Extending Synthetic Data Evaluation Metrics

Andrea Galloni, Imre Lendák, Tomáš Horváth
ELTE – Eötvös Loránd University,
Faculty of Informatics,
Institute of Industry-Academia Innovation,
Department of Data Science and Engineering,
Budapest, Hungary
Email: {*andrea.galloni, lendak, tomas.horvath*}*@inf.elte.hu*

Abstract—**This paper explores the impact of the evolving evaluation metrics used to evaluate synthetic data and the related generative mechanism in the context of Differential Privacy (DP). Specifically, the authors provide an overview of current synthetic data evaluation methodologies and most commonly used metrics, identifying improvement possibilities. As the field continues to evolve, the authors also propose an improved version of a previously introduced comprehensive evaluation metric useful to redefine such evaluation frameworks in order to make them more complete and exhaustive. This proposed composite metric offers new ideas and directions for future comprehensive quantitative analysis of differentially private synthetic data and, thus, of their related generators.**

Index Terms—**Synthetic Data, Evaluation Metrics, Data Evaluation Framework, Differential Privacy.**

1. Introduction

With the increasing prevalence of data collection and sharing, the need for privacy protection has become paramount. Differential Privacy (DP), a concept introduced in 2006 [1, 2], has emerged as a now prominent framework for achieving this goal. Within the privacy preserving synthetic data generation context (and not only), DP provides a rigorous mathematical definition of privacy guarantees, which has been adopted by many government and industry organizations [3, 4, 5, 6, 7].

DP has become an active research area, and many differentially private mechanisms have been proposed in the literature. These mechanisms range from simple random noise addition to more sophisticated algorithms, such as the Laplace Mechanism [2], Exponential Mechanism [8] and Matrix Mechanism [9].

One crucial factor in DP is assessing the efficacy of privacy-preserving mechanisms in terms of both privacy and data utility. Although privacy bounds are mathematically defined, various evaluation metrics and frameworks have been proposed to measure the data utility(-loss) resulting from differentially private mechanisms and algorithms implementing them.

In this paper, we aim to survey different evaluation metrics and frameworks for evaluating differentially private synthetic data and their related generative algorithms. We will provide an overview of the existing metrics used to evaluate the privacy against utility of the generated data. We will also discuss their strengths and limitations and highlight some open research directions for the evaluation of DP mechanisms. Finally, we propose an improved evaluation metric capable of overcoming the identified shortcomings of the metrics currently used in the literature.

2. Related Work and Research Gaps

As the importance of DP has grown over time, there has been a marked increase in the number of publications on the topic with diverse evaluation metrics. In [10, 11], the authors present Privelet and DPCopula and make use of a customized version of the *Relative Error* metric, which includes a *Sanity Bound* that is parametric and relative to the number of tuples within the dataset. The authors provide this modification to avoid division by zero on query counts and range-count queries. Additionally, they use the *Squared Error* evaluation metric, however, it might not be suitable for some applications as it penalizes large errors more heavily. An important aspect to note is that both metrics are sensitive to the dataset value ranges and are not bounded metrics.

Authors of [12] present PrivGene and their data utility measure based in terms of machine learning (ML), making use of SVM and Logistic Regression classifiers measuring the misclassification-rate (bounded metric) as well as k-means clustering measuring the intra-cluster variance. Authors of [13] introduce PrivBayes, a private Bayesian Network data generator with similar evaluations using SVM as classifier and Total Variation Distance (TVD) [14]. The metric used for the ML evaluation is Accuracy, which has several limitations and pitfalls: First, it can be misleading when dealing with imbalanced datasets. Second, it does not take into account the different types of errors a classifier can make, which can lead to misleading results. Precision, Recall and/or F1-score might be better metrics able to adapt to imbalanced datasets.

While TVD is a reliable and appropriate metric, it is important to note that the authors of [13] utilize this metric to compare the original empirical marginal distribution with the model's noisy marginals rather than the empirical marginal distribution of the synthetic data. This approach enables the evaluation of the model but it does have some limitations. Specifically, in this case, this metric has been used to measure how the noisy marginals (of the learned noisy model) differ from the original marginals (original attributes' distributions) and does not assess the quality of the sampling process, thus, ignoring the quality of the final output (the synthetic data). In other words, it does not take into account the sampling error, which may vary from model to model depending on its design. Finally, the authors do not present an aggregate metric to rank the algorithms in a unified evaluation score, leaving room for improvements.

