EÖTVÖS LORÁND UNIVERSITY

FACULTY OF INFORMATICS

EVALUATING SYNTHETIC DATA GENERATORS IN THE CONTEXT OF DIFFERENTIAL PRIVACY

THESES OF THE PH.D. DISSERTATION

ANDREA GALLONI

SUPERVISOR: IMRE LENDÁK, PH.D.

THIS DISSERTATION IS SUBMITTED FOR THE FULFILMENT OF

DOCTOR OF PHILOSOPHY IN DATA SCIENCE AND ENGINEERING

PH.D. SCHOOL OF COMPUTER SCIENCE

HEAD OF SCHOOL: PROF. ERZSÉBET CSUHAJ-VARJÚ

PH.D. PROGRAM OF INFORMATION SYSTEMS

HEAD OF PROGRAM: PROF. ANDRÁS ID. BENCZÚR

OCTOBER 2025, BUDAPEST

DOI: 10.15476/ELTE.2024.310

. . . To Those Whom I Love & Those Who Love Me . . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 24,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 30 figures.

Galloni Andrea

October 2025

Acknowledgements

I would like to express my deepest gratitude to my former supervisor, Professor Tomáš Horváth, for his unwavering support and guidance throughout my doctoral studies. His wisdom, patience, and expert knowledge have been invaluable in shaping my research interests and helping me navigate the complexities of academic life. Working under his supervision has been a great privilege, and I am grateful for the many stimulating discussions we have had that have contributed significantly to sharpening my scientific skills.

I am also grateful to my current supervisor, Professor Lendák Imre, for his continued support and guidance, which has been instrumental in shaping my research direction and providing me with the necessary resources to pursue my research interests. Dr. Imre's expertise and encouragement have been critical in helping me to overcome the challenges of the doctoral studies and to develop my research skills to their fullest potential.

How not to mention Professor Orosz Gábor Tamás as he consistently, politely and gently pushed me to my limits with the most human kindness I've ever experienced in my life. A real motivator and supporter when most needed.

Beyond their academic mentorship, they also acted as mentors and paternal figures in my life, offering understanding and support during my darkest and most difficult moments.

I am deeply indebted to Deutsche Telekom, Ericsson and EIT Digital for providing me with the necessary resources to carry out my research work. Their support has been critical in enabling me to conduct research that has the potential to create significant societal impact. The knowledge and experience gained through these collaborations have been invaluable in helping me to develop my research skills and to broaden my horizons as a researcher.

I would also like to thank the members and former members of the Data Science and Engineering Department at the Faculty of Informatics for providing me with an intellectually stimulating research environment. The guidance and support of my colleagues have been invaluable in shaping my research interests and helping me to develop my research skills.

I am also grateful to my friends for sharing in my joy at the beginning of my doctoral journey and for their continued encouragement and support.

I would like to thank all my friends, with special gratitude to Arianna Bellino, who has always believed in me and my potential; to Gian Marco Canneori, for his valuable insights on life and mathematical theory; and to Edoardo Mascheroni, Damiano Fossa and Irene Pepe for their constant presence and support. Their encouragement and companionship have been invaluable, especially during the most challenging times of this journey.

Finally, I would like to express my gratitude to those who truly wished to be here today but, for various reasons, were unable to join. Your support and presence have been felt in spirit, and I carry it with me always (–· —— –· –· ·– & ·– –· –· ·–).

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 preservation 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.

Table of Abbreviations and Definitions

Table 1: Table of Acronyms and Definitions

Abbreviation	Definition
acc_l	Accuracy Metrics
AE	Absolute Error
AUC	Area Under the Curve
CDF	Cumulative Distribution Function
DAG	Directed Acyclic Graph
DP	Differential Privacy
DP SDG	Differentially Private Synthetic Data Generation
D_p	Private Dataset (containing sensitive data)
D_s	Synthetic Dataset (generated dataset)
D^i	A Generic Database
EFPA	Enhanced Fourier Perturbation Algorithm
EMD	Earth Mover's Distance
ERM	Empirical Risk Minimization
FP	Filter Priority
G_ε	Evaluation Methodology Metric
G_ε^+	Extended Evaluation Methodology Metric
GDPR	General Data Protection Regulation
K	Number of different machine learning models
KL	Kullback-Leibler Divergence (Relative Entropy)
KS	Kolmogorov-Smirnov Test

Continued on next page

Table 1: Table of Acronyms and Definitions (Continued)

