Eötvös Loránd University

Faculty of Informatics

Evaluating Synthetic Data Generators in the Context of Differential Privacy

Ph.D. Thesis Booklet

Galloni Andrea

Supervisor: Imre Lendák, Ph.D.

Ph.D. School of Computer Science

Head of School: Prof. Zoltán Horvat́h

Ph.D. program of information systems

Head of Program: Prof. András Benczúr

October 2025, Budapest

Abstract

The growing reliance on data-driven technologies across various sectors has heightened concerns about data privacy and the ethical use of sensitive information. In response, privacy-preserving synthetic data generation has gained attention as a solution to comply with regulations like the European Union's GDPR while allowing data-driven innovation. Differential privacy, a key framework in this domain, provides mathematical guarantees against the probability of re-identification of individuals in synthetic data. However, creating synthetic datasets that maintain both utility and privacy is challenging, and there is a notable lack of standardized and comprehensive frameworks for evaluating the quality and utility of synthetic data.

This dissertation aims to develop and validate a robust evaluation framework for synthetic data generated under differential privacy constraints. The primary goal is to create criteria and metrics that assess the trade-offs between data utility and privacy, providing a systematic approach to evaluating the effectiveness of various synthetic data generation techniques. This framework seeks to fill the current gap in literature by measuring how well synthetic data replicates statistical properties and machine learning performance of original datasets while ensuring strong privacy protections.

The proposed framework evaluates synthetic data across multiple dimensions: Macro-Statistics for statistical fidelity, Machine Learning utility for predictive performance, and Probability Distribution Distances for measuring the similarity between synthetic and original data attributes' distributions. The framework is applied to various generative models. Additionally, a fairly new correlation coefficient is used for assessing the preser-

vation of dependencies within heterogeneous datasets. This comprehensive framework is rigorously tested across different datasets and generative methods to validate its effectiveness and generalizability.

The implementation of the evaluation framework provides significant insights into the performance of different synthetic data generation techniques. It reveals that while some models, successfully balance privacy with utility, others struggle to preserve critical data characteristics. The new metrics introduced in this work offer a more nuanced and comprehensive evaluation, capturing subtle but essential aspects of data quality often overlooked in traditional assessments. Comparative studies of various generative models, including copulas and hybrid methods, are conducted, highlighting their strengths and limitations while comparative analysis across multiple datasets highlights the variability in model performances.

This dissertation makes a substantial contribution to privacy-preserving data generation by providing a rigorous, comprehensive evaluation framework that can guide both research and practical applications. The methodologies and findings offer valuable insights for improving synthetic data generation techniques and refining evaluation metrics. By addressing gaps in current evaluation practices, this work supports the viability of synthetic data as an alternative to real data, enabling organizations to innovate while upholding the highest standards of data privacy. The impact extends to policy-making, where better evaluation tools can inform the development of regulations and standards for synthetic data use in privacy-sensitive environments.

Motivation

Synthetic data has emerged as a valuable tool for data scientists and researchers who require access to sensitive or confidential data but face legal or ethical constraints that prohibit its use (e.g.: the GDPR regulation within EU) [17]. Differentially private synthetic data is generated using statistical models capturing the key statistical properties of the original data while obfuscating sensitive information [15, 4]. As such, it offers a viable alternative to using real data, particularly in situations where privacy and/or industrial confidentiality is a concern or there is the need of more quantities of data.

The applications of synthetic data are wide-ranging and span across various domains, including healthcare [1, 5, 13, 20], finance [2, 18], social sciences [6, 19], and cybersecurity [14]. In the healthcare sector, synthetic data has been used to facilitate research on sensitive health-related data. In the financial industry, synthetic data has been used to develop risk models and to assess the impact of policy changes on market trends. In the social sciences, synthetic data has been used to study sensitive issues such as racial disparities in education and health outcomes, without compromising the privacy of individuals. In cybersecurity, synthetic data has been used to test the effectiveness of intrusion detection systems and to simulate cyberattacks beyond privacy concerns.

