A Novel Evaluation Metric for Synthetic Data Generation

Andrea Galloni[1(✉)], Imre Lendák[1,3], and Tomáš Horváth[1,2]

[1] Faculty of Informatics, Department of Data Science and Engineering, ELTE – Eötvös Loránd University, Budapest, Hungary
andrea.galloni@inf.elte.hu

[2] Faculty of Science, Institute of Computer Science, Pavol Jozef Šafárik University, Košice, Slovakia
Tomas.Horvath@upjs.sk

[3] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
lendak@uns.ac.rs

Abstract. Differentially private algorithmic synthetic data generation (SDG) solutions take input datasets D_p consisting of sensitive, private data and generate synthetic data D_s with similar qualities. The importance of such solutions is increasing both because more and more people realize how much data is collected about them and used in machine learning contexts, as well as a consequence of newly introduced data privacy regulations, e.g. the EU's General Data Protection Regulation (GDPR). We aim to develop a novel and composite SDG evaluation metric which takes into account macro-statistical dataset similarities and data utility in machine learning tasks against privacy boundaries of the synthetic data. We formalize the mathematical foundations for quantitatively measuring both the statistical similarities and the data utility of synthetic data. We use two well-known datasets containing (potentially) personally identifiable information as inputs (D_p) and existing SDG algorithms PrivBayes and DPGroupFields to generate synthetic data (D_s) based on them. We then test our evaluation metric for different values of privacy budget ϵ. Based on our experiments we conclude that the proposed composite evaluation metric is appropriate for quantitatively measuring the quality of synthetic data generated by different SDG solutions and possesses an expected sensitivity to various privacy budget values.

Keywords: Synthetic data generation · Differential privacy · Evaluation metrics

1 Introduction

The task of synthetic data generation (SDG) tackled in this paper is, given a private dataset $D_p \subset X_1 \times X_2 \times \cdots \times X_m$ of n_p rows and m attributes, to generate a synthetic dataset D_s which has the same number and type of attributes $X_1, X_2, \cdots, X_m$ as D_p and a pre-defined number n_s of rows. The attributes (columns) $X_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ refer to n-dimensional vectors having numeric,

C. Analide et al. (Eds.): IDEAL 2020, LNCS 12490, pp. 25–34, 2020.
https://doi.org/10.1007/978-3-030-62365-4_3

ordinal or nominal values n being equal either to n_p or n_s. The goal is to generate D_s such that it keeps the main statistics and utility of D_p while preserves the individual privacy of its objects (rows). Privacy is especially important when the private data-set holds personally identifiable information (e.g. medical information) and the SDG process must guarantee that it will not be possible to identify any person based on analyzing only the publicly available generated D_s. A well-designed SDG solution generates data with similar statistical characteristics to the private data, e.g. maintains the correlations between inter-related attributes. Synthetic data is often generated as a substitute for the private data to be used in machine learning tasks. Generated data with high utility can be used to train a model, which in turn is expected to have adequate (classification or regression) performance when fed with the private data-set.

A consistent and comprehensive methodology or score for quantitatively measuring and evaluating the quality of the results of SDG is still missing, since at the moment of writing existing techniques only rely on macro-statistics and Machine Learning performances separately. Any such measure would ideally take into consideration more factors, such as *privacy-degree*, *macro-statistics* and *data utility*.

2 Related Work

Differential privacy is a mathematical tool for quantifying and bounding privacy loss [8]. Within the context of statistical disclosure control and cryptography it provides an accurate statistical information about a population while protecting the privacy of each individual within it. A differentially private algorithm holds a series of constraints which are exploited in order to publish information about a statistical database, these constraints limit and quantifies the disclosure of private information of records whose information is present within the database. The most common and widely used approaches to achieve differential privacy are the *Laplace Mechanism* [8] and the *Exponential Mechanism* [15].