Abbreviation	**Definition**
L	Number of different accuracy metrics
Lap	Laplace Distribution/Mechanism
MCC	Matthews Correlation Coefficient
MAE	Mean Absolute Error
MLE	Maximum Likelihood Estimation
MSE	Mean Squared Error
m	Number of different prediction tasks
$MWEM$	Multiplicative Weights Exponential Mechanism
$NIST$	National Institute of Standards and Technology
n_p	Number of Rows of the Private Dataset
n_s	Number of Rows of the Synthetic Dataset
PDF	Probability Density Function
$Pr[e]$	Probability of Event e to Occur
PSD	Private Spatial Decomposition
RE	Relative Error
$RMSE$	Root Mean Squared Error
SDG	Synthetic Data Generation
SVM	Support Vector Classifier
SVM	Support Vector Machine
TVD	Total Variation Distance
US	United States

Continued on next page

Table 1: Table of Acronyms and Definitions (Continued)

Abbreviation	**Definition**
X_i	Dataset Attributes
τ	Kendall's Tau (correlation measure)
ε	Privacy Budget Parameter in Differential Privacy
Δ	Function's Global Sensitivity in Differential Privacy)
Φ_k	Phi Correlation Coefficient
μ	Macro-Statistics Term of G_ε and G_ε^+
δ	Data-Utility Term of G_ε and G_ε^+
ν	Total Variation Distribution Distance Term of G_ε^+
α	Parametric Weight of μ
β	Parametric Weight of δ
γ	Parametric Weight of ν
$\mathcal{M}$	Randomized Algorithm
$\mathbb{N}$	Set of Natural Numbers
$\mathbb{R}$	Set of Real Numbers

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 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) [78]. Differentially private synthetic data is generated using statistical models capturing the key statistical properties of the original data while obfuscating sensitive information [73, 27]. 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 [4, 28, 56, 120], finance [7, 84], social sciences [35, 97], and cybersecurity [58]. 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.

1.2 Identified Research Tasks and Goals

In light of the above discussion in the context of Differential Privacy [38, 39] 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.

1.3 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) [8] 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 [46]. 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; co-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^+ [47]. 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; co-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 [134, 49, 76] 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 [45, 46]. 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; co-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/

1.4 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.

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.

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.

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.

1.5 Dissertation Overview and Structure

The main aim of this dissertation is to build a comprehensive synthetic data evaluation framework with sound theoretical basis in the context of Differential Privacy. Thus, in Chapter 1 we first introduce Differential Privacy, our notation and the anatomy of synthetic data generators. In Chapter 2, we describe the state-of-the-art models used in Synthetic Data Generation especially in the context of Differential Privacy with a focus on the evaluation metrics used for model and synthetic data analysis.

In Chapter 3, we introduce an initial version of a comprehensive evaluation framework and its application in practice. In Chapter 4 we evaluate several interrelated differentially private generative models utilising our developed framework. Chapter 5 presents our work towards improving the previously presented composite metric making it more resilient to specific transformations and more comprehensive adding a probability distance measure to the previously proposed evaluation framework. Chapter 6 draws conclusions, recommendations and future work.

Chapter 2

Differential Privacy and Synthetic Data Generation

2.1 Differential Privacy (DP)

With the exponential growth of data collection and the increasing number of data breaches and privacy violations, there is a growing demand for effective privacy-preserving methods that can protect individual privacy while still allowing for useful data generation and analysis of the data enabling acceptable quality and private data sharing with less concern. Within the context of *Private Data Analysis* [32], given the importance of the topic - in order to use data without privacy concerns - several approaches have been proposed. But before to illustrate the history of data privacy and the introduction of Differential Privacy there is the need to define some important terms:

- **Identifiers**: Identifiers are attributes within a dataset which uniquely distinguish individual entities from each other. These attributes are often directly linked to a human identity (but not only) and can be directly exploited for uniquely identify specific individuals. Examples of identifiers include social security numbers, email addresses, phone numbers, and full names.

- **Quasi-Identifiers**: Quasi-identifiers are attributes within a dataset that, while not directly identifying individuals on their own, can be combined in groups or with other data to potentially re-identify individuals with high degree of probability. Quasi-identifiers contribute to the risk of re-identification when combined in groups or with external information or background knowledge. Examples of quasi-identifiers might be age, gender, ZIP code, occupation, date of birth and not only.