In addition to the aforementioned evaluation metrics, a new metric called Error on the Workload Queries has been proposed in [15]. In this metric, the authors randomly generate subsets of attributes and then generate range queries on the possible combinations of their values. As authors noted, this metric is related to TVD and is equal to it in some special cases, since the variation distance is computed over marginals derived from a combination of values of the different attributes. However, this metric has limitations as it relies on random processes (generation of random triplets of attributes), which makes it less reliable compared to deterministic measures. Moreover, this evaluation framework only provides a partial assessment as it does not consider most of the various statistical properties a dataset may have.

Authors of [16] propose a novel mechanism called *Copula-Shirley*, which is based on differentially private vine copulas. In order to assess the quality of the synthetic data generated by the mechanism, the authors employ several evaluation metrics. To evaluate the marginal distributions of each attribute, the widely used *Kolmogorov-Smirnov test* (KS) is utilized to estimate the fidelity of the distributions of the synthetic data with respect to the original data, taking an average over all attributes. Furthermore, the authors evaluate pairwise correlations between attributes using *Spearman's rank correlation*. The resulting score represents the mean absolute difference of the correlation values between the two correlation matrices. To complement the statistical tests, classification tasks are employed to simulate specific use-cases and evaluate the preservation of correlations between attributes in relation to the considered classification task. In this setting, utility is measured by training two classifiers (Gradient Boosting), with the first trained on a training set of the original data, and the second trained on the synthetic data produced by the generative model. Finally, a disjoint test set (separate from the original data) is used to test the two classifiers, and the authors employ the *Matthews Correlation Coefficient* (MCC) to evaluate classifier performances. The authors do not propose a formal methodology or a unique aggregated score for evaluating and ranking the models.

A benchmark study for differentially private generative mechanisms in a structured manner has been recently proposed in [17]. The study entails evaluating the performance of various differentially private algorithms from several perspectives. The authors measure individual attribute distribution similarity using TVD, which calculates the distance between two one-dimensional distributions. To measure pairwise attribute distribution similarity, TVD is calculated for each two-way marginal and then averaged across all attribute pairs. An improved version of *Cramer's V* [18] is utilized to measure pairwise correlation similarity. Additionally, F1 is used as classification score and is computed by training an XGBoost classifier on the synthetic dataset and then making predictions on the original data. Although the evaluation framework is heterogeneous and encompasses various fundamental aspects of a dataset, it does not include a comprehensive aggregate score to rank the evaluated mechanisms.

Authors of [19] proposed a new comprehensive and heterogeneous evaluation metric G_ϵ, based on correlation analysis and ML performances, which, however, does not take into account statistical measures of data distributions nor analyze individual attributes or pairs of attributes, as in case of TVD or KS. As a result, there is still room for improvement to define a more comprehensive and exhaustive metric.

While surveying the literature we've found that synthetic data evaluation methodologies and metrics are markedly heterogeneous and nearly each contribution seems to use its own set of metrics. This makes hard to choose the right algorithm for a given dataset, based on its properties, or to rank the different DP mechanisms.

3. Our Contribution

As previously stated, our contribution in this work is an extension of G_ϵ [19], an (almost) comprehensive evaluation metric, which will be briefly described here, before introducing our comprehensiveness extension to it.

The task of Synthetic Data Generation (SDG) is to generate a synthetic dataset D_s with a pre-defined number n_s of rows and with the same number and type of attributes $X_1, X_2, \cdots, X_m$ as a given original and private dataset D_p with n_p rows, used as the input to the generative mechanism. The attributes (columns) $X_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ refer to n-dimensional vectors having numeric, ordinal or nominal values where n is equal to either n_p or n_s (in our setting $n_p = n_s$).

3.1. G_ϵ – An Almost Comprehensive Evaluation Metric

A comprehensive method G_ϵ is defined in [19] as:

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

where α and β are weights allowing to designate the importance of the micro-statistics similarity μ and the data utility similarity δ, while ϵ is the privacy budget as defined in the DP framework [2]. It is crucial to emphasize that although the evaluation metric does not directly incorporate the value of ϵ, having knowledge of its magnitude is essential for conducting a comprehensive and equitable comparison of differentially private generative algorithms.