Several differentially private generative approaches were proposed by the scientific community, where lately the most relevant models are utilizing Bayesian networks [12] or Copula functions [18]. These are the following[1]: Privelet+ [20]; PSD (Private Spatial Decomposition) [6]; Filter Priority [7]; P-HP method [1]; PrivateERM [21]; PrivGene [23] using genetic algorithms; PrivBayes [22] utilizing Bayesian networks which was extended with a more complete framework in [17] and [11] with a web interface; DPCopula [13] focusing on Gaussian Copula functions as the Synthetic Data Vault (to date not differentially private) [16]; or an improved and parallelized approach using Copula functions [2].

Different authors use different evaluation metrics for assessing the quality of their SDG solutions. In [22] authors evaluate their results making use of α-*way marginals* derived from query counts of subsets of attributes introduced in [4] measuring the accuracy of each marginal by computing the total variation distance [19] between noisy marginals and the original marginals. Furthermore

[1] Since this paper is focusing on evaluation, most popular SDG approaches are just listed here since their detailed description is out of the scope of this paper.

authors trained multiple SVM classifiers over several attributes of the synthetic dataset, where each classifier predicts one attribute in the data based on all other attributes, in this case the metric used is misclassification rate compared against other differentially private generative algorithms such as [21] and [23] which also apply k-means clustering. In [2] authors evaluate the generated data on all one-way marginal and two-way positive conjunction queries and three-way positive conjunction queries. In [13] authors evaluate the utility of the synthetic data generated by DPCopula answering random range-count queries and compare results against other methods.

In [10] (the most similar contribution) authors provide an evaluation of synthetic datasets under utility perspectives related to machine learning, leaving room for improvements. The authors compare two different approaches including differentially private algorithms which are evaluated under machine learning performance tasks over a single target attribute. Our work wants to provide a more comprehensive approach considering also macro-statistics within our evaluation metric and propose it as a standard methodology.

Based on these, is possible to state that at the moment of writing there is no comprehensive and standard evaluation methodology which evaluates multiple characteristics of the synthetic data against original private data in a unified score. We argue that any synthetic data generation (SDG) solution should possess at least the following three characteristics: i) guarantee a measurable amount of privacy, ii) resemble original data macro-statistics such that correlations, ranges, etc. and iii) maintain data utility for real scenarios usage, meaning that the performance of machine learning algorithms should be similar on both the generated and the original data. A good SDG methodology should maintain a good trade-off between privacy and utility.

3 Proposed Solution

Our metric G_ϵ is a composition of several indicators. Indeed, in the context of SDG there are several aspects, which have to be considered, such that:

- *privacy guarantee* (ϵ);
- the *macro-statistics* between attributes: significant correlation among attributes X_i has to be maintained;
- *data utility* in terms of machine learning performances: we would like to have similar classification performances in terms of accuracy when deploying the same algorithm over the private data-set and the original one.

Privacy Guarantee

The privacy guarantee is a parameter to quantify the privacy budget and it is direct consequence of ϵ-*differentially private* mechanism definition introduced in [8] and well formalized in [9]. Namely epsilon is a guaranteed boundary of privacy loss.

Definition 1 (Differential Privacy (DP) [8]). *A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|X|}$ is (ϵ)-differentially private if for all $\mathcal{S} \subseteq$ Range $(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$*

$$\Pr[\mathcal{M}(D_p) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D_s) \in \mathcal{S}]$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$ we say that $\mathcal{M}$ is ϵ-differentially private; thus the following privacy boundary is guaranteed:

$$log(\Pr[\mathcal{M}(D_p) \in \mathcal{S}]) - \log(\Pr[\mathcal{M}(D_s) \in \mathcal{S}]) \leq \epsilon \tag{1}$$

The term ϵ represents the so called *privacy budget* and it is a parameter of any differentially private mechanism. As its value decreases the more privacy is guaranteed through the injection of random noise when learning the probability distributions of the data-points; it is the direct measure of privacy boundary within the context of differential privacy. Thus for achieving a fair and consistent evaluation of SDG methods ϵ has to hold the same fixed value among those mechanisms to be compared at time of model creation and/or data generation.[2]

Macro-statistics

In [3], authors provide a new and practical correlation coefficient ϕ_k. It is based on several refinements to Pearson's hypothesis test of independence of two variables which works consistently between categorical, ordinal and interval variables. It also captures non-linear dependency. Moreover, it reverts to the Pearson correlation coefficient in case of a bi-variate normal input distribution. These are useful features when studying the correlation between variables with mixed types. Particular emphasis is paid to the proper evaluation of statistical significance of correlations and to the interpretation of variable relationships in a contingency table, in particular in case of low statistics samples and significant dependencies.