- **Sensitive Information**: Sensitive information refers to data elements within a dataset that, if disclosed or compromised, could result in harm, loss of privacy, or other adverse consequences for individuals. Examples of sensitive information might be details about a person's health, financial status, criminal history, sexual orientation, religious beliefs, political orientation/identity or other personal matters. Protecting sensitive information is crucial for preserving individual privacy and preventing potential harm.

- **Privacy Breach**: A privacy breach occurs when there is unauthorized access to, disclosure of, or loss of control over strategic, confidential *sensitive information*. This breach may involve the exposure of personal data to individuals or entities who are not authorized to access them, potentially leading to privacy violations, identity theft, financial loss, reputational damage, or other adverse outcomes for the affected individuals. In general privacy breaches can occur due to various factors, including cyber attacks, data leaks, insider threats, inadequate security measures, human error or improper data anonymization techniques and/or procedures.

2.1.1 The Evolution of Data Anonymization Techniques

One of the early attempts to achieve privacy beyond mere unique identifiers removal or identifiers masking is *K-anonymity* [110, 117]. It was first introduced in 1998 by Sweeney, who identified the problem of *re-identification attacks* on *anonymized datasets* where unique identifiers are removed or masqueraded. The basic idea of K-anonymity is to make sure that each individual in a dataset cannot be uniquely identified based on the values

of their *quasi-identifiers*, which are attributes that can potentially identify them but not necessarily unique.

To achieve K-anonymity, a dataset is transformed in such a way that every record is indistinguishable from at least $k-1$ other records with respect to their quasi-identifiers. This means that any combination of quasi-identifiers that can be used to identify an individual in the dataset must occur in at least K records.

The most common approach to achieving K-anonymity is generalization [110] (but also other methods are possible [42]), which involves replacing specific values of quasi-identifiers with less precise or more general values. For example, if the age of an individual is a quasi-identifier, it can be generalized by replacing the specific age with an age range. Another approach is suppression, which involves removing quasi-identifiers from a record to prevent them from being linked to an individual but this process might imply the removal of important if not fundamental features useful for data exploitation depending on the context and the scope of data sharing or usage.

While K-anonymity provides a basic level of privacy protection, it has some limitations [125, 123] as authors note that K-anonymity only prevents association between individuals and tuples (protects against identity disclosure, but it fails to protect against sensitive attribute disclosure), instead of association between individuals and sensitive values such characteristics lead K-anonymity to be vulnerable to *Homogeneity Attack* and the *Background Knowledge Attack* [82, 116]. K-anonymity offers straightforward and intuitive protection. When a dataset adheres to K-anonymity with a specified value of k, it ensures that even if someone possesses only the quasi-identifier values of an individual, they cannot confidently pinpoint the corresponding record with probability higher than $1/k$. This latter characteristic of K-anonimity represents a limitation: it assumes that all individuals in a dataset have the same risk of re-identification, which may not be the case in practice. Additionally, it does not provide any guarantees about the information that can be inferred from the data, such as correlations between quasi-identifiers and sensitive attributes. To be more specific even though individuals within a group are grouped together and appear the same in terms of non-sensitive attributes or quasi-identifiers (like age, ZIP code), they

might still share the same sensitive attributes (e.g.: medical conditions). A homogeneity attack takes advantage of this shortcoming. Since *k-anonymity doesn't enforce diversity among individuals within a group in terms of sensitive attributes*, an attacker might exploit this lack of diversity to infer further information about all the individuals within the group thus compromising their privacy. In Background Knowledge Attacks the assumption is that the attacker possesses additional information (background knowledge) that is not part of the dataset itself but can be used to infer sensitive information about individuals. This information might be sourced from various sources (e.g.: public records, social media, or other datasets). The attacker might use inference by combining the anonymized dataset with his/her background knowledge, and thereby make inferences about specific individuals within the dataset. For example, even though the dataset might not directly reveal someone's medical condition, if the attacker knows that a particular individual has a certain age, lives in a specific area, and has visited a hospital with a known specialty, they might be able to infer the medical condition of that individual. In this case the attacker's ability to infer sensitive information about individuals despite the anonymization measures constitutes a privacy breach. This possibility undermines again the effectiveness of K-anonymity in protecting individual privacy leading to potential privacy breaches.