Given the potential benefits of synthetic data, there has been a growing interest in its use among researchers and data scientists. However, the evaluation of the validity of synthetic data remain a concern, as it is essential to ensure that the synthetic data accurately represents the original data and preserves its key statistical properties. Several methods have been proposed to evaluate synthetic data but at the same time further research is needed to develop more complete models and evaluation frameworks which can facilitate its use in several different contexts. The main open issues are:

1. *Inconsistencies in evaluating the quality of synthetic data have been observed in the current literature.* While synthetic data has become an increasingly valuable tool for data scientists and researchers facing legal or ethical constraints in using sensitive or confidential data, the lack of standardized evaluation methodologies and metrics poses a challenge to effectively assessing the strengths and weaknesses of generative mechanisms.

2. Synthetic data is generated using statistical models which aim to preserve the key statistical properties of the original data while blurring identifiers and/or sensitive information. However, *the accuracy of these models and their ability to replicate the underlying distribution of the original data can vary depending on a range of*

factors, including the domain of the attributes and the complexity of the data, the chosen generative method, and the specific privacy requirements.

3. *The nature of the data and it's composition may introduce constraints* on the type of tests and metrics that the scientist could and should use to carry out a proper evaluation of the generative mechanisms.

4. To evaluate the quality of synthetic data, researchers have developed and used a variety of metrics, ranging from simple measures such as mean squared error to more complex measures such as making use of machine learning algorithms and comparing performances. *However, these methodologies and metrics are often highly specific to the particular use case or research, and there is the lack standardization across different studies.*

5. *The lack of standardization and a unified, comprehensive framework for evaluating synthetic data* complicates the ability to draw meaningful comparisons across studies. This deficiency makes it difficult to identify the most effective generative mechanisms for specific use cases and data characteristics, and hinders the fair ranking of different generative methodologies on a given dataset. Without standardized evaluation criteria, assessing the strengths and weaknesses of various generative methods and making informed decisions about their practical application becomes a significant challenge.

To address the issues outlined above, a standardized framework for evaluating synthetic data needs to be developed. Such a framework could include a set of sound and commonly agreed upon evaluation metrics which are tailored to different use cases and data characteristics. This would ensure that different studies can be compared effectively and that the most effective generative mechanisms for specific use cases can be identified.

In addition, the development of a standardized framework could facilitate the ranking of different generative methodologies in a fair and objective manner. This could be achieved by incorporating a scoring system that considers the strengths and weaknesses of different generative methods and evaluates their performance against a set of agreed upon metrics.

Furthermore, the framework could include guidelines for selecting appropriate generative methods based on the nature of the data and specific use cases. This would help ensure that the most appropriate generative methods are selected for a given dataset, improving the quality of the resulting synthetic data.

Overall, the development of a standardized framework and a unified scoring system for evaluating synthetic data quality would represent a significant step forward in the field of data science and research. By establishing a common set of evaluation metrics and guidelines tied together for selecting appropriate generative methods, this framework could improve the reproducibility and comparability of different studies, while also ensuring that the resulting synthetic data is of high (or at least acceptable, depending on the context) usefullness.

Identified Research Tasks and Goals

In light of the above discussion in the context of Differential Privacy [7, 8] the following research tasks had been identified:

1. **Task one** (T_1)**:** Survey the literature in order to identify the most common and relevant aspects to be considered in order to evaluate synthetic data. **Research Question 1** (Q_1)**:** Which aspects are appropriate to be used to evaluate synthetic data? **Goal 1:** (G_1) Identify the most important and appropriate aspects to be considered when evaluating synthetic data.

2. **Task Two** (T_2)**:** Find a standardized methodology as a set of fundamental metrics to properly and consistently evaluate synthetic data and their generative mechanisms. **Research Question Two** (Q_2)**:** Is there a standardized methodology and a set of metrics to evaluate synthetic data in all its aspects? **Goal Two:** (G_2) For each property of the data define a set of reliable and generalized evaluation metrics.