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

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

Data Utility

D_s is intended to be used mostly for analytic purposes on which various machine learning (ML) tasks might be performed.

Since, in the time of generation of D_s we can not be sure which of the attributes from $X_1, X_2, \ldots, X_m$ would serve as labels in the future, we consider m different prediction tasks on D_s and D_p, respectively. The corresponding models are denoted as $M_{X_i,D}, M_{X_2,D}, \ldots, M_{X_m,D}$. Here, $M_{X_i,D}$, with $1 \leq i \leq m$, denotes a ML model learned/optimized using the data D (D_p or D_s) using the attributes $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m$ to predict the attribute X_i. Within our experimental framework the machine learning task is classification over categorical attributes.

[2] The important role of ϵ in DP justifies its presence as subscript of G in our evaluation metric definition since we evaluate G at varying of ϵ.

It is important that different classes/types of ML algorithms should be used due to their different biases. Thus, for a more generic model, we allow K different ML models deployed over D, i.e. the synthetic (D_s) and the private (D_p) dataset. We denote these models by $M^1_{X_i,D}, M^2_{X_i,D}, \ldots, M^K_{X_i,D}$, where $1 \leq i \leq m$.

Let $acc(M_{X_i,D})$ denote the performance of $M_{X_i,D}$ measured on D (D_p or D_s). acc can be an arbitrary accuracy measure such that miss-classification rate or AUC, to name a few. For a more generic model we allow L different accuracy measures which will be 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 will be computed as

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

The Combined Metric

Our proposed formula for a combined evaluation metric considering privacy-guarantee, macro-statistics and data utility is defined as

$$G_\epsilon = \alpha\mu(D_s, D_p) + \beta\delta(D_s, D_p) \tag{4}$$

where α and β are weights allowing the user to define the importance of micro-statistics similarity and data utility similarity, while ϵ represents the value fed to the SDG algorithm while implementing DP while building the model.

4 Experimental Setup and Results

In order to evaluate our method we've deployed *PrivBayes* developed in *Python3.7* [22] based on *Bayesian Networks* and *DPGroupFields* mentioned in [5] and among the winners of *Differential Privacy Synthetic Data Challenge 2019* developed in *Java* and based on histogram sampling techniques (both algorithms are publicly available on *github.com*). Since we expect a wider application in real use-cases of DP SDG techniques by the industry, the selection of the two datasets is contextual to the most probable fields of application: indeed both our datasets do contain information about individuals in two main areas: *healthcare data* and *census data*. Namely *diabetes* is a well known dataset holding 8 real attributes and a categorical-binary one (for classification) for a total of 9 attributes and holding 768 records and the *adults* data-set composed by 14 numerical and categorical attributes (majority) and a categorical-binary one (for classification). It is important to mention that at generation time the same number of records of the original databases have been generated, namely: $n_s == n_p$ for all experiments. All the algorithms have been run using a computer equiped with an Intel *CPU i7-7500U@2.70 GHz* and *16 GB RAM DDR4*. All the ML tasks have been deployed using Python's *Scikit-Learn 0.22.2*. For a matter of brevity we're going to show the most relevant results.