L-diversity [82] is a privacy model that extends K-anonymity to address the issue of low diversity within the same groups. It was first introduced by Machanavajjhala et al. in 2007 as a means of addressing the limitations of K-anonymity. The basic idea of L-diversity is to ensure that sensitive attributes are well represented among the k-anonymized records, in addition to ensuring that each record is indistinguishable from at least $k-1$ other records with respect to their quasi-identifiers. As per the authors L-Diversity can provide privacy even when the data publisher does not know what kind of knowledge the adversary possesses. To achieve L-diversity, a dataset is transformed in a way that ensures that for each set of quasi-identifiers, there are at least L well-represented values for each sensitive attribute. This means that an attacker who knows the quasi-identifiers of an individual cannot infer their sensitive attribute with high probability. The most common approach to achieving L-diversity is by adding a diversity requirement to the generalization step [110]

of K-anonymity on the sensitive attributes set. For example, if the age of an individual is a quasi-identifier, it can be generalized by replacing the specific age with an age range, and ensuring that each age range group has at least L distinct values for each sensitive attribute or adding noise to the sensitive attributes to make it difficult to infer their exact values. While L-diversity provides a stronger level of privacy protection than K-anonymity, it also has some limitations. One problem with L-diversity is that it is limited in its assumption of adversarial knowledge. It is possible for an adversary to gain information about a sensitive attribute as long as she has information about the global distribution of this attribute, also, a limitation is that the privacy mechanism assumes that all the attributes are categorical [79].

To overcome the limitations of L-diversity, in [79], the authors, introduce *T-closeness*, which addresses the limitations of previous techniques such as K-anonymity and L-diversity in ensuring privacy in data release. T-closeness proposes a new criterion for privacy preservation where t is a threshold that determines how close the distribution of sensitive attribute values in any subset of the data needs to be to the overall distribution of the sensitive attribute values in the entire dataset in order to consider the dataset sufficiently protected against privacy breaches. The data is first divided into groups of records called equivalence classes. These classes are formed based on certain attributes (e.g., age group, zip code), then T-closeness stipulates that the distribution of sensitive attributes within any equivalence class (e.g.: a set of records that are indistinguishable from each other with respect to certain "identifying" attributes) should closely resemble the overall distribution in the dataset, with the disparity between the two distributions limited by a threshold value, denoted as t. This novel approach aims to prevent attribute disclosure by constraining the ability of malicious actors to infer individual-specific information from released data based on background knowledge on the sensitive attribute (t-closeness contemplates only one sensitive attribute in its definition). While K-anonymity protects against identity disclosure by ensuring that each record is indistinguishable from at least $k-1$ other records in terms of quasi-identifiers, it falls short in preventing attribute disclosure. T-closeness overcomes these limitations by focusing on the distribution of sensitive attributes within equivalence

classes, thereby providing a more robust criterion for privacy preservation in data release. By ensuring that the distribution of sensitive attributes in each equivalence class closely matches the overall distribution in the dataset, T-closeness aims to limit the ability of adversaries to infer individual-specific information from released data, thereby enhancing privacy protection.

T-closeness offers a significant improvement over K-anonymity and L-diversity for data privacy, but it has its own limitations. The original T-closeness paper did not propose a computational procedure to achieve this property and did not mention the large utility loss that this property is likely to inflict on the original data [113]. *Data Distortion*: achieving a strong T-closeness guarantee often requires significant data modification through techniques like adding noise or generalization. This can distort the data and make it less useful for its intended purpose (e.g., statistical analysis, machine learning). Finding the right balance between privacy and data utility can be challenging. T-closeness is effective only for protecting a single pre-defined sensitive attribute. If there are multiple sensitive attributes T-closeness might not be sufficient/applicable. Computing the distance between distributions (Earth Mover Distance [109]) for T-closeness can be computationally expensive for large datasets, making it less scalable for real-world applications. It doesn't necessarily prevent attackers from learning statistical properties of the entire dataset or discovering correlations between different attributes. *To conclude, given the evolution of the aforementioned techniques we can still assert that the main limitation in common to those techniques is that K-anonymity, L-diversity and T-closeness are heuristic approaches that do not offer a precise measure of privacy protection, nor a strict and formal definition of privacy.*

2.1.2 Differential Privacy

Differential Privacy [37, 38, 40] has gained increasing attention in recent years [128, 142] due to growing concerns regarding *data privacy* and the need to quantitatively protect sensitive information in the age of big data [131] in a more formal and complete framework than previously introduced methods.

