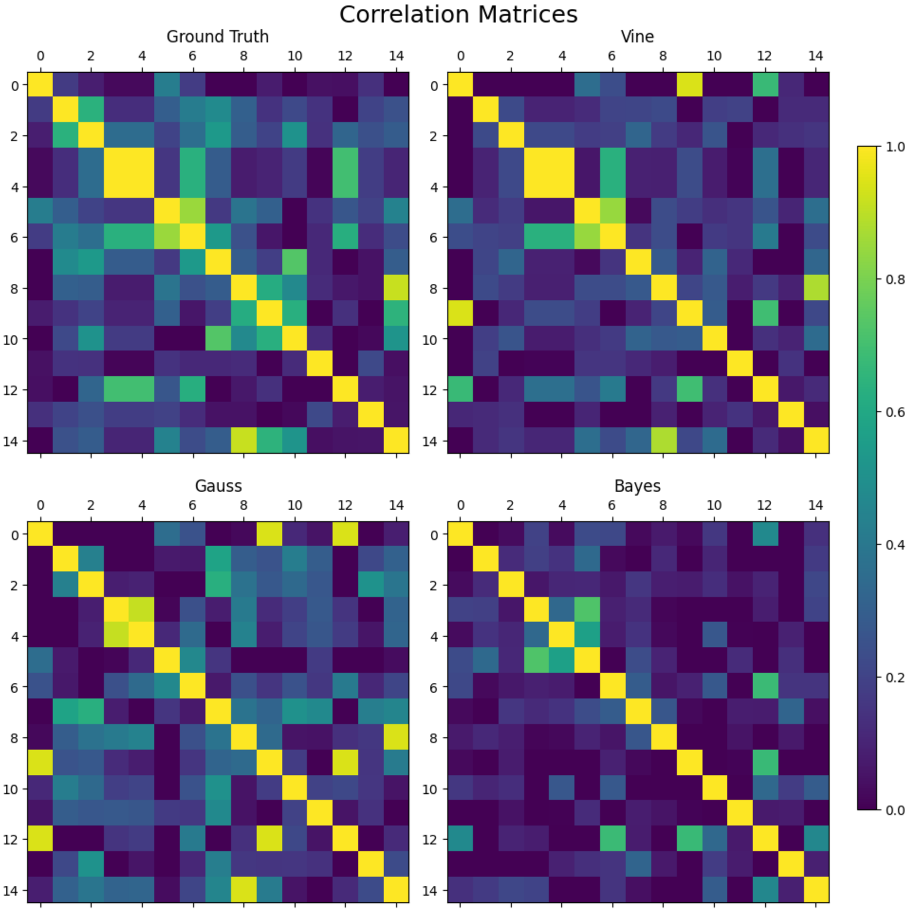

Fig. 5. Correlation matrices over Fin Services dataset given a fixed value of $\epsilon = 0.8$, x and y axes are the attributes while the values of ϕ_k between each attribute are represented by the colored cells of the matrix.

6 Conclusions and Future Work

We have performed an empirical evaluation study on DP SDG. Specifically, we have performed a benchmark of differentially-private synthetic data generation

(DP SDG) algorithms for tabular data with heterogeneous attributes in the field of finance and credit. We evaluated a specific set of algorithms that are related, and found that their overall performances confirm this. Our evaluation considered their utility in terms of machine learning and macro-statistics, such as pairwise correlations in a balanced setup. However, we found that on numeric data, these algorithms tend to be weak or require further pre-processing to improve their performances. Our research raises important questions for future research, including exploring different binning techniques and/or encoding methods as a form of data pre-processing, and potentially developing a framework to select the best algorithms for a given dataset.

Acknowledgments

We are thankful to OTP Bank for sharing their anonymized data.

References

1. Sdgym. https://github.com/sdv-dev/SDGym
2. Acs, G., Castelluccia, C., Chen, R.: Differentially private histogram publishing through lossy compression. In: 2012 IEEE 12th International Conference on Data Mining. pp. 1–10. IEEE (2012)
3. Asghar, H.J., Ding, M., Rakotoarivelo, T., Mrabet, S., Kaafar, D.: Differentially private release of datasets using gaussian copula. Journal of Privacy and Confidentiality **10**(2) (Jun 2020)
4. Baak, M., Koopman, R., Snoek, H., Klous, S.: A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. Computational Statistics & Data Analysis **152**, 107043 (2020)
5. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 273–282 (2007)
6. Cormode, G., Procopiuc, C., Srivastava, D., Shen, E., Yu, T.: Differentially private spatial decompositions. In: 2012 IEEE 28th International Conference on Data Engineering. pp. 20–31. IEEE (2012)
7. Cormode, G., Procopiuc, C., Srivastava, D., Tran, T.T.: Differentially private summaries for sparse data. In: Proceedings of the 15th International Conference on Database Theory. pp. 299–311 (2012)
8. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
9. Dwork, C.: Differential privacy. In: Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33. pp. 1–12. Springer (2006)
10. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. pp. 265–284. Springer (2006)
11. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3-4), 211–407 (2014)

12. Galloni, A., Lendák, I., Horváth, T.: A novel evaluation metric for synthetic data generation. In: Analide, C., Novais, P., Camacho, D., Yin, H. (eds.) Intelligent Data Engineering and Automated Learning – IDEAL 2020. pp. 25–34. Springer International Publishing, Cham (2020)
13. Gambs, S., Ladouceur, F., Laurent, A., Roy-Gaumond, A.: Growing synthetic data through differentially-private vine copulas. Proceedings on Privacy Enhancing Technologies **2021**(3), 122–141 (2021)
14. Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y., Zhang, D.: Principled evaluation of differentially private algorithms using dpbench. In: Proceedings of the 2016 International Conference on Management of Data. pp. 139–154 (2016)
15. I-Cheng Yeh: Default of credit card clients data, https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients, [Online; accessed 31-March-2023]
16. Li, H., Xiong, L., Jiang, X.: Differentially private synthesization of multidimensional data using copula functions. In: Advances in database technology: proceedings. International conference on extending database technology. vol. 2014, p. 475. NIH Public Access (2014)
17. OTP Bank: Financial services, proprietary Data Set
18. Quinlan: Credit approval data set, https://archive.ics.uci.edu/ml/datasets/Credit+Approval, [Online; accessed 31-March-2023]
19. Ronny Kohavi and Barry Becker: Adult data set, https://archive.ics.uci.edu/ml/datasets/adult, [Online; accessed 31-March-2023]
20. Tsybakov, A.B.: Introduction to nonparametric estimation. Springer Science & Business Media (2008)
21. Xiao, X., Wang, G., Gehrke, J.: Differential privacy via wavelet transforms. IEEE Transactions on knowledge and data engineering **23**(8), 1200–1214 (2010)
22. Zhang, J., Zheng, K., Mou, W., Wang, L.: Efficient private erm for smooth objectives. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. p. 3922–3928. IJCAI'17, AAAI Press (2017)
23. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: Privbayes: Private data release via bayesian networks. ACM Transactions on Database Systems (TODS) **42**(4), 1–41 (2017)
24. Zhang, J., Xiao, X., Xie, X.: Privtree: A differentially private algorithm for hierarchical decompositions. In: Proceedings of the 2016 international conference on management of data. pp. 155–170 (2016)
25. Zhang, J., Xiao, X., Yang, Y., Zhang, Z., Winslett, M.: Privgene: differentially private model fitting using genetic algorithms. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 665–676 (2013)