For sake of simplicity, we have used the following settings: $K = 1$ and $L = 1$ in Eq. (3), $\alpha = 1$ and $\beta = 1$ in Eq. (4). While for what concerns the values of the privacy budget ϵ (parameter of the SDG routines at generation time) the following values have been selected: (2.0, 1.5, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.01). Those values have been chosen taking into consideration the public literature describing experiments and this value interval represent the most common setup, namely we've extended the ranges proposed within [22] and [13]. Our K task is going to be classification over all the categorical attributes ($K = 1$) of each dataset solved through the well known SVM algorithm with *RBF kernel* and $C = 1$ (regularization parameter) and $\gamma = 1/(m * variance(X))$ over the categorical attributes (default value of the SKLearn SVC classifier). In case of categorical target attributes. For what concerns L in our settings we set its value is 1 and is going to be *miss-classification rate* over categorical attributes.

4.1 Macro-statistics μ

In our first experiment our goal was to measure the effects of the chosen privacy budget on the macro statistics $\phi(D_s)$ against $\phi(D_p)$. Given the random nature at the root of the generative algorithms for each value of ϵ we've generated three different synthetic D_s for each input datasets both making use of PrivBayes and DPFieldGroups, then we've computed the average values of μ for each value of ϵ as plotted in Fig. 1. As expected the distance in terms of macro-statistics defined in Eq. 2 grows as the magnitude of the matrices of the difference among correlation coefficients defined as ϕ in 2 grows. Generally, as expected, we might observe that δ tends to grow at decreasing of ϵ (differential privacy budget), indeed at decreasing ϵ more noise is introduced within the learned model thus correlation coefficients tends to differ more and more between the original dataset and the synthetic ones. In Fig. 1 it is possible to observe the behaviour of the δ term over ϵ against two datasets (Adults and Diabetes). It is possible to note that for Diabetes the computed values of μ appear more unstable if compared to the ones produced by the Adults dataset, This behavior is due to the splitting of continuous values performed by PrivBayes. Still the value range looks acceptable thus in this case the values of δ will play a decisive role when computing G_ϵ (at same values of α and β).

4.2 Data Utility δ

Also the δ factor as expected holds a similar behavior as μ but with a slightly different magnitude this characteristic justifies the presence of the two constants α and β in Eq. 4. Also in this case over our experimental setup we've run three times the data generation algorithm *PrivBayes* and plotted the average of the measures of δ for each value of ϵ. The values of δ are the result of repeating ten times the same machine learning task (classification in our setup) over each attribute, randomly selecting training and testing records with a ratio of 0.2 for testing. As expected the measure in terms of data utility defined in Eq. 3 tends to grow. Generally, it is possible to observe that values of δ tends to grow at

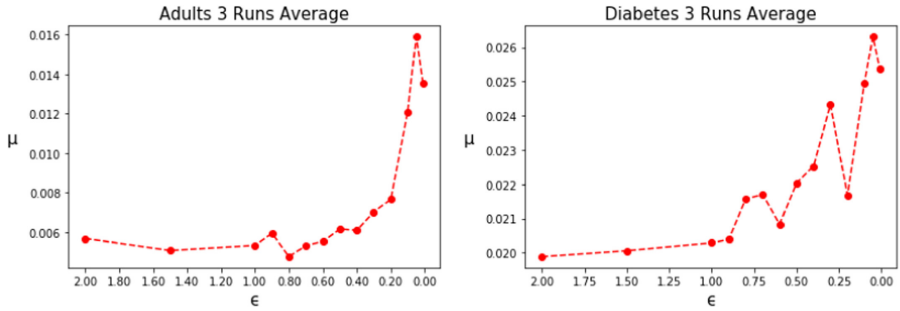

Fig. 1. Values of μ against the two datasets Adults and Diabetes using PrivBayes synthetic data generation method.

decreasing of ϵ (differential privacy budget), indeed at decreasing ϵ more noise is introduced and performances in terms of accuracy tends to differ. In Fig. 2 it is possible to observe the behaviour of the δ term over ϵ against two datasets (Adults and Diabetes). It is important to note that within this setup we can observe that values of δ over ϵ tend to be more stable holding a more clear trend in the case of *adults* (where the majority of the attributes is categorical) when compared to the *Diabetes* graph in which the majority of attributes is numerical. This outcome is due to the splitting of the continuous values within the Bayesian Networks model. This observation represents an important/insight for the user, indeed our metric could suggest the scientist to alter (augment) only the number of splits for a continuous or several continuous variables *obtaining better scoring results without altering the privacy budget magnitude* ϵ. In this case the values of δ grow *"faster"* than μ for the same dataset, thus this factor could dominate smaller values of ϵ.