Differential Privacy is a powerful privacy-preserving technique which provides a rigorous mathematical framework [40] for ensuring that the privacy of individuals (and individuals' records) in a dataset is protected in a quantitative approach. The concept of Differential Privacy was first introduced by Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith in 2006 [37], and has since become a fundamental tool in the field of privacy-preserving data analysis.

Differential privacy has been demonstrated to be superior to K-anonymity, L-diversity and T-closeness under several aspects (even though under some restricted conditions L-diversity might be equivalent to Differential Privacy [33]). Firstly, *Differential Privacy provides a quantifiable measure of privacy protection*, whereas K-anonymity, L-diversity and T-closeness are heuristic approaches that do not offer a precise measure of privacy protection. Differential privacy provides a formal framework for measuring the privacy guarantees of a system, which allows for a more accurate assessment of privacy protection.

Secondly, Differential Privacy is more robust to re-identification attacks compared to K-anonymity and L-diversity or T-closeness. While K-anonymity and L-diversity and T-closeness can provide some level of protection against identity disclosure, they do not guarantee a degree of probability that an attacker will not be able to re-identify individuals in the dataset while differential privacy is capable of that by its definition. In contrast, Differential Privacy provides a stronger privacy guarantee by ensuring that the probability of identifying an individual in the dataset remains roughly the same, whether or not their data is included.

Finally, Differential Privacy is a more flexible approach to privacy protection compared to K-anonymity and L-diversity. While K-anonymity and L-diversity are designed to protect against specific types of privacy attacks, Differential Privacy can be tailored to provide protection against a wide range of attacks. Additionally, Differential Privacy also allows for the release of aggregate statistics while protecting individual privacy, whereas K-anonymity and L-diversity only allow for the release of sanitized datasets.

At its core, *Differential Privacy provides a formal definition of privacy* that captures the intuition that the inclusion or exclusion of any individual's data in a dataset should not

have a significant impact on the results of any analysis performed on that dataset. In other words, Differential Privacy aims to protect individual privacy by ensuring that the analysis of a dataset does not reveal any information about any particular individual record that could not be learned without their data. This means that the analysis should not disclose any specific details about an individual that would not be discoverable even if their data were or were not included in the dataset.

The key idea behind Differential Privacy is to add noise to the data in such a way that the results of any analysis on the data are still accurate, but the contribution of any individual's data to the analysis is effectively randomized. This noise ensures that the inclusion or exclusion of any individual's data has a minimal impact on the results of any analysis, thereby preserving individual privacy. We can identify two kinds of differential privacy:

Interactive Differential Privacy [107] enables privacy-preserving analysis of data in a centralized environment. In interactive differential privacy, multiple parties can access their private data through queries from a central entity, which is responsible to provide function's output while preserving the privacy of the individual data subjects. Interactive differential privacy enables privacy-preserving collaboration between different organizations and individuals, allowing them to jointly analyze sensitive data while preserving the privacy of the underlying data sources. But it erodes privacy budget at each different function call making it hard to use in practice. Figure 2.1 provides a graphical idea of the interactive privacy framework's schema.

Non-interactive Differential Privacy which represents the branch we do focus in this dissertation [74], on the other hand, is focused on preserving privacy in the analysis of data that can be stored anywhere or published. In this model, a data collector or custodian applies Differential Privacy techniques to the data before releasing it. *This approach has been widely used in many real-world applications* such as the analysis of location data [72, 127]. Figure 2.2 provides a graphical idea of the interactive privacy framework's schema.