3.1.1. Macro-statistics (μ). The macro-statistics term is mainly based on [20], where a practical and versatile correlation coefficient ϕ_k has been introduced as an improvement over different correlation analysis metrics of two variables, as it can be applied consistently across categorical, ordinal, and interval variables. Authors of [20] demonstrate that it is capable of capturing nonlinear dependencies and that it relates to the Pearson's correlation coefficient in specific cases. These characteristics make it particularly useful when correlating variables of mixed types. The authors of [20] also take into account the importance of properly evaluating the statistical significance of correlations, especially when dealing with low statistics samples. The value interval of ϕ_k is bounded to $[0, 1]$ where 0 means no correlation and 1 means perfect correlation.

The proposed overall macro-statistics measure μ between D_s and D_p, both having m attributes $X_1, X_2, \ldots, X_m$ is defined as:

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

where $\phi_k(D)$ refers to the pairwise correlation coefficient matrix of the attributes X_i of dataset D on the input.

3.1.2. Data Utility (δ). Assume the dataset D_s is primarily intended for analytical purposes and can be used for various machine learning tasks. As authors of [19] state, at the time of generating D_s it might not be known which of the attributes $X_1, X_2, \ldots, X_m$ may be used later as a label (target variable). Thus, m different prediction tasks are considered for D_s and a private dataset D_p. The ML models are denoted as $M_{X_i,D}, M_{X_2,D}, \ldots, M_{X_m,D}$, where $M_{X_i,D}$, with $1 \leq i \leq m$, represents a ML model learned on a training subset of D using attributes $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m$ to predict attribute X_i.

For a more general model, K different ML models are allowed to be deployed on both the synthetic (D_s) and private (D_p) datasets. These models are denoted as $M^1_{X_i,D}, M^2_{X_i,D}, \ldots, M^K_{X_i,D}$, where $1 \leq i \leq m$.

The performances of the ML models $M_{X_i,D}$ on D are measured using an arbitrary accuracy measure, denoted as $acc(M_{X_i,D})$, where L different accuracy measures are allowed and are denoted as $acc^1(M^k_{X_i,D}), acc^2(M^k_{X_i,D}), \ldots, acc^L(M^k_{X_i,D})$, where $1 \leq i \leq m$ and $1 \leq k \leq K$.

The proposed overall data utility measure δ between D_s and D_p is computed as follows:

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

3.2. Our Improvements from G_ϵ to G_ϵ^+

In light of the limitations observed in G_ϵ, we propose an enhanced version of the generative mechanism, denoted as G_ϵ^+. Specifically, we advocate for the inclusion of attribute distribution distance metrics, such as the TVD, in the evaluation process. This will provide a more comprehensive and robust evaluation of the synthetic data, covering all relevant aspects of the datasets' properties.

In order to provide a contextualization of our contribution, it is necessary to first elucidate the limitations of G_ϵ. Specifically, authors of [19] did not take into account the range of attributes or the distribution shapes. Although the term δ may eventually, in specific cases, capture some of these factors, the metrics used to evaluate the data are generally not impacted by linear transformations, or are only partially affected.

To illustrate, consider a generic numerical dataset, which comprises mostly of numerical attributes and a binary target class. Applying a linear transformation, such as a translation (i.e., adding a constant to all attribute tuples) while retaining the original classes, the evaluation metric proposed in [19] would not properly capture this change in statistical properties, resulting in a very low score of 0 which would be misleading. While the case of a rotation of the dataset (e.g. using a rotation matrix) the only term affected would be μ (except for 90 degrees rotations and their multiples) while most of the data utility ML algorithms would still fully capture data characteristics leading the δ term to be close to 0. Indeed, in this case as well as similar contexts, G_ϵ would be describing the output as genuine (0 means perfect replication of the data characteristics) while it would not be the case for at least two reasons: First, to define a synthetic realistic dataset, the range of attribute values should be reasonably preserved. Second, in the context of DP, a pure linear or similar transformation would not guarantee privacy due to linkage and probabilistic attacks [21]. Thus, Pearson's correlation and ϕ_k [20] are not significantly affected by linear transformations, leaving room for improvements.

Furthermore, authors of [22] (*micro.*) demonstrate that it is possible to generate synthetic data starting from a pseudo-random matrix while maintaining intact mean, variance, correlation and covariance, by using Cholesky decomposition and then solving a linear system to drive/correct the statistics to their original values. Within this framework, given a regression dataset, the metric proposed in [19] would not be able to capture this change in values or at least the μ term would not be affected leading G_ϵ to a value close to 0. Based on this latter example we advocate against introducing such metrics while extending the evaluation metric (mean, variance, covariance) mainly for two reasons, such that, these metrics are not bounded in a specific interval and, as showed in [22], there might be a certain lack of comprehensiveness.