3. **Task Three** (T_3)**:** Define a unique scoring system and define an evaluation framework which is comprehensive and heterogeneous in such a way that covers the main properties of the data characteristics. **Research Question Three** (Q_3)**:** Is there the possibility to define an evaluation framework such that we can build and validate a standardized, comprehensive and heterogeneous scoring system for quantitatively evaluating the generated data? **Goal Three:** (G_3) Build a reliable and complete evaluation framework which is comprehensive and heterogeneous in such a way that its metrics cover all the main aspects of the data and their properties as a unique score.

4. **Task Four** (T_4)**:** Given a proper evaluation methodology and and established framework perform an extensive data evaluation on a set of interrelated differentially private algorithms. **Research Question Four** (Q_4)**:** How do similar algorithms compare to each other in terms of data quality and utility? **Goal Four:** (G_4) Perform a data quality empirical study on a set of interrelated differentially private algorithms.

Scientific and Engineering Contribution

Based on the above identified research tasks, the following scientific and engineering contributions are presented in this dissertation:

1. **Contribution on Data Science: Literature Review on Data Evaluation and Analysis Fundamental Criteria.**

   The authors carried a deep literature review and have identified at least two fundamental aspects of synthetic data which are necessary for carrying an exhaustive data evaluation process:

   (a) **Data Quality:** this aspect encompasses various properties: the completeness, consistency, and reliability of the synthetic data. It involves assessing how well the synthetic data reflects the statistical properties and distributions of the original data while maintaining integrity with comparable characteristics to the original data.

(b) **Data Utility:** this aspect refers to the property as the usefulness of the synthetic data for downstream tasks. It involves evaluating the synthetic data ability to support and perform when fed to various models and algorithms, such as machine learning and data mining tasks, with comparable performance to the original data.

Regarding the assessment of data quality aspect, the authors at this stage identified correlation analysis as a useful and reliable property as a good synthetic data generator should produce an output which leads pairwise correlations to be comparable to those obtained with the original data making synthetic data reliable in terms of relation between attributes. For what concerns the assessment of data utility aspect, the authors used a solid property: machine learning performances (and in some cases leaving room for data mining performances in the case of clustering analysis as evaluation method) as a good synthetic data generator should produce an output which leads to machine learning results that are comparable to those obtained with the original data making synthetic data usable for most common tasks.

As contribution authors first released an initial version of their synthetic data evaluation method including metrics aimed to measure both aforementioned aspects proposing G_ε. The introduced method for evaluating data quality includes correlation analysis (as data quality aspect) between the attributes of the private and synthetic datasets using (ϕ_k) [3] as the correlation coefficient. It also encompasses machine learning performances (as data utility aspect) over a set of (K) different machine learning models (M_k) using a set of (L) different accuracy metrics (acc_l) on (m) different prediction tasks [10]. The results of this research showed that results are coherent with other related publications validating the reliability of the results. With this contribution authors carried and completed T_1 answering Q_1 identifying as the main aspects for synthetic data evaluation to be **the data quality**

aspect and **the data utility aspect** contributing to the initial proposal of a general data evaluation method considering the most important aspects thus meeting G_1. In this contribution the candidate provided the core ideas and carried out the whole experimental part; other authors contributed on refining the main ideas, validated results and proof read the final manuscript.

2. **Contribution on Data Science: Methodology Design, Development, Validation and Extension.**

   The authors carried further research finding limitations to their previously presented evaluation method G_ε and found room for improvement presenting a definitive, and complete methodology leading to a more reliable and comprehensive synthetic data evaluation framework. The authors acknowledged and empirically demonstrated that G_ε suffered an insensitivity to linear transformations: as they demonstrate that G_ε is not significantly affected by linear transformations such as translations or rotations, which can alter the statistical distribution properties of the data without significantly affecting the evaluation metric. Authors also noted a limited consideration of attribute ranges: G_ε does not fully capture the range of attribute values, which is fundamental for generating realistic synthetic data. Furthermore authors discovered the insensitivity of G_ε to certain specific data generation techniques indeed, it may not accurately evaluate synthetic data generated using techniques which preserve pairwise correlation (e.g.: making use of well computed outliers which can preserve statistical correlations but altering most of the statistical distribution values).