Fig. 2.1 Interactive Differential Privacy Schema

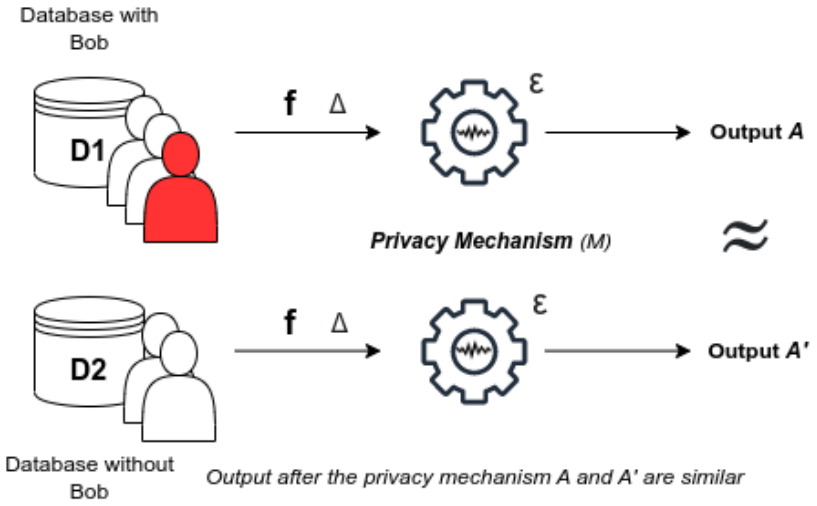
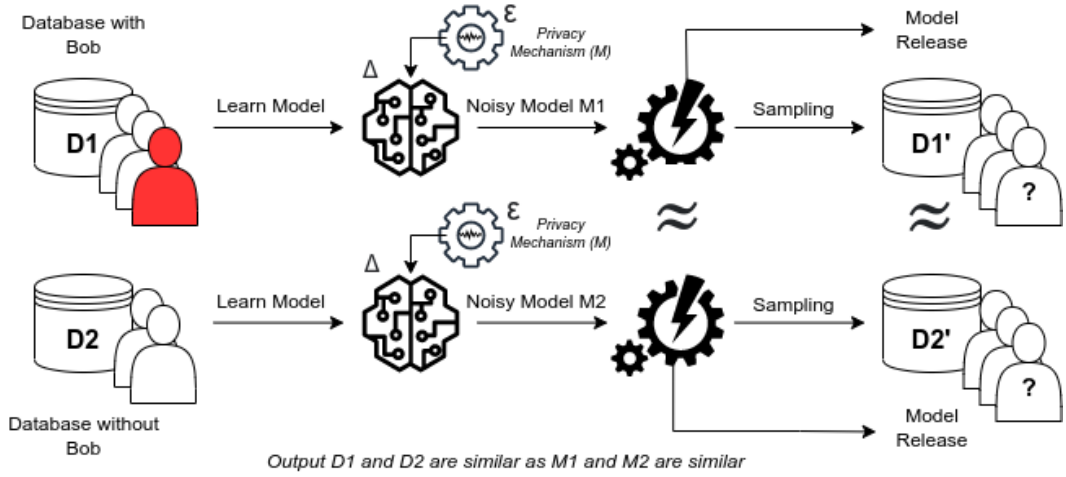


Fig. 2.2 Non-Interactive Differential Privacy Schema



One of the key benefits of Differential Privacy is its flexibility, as it can be applied to a wide range of data analysis tasks, including machine learning [112, 121], data mining [80, 92], and statistical analysis [38]. Additionally, Differential Privacy can be used in both centralized and decentralized settings [23, 52, 115], making it applicable to a variety of scenarios.

It is worth to note that Differential Privacy has been recognized as a sound and powerful tool also from different fields [5, 17, 18, 28, 143], industries and governmental organizations including companies such as US Census Bureau, Google, Apple, Facebook,

Microsoft and Uber [1, 43, 55, 91, 94, 105, 118, 119].

The following are the unique characteristics of Differential Privacy [38, 39]:

1. **It defines formal privacy guarantee:** provides a rigorous and formal privacy guarantee that can be mathematically proven. This allows for a principled approach to designing privacy-preserving systems and evaluating their effectiveness.

2. **Calibrated Statistical Randomness:** Differential privacy relies on the use of statistical randomness to inject calibrated noise into the laerned model in the case of non-interactive DP and in the output of functions in the case of interactive DP. This makes it difficult for an attacker to determine whether a particular individual's data was included in the query.

3. **Preservation of utility:** despite adding noise to the query output, differential privacy aims to preserve the utility of the data as much as possible. This is achieved by carefully *controlling the amount and type of noise* that is added.

4. **Composability:** Differential privacy is compositional, which means that privacy guarantees can be combined even when multiple different functions are applied to the same dataset.

5. **Transparency and Measurability:** Differential privacy is transparent, which means that users can be made aware of the privacy risk associated with their data before they choose to share it. This is particularly important in settings where users may be reluctant to share their data due to privacy concerns.

6. **Flexibility:** Differential privacy can be applied to a wide range of data analysis tasks, from simple counting queries to more complex machine learning algorithms training.

While Differential Privacy has gained widespread attention and adoption in recent years, there are still many challenges that need to be addressed. These include developing more efficient and scalable algorithms for implementing Differential Privacy, as well as

developing methods for evaluating the effectiveness of Differential Privacy mechanisms in real-world industrial settings [54]. There is still on open issue on how to define the balance between privacy guarantees (noise injection) [66] and data quality. Despite these challenges, Differential Privacy remains the most used and studied approach for protecting individual privacy to date.