TVD would be able to compensate for such shortcomings of G_ϵ, thus, including attribute distribution distance metrics into G_ϵ (denoting the refined g_ϵ as G_ϵ^+) improves the evaluation process, making it more reliable and resilient to this kind of transformations and similar methods as in [22]. We've identified that a more comprehensive and heterogeneous evaluation should cover also single-attribute distribution distance. Furthermore, the final score should be composed of bounded metrics. Thus, we do exclude mean, variance and covariance of data evaluation and include the following terms to the unified score:

1) *Macro-Statistics* μ such as correlation analysis;
2) *Data Utility* δ in terms of ML performances if applicable;
3) *Single-Attribute Distribution Distance* (1-TVD) ν.

3.3. Total Variation Distance (TVD)

TVD is defined as the sum of the absolute differences between the probabilities of corresponding events in terms of their two probability distributions. In other words, it is a measure of how much one probability distribution diverges from another. A lower total variation distance indicates a greater similarity between the two distributions.

Definition 1. Formally, let P and Q be two probability distributions over the same sample space. The total variation distance between P and Q is defined as $TVD(P, Q) = \sup_{A\in\mathcal{A}} |P(A) - Q(A)| = \sup_{A\in\mathcal{A}} \left|\int_A (p-q)d\nu\right|$ where $\mathcal{A}$ is any measurable space and A is any subset of $\mathcal{A}$.

As per definition, the value range of *TVD* is bounded to $[0, 1]$ and, worth to note is that, it works in case of both continuous as well as discrete probability distributions [14].

When dealing with discrete PDFs, normalized histograms or marginals $P(x)$ and $Q(x)$, this metric can be computed as half of the *L1-distance* between the two marginals of attribute X_i, where both histograms are treated as probability distributions, i.e.

$$\text{TVD}(P, Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$$

The *Univariate TVD* (ν) of G_ϵ^+ aims to measure how, on average, the distribution of each specific attribute is affected on the generated values (X_i^s), computing this metric against the original (X_i^p), such as

$$\nu(D_p, D_s) = \frac{1}{m}\sum_{i=1}^{m} \text{TVD}(P(X_i^p), Q(X_i^s)) \qquad (4)$$

3.4. The Combined Metric G_ϵ^+

As in [19], we do aim to construct a combined overall metric, such that

$$G_\epsilon^+ = \alpha\mu(D_s, D_p) + \beta\delta(D_s, D_p) + \gamma\nu(D_p, D_s) \qquad (5)$$

where α, β and γ are weights allowing to designate the importance of micro-statistics similarity (μ), data utility similarity (δ) and total variation distance (ν), while ϵ is the privacy budget as defined in the DP framework [2].

4. Experimental Results

Within this experimental framework, due to space constraints, we do present only the most relevant findings. 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 *Scikit-Learn 0.24.1* and *NumPy 1.20.1*.

We carry two different kind of experiments: First, we compare G_ϵ and G_ϵ^+ over a generic numerical bi-variate synthetic dataset to highlight the main differences in their outputs, when applying specific transformations, and the impact of ν on the overall score. Second, we compare the two scores over a real dataset and two DP SDG algorithms, namely, *PrivBayes* and *Copula-Shirley* over the *Adults* dataset. For all the experiments we did set $\alpha = \beta = \gamma = 1$.

4.1. Synthetic

In order to emphasize the need of a probability distribution distance to create a more comprehensive metric, we do carry out the following specific transformations (Table 1 and Figure 1): two kind of rotations, $\pi/2$ and π, a translation of the vector space (we do add a constant to the whole dataset, namely the sum of the averages of the two variables) and, finally, we do apply a nonlinear transformation using the algorithm presented in [22]. For the sake of completeness we do include also the Pearson's correlation coefficient even if it is not used to compute the two metrics as it relates to ϕ_k. Indeed, it is interesting to note that because of the value range of ϕ_k ($[0, 1]$) the rotation affects only Pearson's correlations which results in having the same magnitude but different signs, as ϕ_k gets the same exact values, and the μ member of both G_ϵ and G_ϵ^+ gets to 0. While, for the π rotation, both coefficient converge to the same value, leading to a distance of 0 value for both. For what concerns the ML model utilized in this context we opted for linear regression as it is simple and interpretable. Also, in order to carry an unbiased ML evaluation, the use of bounded metrics is preferable, so we decided to use R^2 as an accuracy metric for the linear regression task over one target variable.