4.3 Composite Measure G

Finally we calculated the values of composite measure G defined by Eq. 4 for $\alpha = 1$ and $\beta = 1$. We find that the evaluation method provides coherent results. Figure 4 shows results of our experiments and G_ϵ for the Adults dataset.

Within this setup we can note that at the same privacy budget ϵ PrivBayes clearly better preserves data-utility (lower values of G) when compared to DPFieldGroups. This is due to the fact that histogram sampling (DPFieldGroups) performs worse when applied over dataset holding a double digit number of attributes. Thus, we state that our method reflects and confirms earlier literature results [13,14,22]. Figure 3, instead, shows results of G_ϵ regarding just PrivBayes since DPFieldGroups has not been designed to handle attributes holding floating numbers (not integers). Also in this case the effect of continuous attributes splitting has a noticeable impact on the stability of the score.

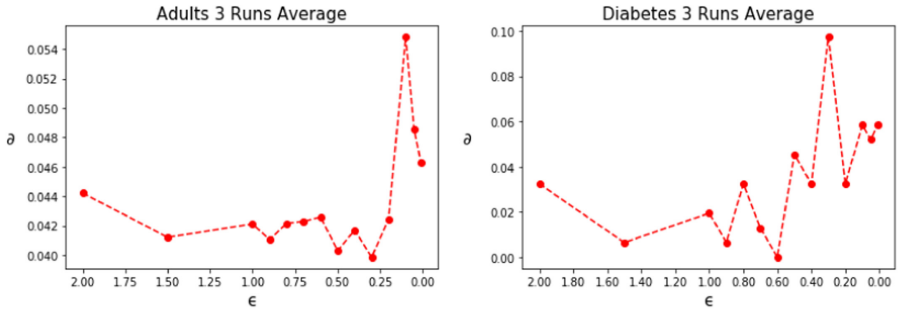

Fig. 2. Values of δ against the two datasets Adults and Diabetes using PrivBayes for both as synthetic data generation method. Values of δ tend to grow *"faster"* due to the nature of PrivBayes which splits continuous intervals.

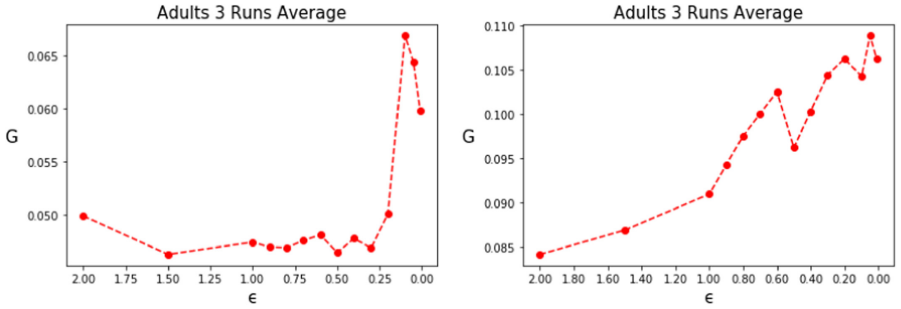

Fig. 3. Values of G_ϵ over the two algorithms PrivBayes (left) and DPFieldGroups (right) deployed on Adults dataset. Within this setup PrivBayes clearly keeps a better data utility over varying ϵ (lower G).