2.1.3 DP Notation and Theoretical Concepts

Differential Privacy is based on two fundamental concepts: *Function Sensitivity and Privacy Budget*. In order to define Differential Privacy there is the need of formal definitions.

2.1.1. Definiton. (Random Mechanism) *A randomized algorithm $\mathcal{M}$ with domain A and discrete range B is associated with a mapping $M : A \to \Delta(B)$. On input $a \in A$, the algorithm $\mathcal{M}$ outputs $\mathcal{M}(a) = b$ with probability $(\mathcal{M}(a))_b$ for each $b \in B$. The probability space is over the coin flips of the algorithm $\mathcal{M}$. [40]*

Let define a generic statistical database as D. We will think of a database D as being collections of records from a universe $\mathcal{D}$. It is convenient to represent databases by their histograms: $D \in \mathbb{N}^{|\mathcal{D}|}$, in which each entry D_i represents the number of elements in the database D of type $i \in \mathcal{D}$ (letting the symbol $\mathbb{N}$ denote the set of all non-negative integers, including zero). In this representation, a measure of the distance between two databases D^1 and D^2 will be their ℓ_1 distance:

2.1.2. Definiton. (Databases Distance) *The ℓ_1 norm of a database D is denoted as $\|D\|_1$ and it is defined to be:*

$$\|D\|_1 = \sum_{i=1}^{|\mathcal{D}|} |D_i|$$

The ℓ_1 distance between two databases D^1 and D^2 is $\|D^1 - D^2\|_1$.

Note that $\|D\|_1$ is the size of a database D (i.e., the number of records it contains), and $\|D^1 - D^2\|_1$ is a measure of how many records differ between D^1 and D^2.

Databases may also be represented by multi-sets of rows (elements of D), in this case distance between databases is typically measured by the Hamming distance, i.e., the number of rows in which they differ.

2.1.3. Definiton. (Adjacent Databases) *Two databases D^1, D^2 are called adjacent (or neighbour) if their l_1 distance is bounded by 1, that is:*

$$\|D^1 - D^2\|_1 \leq 1$$

2.1.4. Definiton. (Differential Privacy) *A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{D}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $D^1, D^2 \in \mathbb{N}^{|\mathcal{D}|}$ such that two adjacent databases ($\|D^1 - D^2\|_1 \leq 1$):*

$$\Pr[\mathcal{M}(D^1) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D^2) \in \mathcal{S}] + \delta$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, follows that $\mathcal{M}$ is ε-differentially private.

Intuitively, Definition 2.1.4 states that for any two adjacent databases D^1 and D^2, the probability of the mechanism $\mathcal{M}$ returning any particular output $\mathcal{S}$ on D^1 is not much different from the probability of the algorithm returning $\mathcal{S}$ on D^2, up to a multiplicative factor of $\exp(\varepsilon)$ and an additive factor of δ. In other words, the algorithm $\mathcal{M}$ is *Differentially Private* because it does not reveal much about any particular individual's data, even when an attacker has access to the output of the algorithm on adjacent databases.

Let's suppose we have a query which is a function f, and the database is D, the true answer is $f(D)$. The mechanism $\mathcal{M}$ adds calibrated random noise to the true answer to produce the *response* $f(D) \in R$. The main concept of preserving privacy is by providing as a response a noisy version of the true answer. Given this premise the amount of noise should be calibrated based on the sensitivity of the function f to any input.

2.1.5. Definiton. (Global Sensitivity Δf) *For $f : \mathcal{D} \to R^k$, the global sensitivity of f is:*

$$\Delta f = \max_{D^1, D^2} \|f(D^1) - f(D^2)\|_1$$

for all D^1, D^2 differing in at most one element.

In the case where $k = 1$, the sensitivity of the function f corresponds to the maximum difference in the output of f between two neighbor databases (any two databases that differ in only a single entry). Notably, the sensitivity is a property of f alone, and it is completely independent of the input database. By capturing how much the output of f changes when the input database changes by only one element, the sensitivity serves as a crucial parameter in designing an effective differentially private data generator, by determining the amount of additive noise required to ensure Differential Privacy.

2.1.4 Privacy Mechanisms

It is possible to achieve Differential Privacy through different techniques based on the distribution from which the noise is generated from [96, 99]. The most used are:

1. **Exponential Mechanism** selects an output from a set of possible outputs from an exponentially distributed noise function satisfying a Lipschitz condition [38].