Figure 1 illustrates the nature of the synthetic dataset before and after 90 degrees rotation and gives a visual illustration of the *TVD* of the two variables after the transformation.

It is important to note that the only metric capable of detecting changes in both rotations and translation is the *TVD*. In the case of nonlinear transformations, it is noteworthy that the Pearson's correlation coefficient distance μ becomes null, as stated in [22], as it preserves mean, variance, covariance, and correlation coefficients by construction. However, in this case, the R^2 metric and ϕ_k are capable of capturing the changes in dataset properties compensating ν which gets a lower value if compared to other transformations as in this case the PDFs and the ranges of the variables resembles the original private ones.

4.2. Evaluating DP SDG Mechanisms

For the purpose of testing DP SDG algorithms, we use a classification task on the Adults dataset. We use the following configuration: $K = 3$ (*SVC, Logistic Regression, and Decision Trees*) and $L = 3$ in (3). In this context, in the data utility metric δ, the parameter is $m = 1$, which refers to the specific target class of the dataset (*salary*). We set the *PrivBayes* degree of the network to $n = 3$.

Regarding the privacy budget ϵ (for SDG mechanisms), we have selected the following values: 0.05, 0.1, 0.2, 0.4, 0.8, and 1.6. In this illustrative case, it is worth noting that the added term ν makes the difference in the retention of the original data characteristic more clear, highlighting the difference between the two algorithms making G_ϵ^+ more reliable and complete when evaluating two privacy mechanisms.

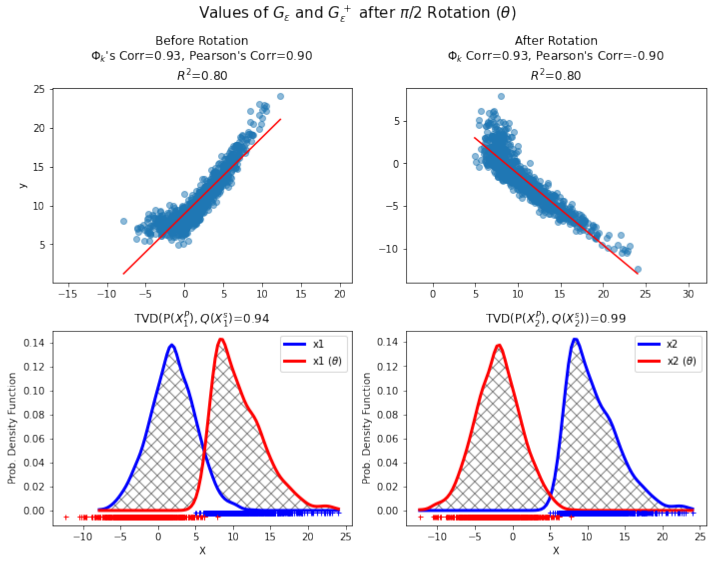

Figure 1. Illustrative example of dataset rotation and the TVD distance.

TABLE 1. METRICS AND SCORES (SYNTHETIC DATASET)

Test.	$\underline{\phi_k}$	C	$\underline{R2}$	RSE	$\underline{TVD}$	G_ϵ	G_ϵ^+
Rot. ($\pi/2$)	0	1.8	0	0	0.97	0	0.97
Rot. (π)	0	0	0	0	0.82	0	0.82
Trans.	0	0	0	0	0.94	0	0.94
Micro.	0.11	0	0.02	0	0.27	0.13	0.4