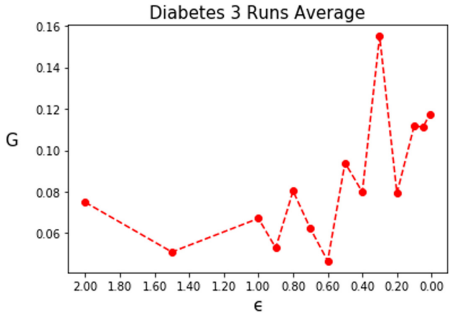

Fig. 4. Values of G_ϵ over Diabetes dataset using PrivBayes as generative algorithm.

5 Conclusions and Future Work

We propose a novel composite and comprehensive evaluation metric G_ϵ for quantitatively measuring synthetic data generation solutions. The metric takes into account dataset similarities and data utility, against privacy budget ϵ, between synthetically generated data and the original data. We test the introduced evaluation metric against two datasets comparing two different differentially private synthetic data generation algorithms. The results are consistent with literature and will open the path for further investigation and possibly it will be used as a base-model and standardized methodology to assess and evaluate the trade-off between privacy and data utility within the context of differentially private synthetic data generation.

As for the next steps there is room for testing this evaluation metric against several different DP SDG techniques, possibly find an empirical way to set the best values of α and β based on the objective of the synthetic data generation task and the properties of the dataset.

Acknowledgments. This research was co-funded by EIT Digital Industrial Doctorate and Ericsson Hungary. Project no. ED_18-1-2019-0030 (Application domain specific highly reliable IT solutions subprogramme) has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme funding scheme. We are thankful to Gian Marco Canneori for the fruitful discussions leading to the final mathematical foundations presented in this paper.

References

1. 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)
2. Asghar, H.J., Ding, M., Rakotoarivelo, T., Mrabet, S., Kaafar, D.: Differentially private release of datasets using Gaussian copula. J. Priv. Confidentiality **10**(2) June 2020
3. Baak, M., Koopman, R., Snoek, H., Klous, S.: A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. Comput. Stat. Data Anal. **152**, 107043 (2020)
4. 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)
5. Bowen, C.M., Snoke, J.: Comparative study of differentially private synthetic data algorithms and evaluation standards (2019). arXiv preprint arXiv:1911.12704
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. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
9. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput.Sci. **9**(3–4), 211–407 (2014)
10. Hittmeir, M., Ekelhart, A., Mayer, R.: On the utility of synthetic data: an empirical evaluation on machine learning tasks. In: Proceedings of the 14th International Conference on Availability, Reliability and Security, pp. 1–6 (2019)
11. Howe, B., Stoyanovich, J., Ping, H., Herman, B., Gee, M.: Synthetic data for social good (2017). arXiv preprint arXiv:1710.08874
12. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT press, Cambridge (2009)
13. Li, H., Xiong, L., Jiang, X.: Differentially private synthesization of multi-dimensional data using copula functions. In: Advances in Database Technology: Proceedings. International Conference on Extending Database Technology, vol. 2014, p. 475. NIH Public Access (2014)
14. Li, H., Xiong, L., Zhang, L., Jiang, X.: Dpsynthesizer: differentially private data synthesizer for privacy preserving data sharing. In: Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, vol. 7, p. 1677. NIH Public Access (2014)
15. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), pp. 94–103. IEEE (2007)
16. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 399–410, IEEE (2016)
17. Ping, H., Stoyanovich, J., Howe, B.: Datasynthesizer: privacy-preserving synthetic datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pp. 1–5 (2017)
18. Sklar, A.: mfonctions de répartition à n dimensions et leurs marges, n publ. Inst. Statist. Univ. Paris **8**, 229–231 (1959)
19. Tsybakov, A.B.: Introduction to Nonparametric Estimation. Springer Science & Business Media, Berlin (2008)
20. Xiao, X., Wang, G., Gehrke, J.: Differential privacy via wavelet transforms. IEEE Trans. Knowl. Data Eng. **23**(8), 1200–1214 (2010)
21. 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, IJCAI 2017, pp. 3922–3928. AAAI Press (2017)
22. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: Privbayes: private data release via bayesian networks. ACM Trans. Data. Syst. (TODS) **42**(4), 1–41 (2017)
23. 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)