   With this contribution the authors thus extended their research overcoming the aforementioned limitations keeping machine learning performances as a property for data utility aspect and identified two properties which when used together lead to a more comprehensive methodology for evaluating the data quality aspect:

   (a) **Macro-Statistical Properties:** this property includes evaluating the synthetic data ability to preserve key statistical measures as correlation analysis.

(b) **Statistical Distribution Properties:** this property focuses on assessing how accurately the synthetic data replicates the statistical distributions of the original data.

The authors introduced a definitive methodology and framework for evaluating synthetic data and their related generators G_{ε}^{+} [11]. They identified that the previously proposed methodology could be more comprehensive and resilient thanks to the introduction of probability distribution function distance, specifically Total Variation Distance (TVD) proposing G_{ε}^{+} which can be considered an improved version of G_{ε}. This enhancement addresses the limitations of G_{ε} in capturing the impact of linear transformations or changes in attribute distributions, making the evaluation methodology and framework more robust and reliable. The results of this research showed that results are coherent with other related publications validating the reliability of the results. With this contribution authors carried and completed T_2 proposing a complete and sound evaluation methodology assessing three fundamental data properties:

(a) **Macro-Statistics Property**

(b) **Statistical Distribution Property**

(c) **Machine Learning Performances Property**

The authors answered Q_2 and met G_2 making use of ϕ_k for what concerns correlation analysis (data quality aspect), making use of *accuracy scores* such as misclassification rate (data utility aspect) for what concerns machine learning performances and finally introducing the TVD for what concerns the statistical distribution property (data quality aspect). Finally with this contribution the authors also carried out T_3 aggregating all the properties (thus also their related metrics) under a single (parametric) score which can be used to fairly compare different data generators covering all the main aspects for a proper and complete synthetic data evaluation process. The authors proposed an evaluation framework validating it with empirical experiments which reflects their expectation and are in line with the current literature results thus affirmatively answering $Q3$ thus meeting G_3. In this contribution the candidate provided the core ideas and carried out the whole experimental part; other authors contributed on refining the main ideas, validated results and proof read the final manuscript.

3. **Contribution on Software Engineering and Data Science: Experiment Design and Validation.**

   The authors provided an extensive evaluation of a set of Differentially Private generative mechanisms namely based on Bayesian Networks, Vine-Copula and Gaussian-Copula [21, 12, 16] and identified the strengths and limitations of such methods when fed with datasets of varying nature over different privacy requirements, providing a fair and real use-case scenario. They utilized the previously proposed (G_ε) metric to benchmark these algorithms on datasets from various domains, including finance and credit risk. The findings revealed the strengths and weaknesses of each approach, highlighting the importance of tailoring generative techniques to specific data characteristics and use cases. For instance, copula-based methods demonstrated superior performance in preserving complex numerical dependencies, while DAG-based (Directed Acyclic Graphs) methods excelled in capturing the underlying structure of categorical data [9, 10]. With this contribution authors carried out T_4 and answered Q_4 meeting G_4. In this contribution the candidate provided the core ideas and carried out the whole experimental part; other authors contributed on refining the main ideas, validated results and proof read the final manuscript.

   The code developed to carry out experiments is publicly available at the following link: https://github.com/andreagalloni92/SDGEvalMETH/

Publications

List of publications, in chronological order, used in the dissertation:

1. Galloni, A., Lendák, I., & Horváth, T. (2020, October). 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 (pp. 25-34). Cham: Springer International Publishing. In this contribution the candidate provided the core ideas and carried out the whole experimental part; co-authors contributed on refining the main ideas, validated results and proof read the final manuscript.