The underlined metrics are used to compute G_ϵ and G_ϵ^+

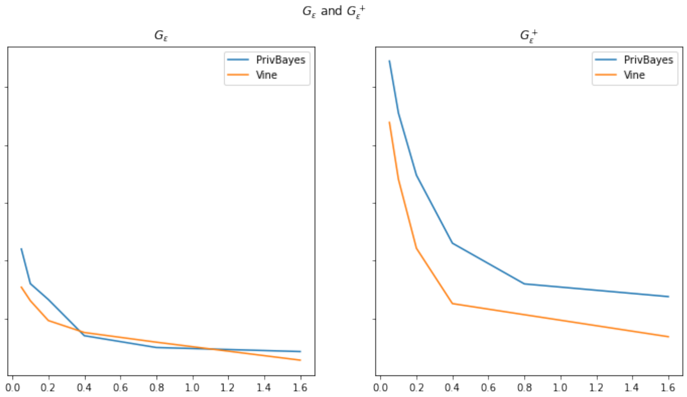

Figure 2. The two scores over the same datasets

5. Conclusions and Future Work

We have proposed an improved version of G_ϵ [19], called G_ϵ^+, which is a composite and more comprehensive synthetic data evaluation metric. Its purpose is to quantitatively measure synthetic data generation mechanisms. It takes into account dataset macro-statistics, data utility, and attribute distribution distances against privacy budget between differentially private synthetic data and the original private data. Our improvement and contribution are supported by theoretical and practical examples. We have tested the introduced evaluation metric against two datasets, comparing two distinct differentially private synthetic data generation algorithms. The results are consistent with previous work, showing that G_ϵ^+ has the potential to become a standard for synthetic data evaluation. As future work, we aim to compare several algorithms against different datasets to assess if there is some discriminant in the data to choose the best privacy mechanisms.

Acknowledgments

"Application Domain Specific Highly Reliable IT Solutions" project has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme. This research was co-funded by EIT Digital Industrial Doctorate and Ericsson Hungary. Authors are thankful to Telekom Innovation Laboratories (T-Labs) and the DigitalTech European Digital Innovation Hub (EDIH) project.

References

[1] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[3] "Understanding differential privacy," https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/differential-privacy.html, accessed: 2023-03-30.

[4] A. Thakurta, A. H. Vyrros, U. S. Vaishampayan, G. Kapoor, J. Freudiger, V. R. Sridhar, and D. Davidson, "Learning new words," https://www.google.com/patents/US9594741, 2017, uS Patent 9,594,741.

[5] ——, "Learning new words," https://www.google.com/patents/US9645998, 2017, uS Patent 9,645,998.

[6] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 2014, pp. 1054–1067.

[7] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275.

[8] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 2007, pp. 94–103.

[9] C. Li, G. Miklau, M. Hay, A. Mcgregor, and V. Rastogi, "The matrix mechanism: Optimizing linear counting queries under differential privacy," *The VLDB Journal*, vol. 24, no. 6, p. 757–781, dec 2015. [Online]. Available: https://doi.org/10.1007/s00778-015-0398-x

[10] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Transactions on knowledge and data engineering*, vol. 23, no. 8, pp. 1200–1214, 2010.

[11] H. Li, L. Xiong, and X. Jiang, "Differentially private synthesization of multi-dimensional data using copula functions," in *Advances in database technology: proceedings. International conference on extending database technology*, vol. 2014. NIH Public Access, 2014, p. 475.

[12] J. Zhang, X. Xiao, Y. Yang, Z. Zhang, and M. Winslett, "Privgene: differentially private model fitting using genetic algorithms," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 665–676.

[13] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, pp. 1–41, 2017.

[14] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[15] R. McKenna, D. Sheldon, and G. Miklau, "Graphical-model based estimation and inference for differential privacy," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4435–4444.

[16] S. Gambs, F. Ladouceur, A. Laurent, and A. Roy-Gaumond, "Growing synthetic data through differentially-private vine copulas," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 3, pp. 122–141, 2021.

[17] Y. Tao, R. McKenna, M. Hay, A. Machanavajjhala, and G. Miklau, "Benchmarking differentially private synthetic data generation algorithms," *arXiv preprint arXiv:2112.09238*, 2021.

[18] W. Bergsma, "A bias-correction for cramér's v and tschuprow's t," *Journal of the Korean Statistical Society*, vol. 42, no. 3, pp. 323–328, 2013.

[19] A. Galloni, I. Lendák, and T. Horváth, "A novel evaluation metric for synthetic data generation," in *Intelligent Data Engineering and Automated Learning–IDEAL 2020: 21st International Conference, Guimaraes, Portugal, November 4–6, 2020, Proceedings, Part II*. Springer, 2020, pp. 25–34.

[20] M. Baak, R. Koopman, H. Snoek, and S. Klous, "A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics," *Computational Statistics & Data Analysis*, vol. 152, p. 107043, 2020.

[21] M. M. Merener, "Theoretical results on de-anonymization via linkage attacks," *Trans. Data Privacy*, vol. 5, no. 2, p. 377–402, aug 2012.

[22] J. M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer, "Fast generation of accurate synthetic microdata," in *Privacy in Statistical Databases*, vol. 3050. Springer, 2004, pp. 298–306.