2. Galloni, A., & Lendák, I. (2023, September). Differentially Private Copulas, DAG and Hybrid Methods: A Comprehensive Data Utility Study. In International Conference on Computational Collective Intelligence (pp. 270-281). Cham: Springer Nature Switzerland. In this contribution the candidate provided the core ideas and carried out the whole experimental part; co-authors contributed on refining the main ideas, validated results and proof read the final manuscript.

3. Galloni, A., Lendák, I., & Horváth, T. (2023, July). Extending Synthetic Data Evaluation Metrics. In 2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES) (pp. 209-214). IEEE. In this contribution the candidate provided the core ideas and carried out the whole experimental part; co-authors contributed on refining the main ideas, validated results and proof read the final manuscript.

Other publications of the author:

1. Galloni, A., Horváth, B., & Horváth, T. (2018, September). Real-time Monitoring of Hungarian Highway Traffic from Cell Phone Network Data. In ITAT (pp. 108-115). CEUR Proceedings. In this contribution the candidate carried out the whole implementation and experimental part; co-authors contributed on providing the core idea, validated results and proof read the final manuscript.

Bibliography

[1] Arno Appenzeller, Moritz Leitner, Patrick Philipp, Erik Krempel, and Jürgen Beyerer. Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, 12(23):12320, 2022.

[2] Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8, 2020.

[3] 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*, 152:107043, 2020.

[4] Rachel Cummings and Deven Desai. The role of differential privacy in gdpr compliance. In *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*, volume 20, 2018.

[5] Fida Kamal Dankar and Khaled El Emam. Practicing differential privacy in health care: A review. *Trans. Data Priv.*, 6(1):35–67, 2013.

[6] Vito D'Orazio, James Honaker, and Gary King. Differential privacy for social science inference. *Sloan Foundation Economics Research Paper*, (2676160), 2015.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory*

of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3, pages 265–284. Springer, 2006.

[8] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[9] Andrea Galloni and Imre Lendák. Differentially private copulas, dag and hybrid methods: A comprehensive data utility study. In *International Conference on Computational Collective Intelligence*, pages 270–281. Springer, 2023.

[10] Andrea Galloni, Imre Lendák, and Tomáš 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*, pages 25–34. Springer, 2020.

[11] Andrea Galloni, Imre Lendák, and Tomáš Horváth. Extending synthetic data evaluation metrics. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pages 000209–000214, 2023.

[12] Sébastien Gambs, Frédéric Ladouceur, Antoine Laurent, and Alexandre Roy-Gaumond. Growing synthetic data through differentially-private vine copulas. *Proceedings on Privacy Enhancing Technologies*, 2021(3):122–141, 2021.

[13] Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Matilde Karakachoff, Sophie Limou, Nicolas Vince, Matthieu Wargny, et al. Patient-centric synthetic data generation, no reason to risk re-identification in the analysis of biomedical pseudonymised data. 2022.

[14] Mehmet Emre Gursoy, Acar Tamersoy, Stacey Truex, Wenqi Wei, and Ling Liu. Secure and utility-aware data collection with condensed local differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2365–2378, 2019.

[15] Shalini Kurapati and Luca Gilli. Synthetic data: A convergence between innovation and gdpr. *J. Open Access L.*, 11:1, 2023.

[16] Haoran Li, Li Xiong, and Xiaoqian Jiang. Differentially private synthesization of multi-dimensional data using copula functions. In *Advances in database technology: proceedings. International conference on extending database technology*, volume 2014, page 475. NIH Public Access, 2014.

[17] He Li, Lu Yu, and Wu He. The impact of gdpr on global technology development, 2019.

[18] Tabish Maniar, Alekhya Akkinepally, and Anantha Sharma. Differential privacy for credit risk model. *arXiv preprint arXiv:2106.15343*, 2021.

[19] Daniel L Oberski and Frauke Kreuter. Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review*, 2(1):1, 2020.

[20] Amirsina Torfi, Edward A Fox, and Chandan K Reddy. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500, 2022.

[21] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.