2. **Laplace Mechanism** selects an output from a set of possible outputs from a Laplace noise distribution to provide differential privacy [38].

3. **Gaussian mechanism** adds Gaussian noise to provide differential privacy [9, 34].

4. **Randomized Response** used for binary queries (e.g., yes/no questions) and adds random noise to the response to protect privacy [40].

5. **Matrix Mechanism** it is a generalization of the Laplace mechanism that can handle multiple correlated outputs [75].

Hereby we are going to describe only the mechanisms which are most used and relevant within this context: *Exponential Mechanism*, *Laplace Mechanism*.

Exponential Mechanism

The Exponential Mechanism has at his core the exponential distribution.

2.1.6. Definiton. (The Exponential Distribution) *The Exponential Distribution with scale* λ *is the distribution with probability density function:*

$$\exp(\lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & otherwise \end{cases}$$

The exponential distribution equals zero for non-positive values.

The **Exponential Mechanism** [88] is made for query functions with arbitrary utility and arbitrary range, while guarantees Differential Privacy. Given some arbitrary range $\mathcal{R}$, the Exponential Mechanism is defined with respect to the utility function $u : \mathbb{N}^{|\mathcal{D}|} \times \mathcal{R} \to \mathbb{R}$, which maps database/output pairs to utility scores. Intuitively, for a database D, the mechanism outputs elements of $\mathcal{R}$ with the maximum possible utility score. Note that we refer to the sensitivity of the utility score $u : \mathbb{N}^{|\mathcal{D}|} \times \mathcal{R} \to \mathbb{R}$. Only the sensitivity of u with respect to its database argument is relevant and it can be arbitrarily sensitive in its range argument [40].

The Exponential Mechanism $\mathcal{M}_E(x, u, \mathcal{R})$ selects and outputs an element $r \in \mathcal{R}$ with probability proportional to $\exp\left(\frac{\varepsilon u(x,r)}{2\Delta u}\right)$.

2.1.7. Theorem. (The Exponential Mechanism is Differentially Private [40, 88]) *The Exponential Mechanism as per its definition preserves* ε*-Differential Privacy.*

The Exponential Mechanism can offer significant utility guarantees due to its exponential discounting of outcomes with decreasing quality scores. This feature allows the mechanism to rapidly decrease the selection probability of items with low scores, thereby ensuring that high-scoring items are chosen with a greater likelihood.

Laplace Mechanism

The mechanism which is most commonly used in applied DP schemes is the ***Laplace Mechanism***. It is based on the *Laplace Distribution* with standard deviation $\sqrt{2}\Delta f/\varepsilon$, it is symmetric and can be denoted as $\text{Lap}(\Delta f/\varepsilon)$.

2.1.8. Definiton. (The Laplace Distribution) *The Laplace Distribution with 0 mean and scale b is the distribution with probability density function:*

$$\text{Lap}(x \mid b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right).$$

The Laplace distribution is a symmetric version of the exponential distribution.

Given a function f the privacy mechanism $\mathcal{M}$ responds with

$$f(D) + (\text{Lap}(\Delta f/\varepsilon))^k$$

namely adding noise with distribution $\text{Lap}(\Delta f/\varepsilon)$ independently on each of the k components of $f(D)$. Lower ε values flattens the $\text{Lap}(\Delta f/\varepsilon)$ distribution, leading to a larger magnitude of the noise. Once ε is fixed, assuming f having high sensitivity leads to flatter curves, leading the noise to be bigger in magnitude.

2.1.9. Theorem. (The Laplace Mechanism is Differentially Private [38]) *For $f : \mathcal{D} \to R^k$, the mechanism $\mathcal{M}_f$ that adds independently generated noise with distribution* $\text{Lap}(\Delta f/\varepsilon)$ *to each of the k output terms satisfies ε-Differential Privacy.*

The mechanism denoted as $\mathcal{M}$ in the aforementioned context exhibits high accuracy for counting queries. This is primarily attributed to the fact that the noise added to preserve Differential Privacy is solely dependent on the sensitivity of the function and the parameter ε, while being independent of the size and structure of the underlying database. As a result, the introduced errors for typical queries in a large database are relatively low, further substantiating the effectiveness of the Differential Privacy mechanism.

2.1.5 Properties of Differential Privacy

Differential Privacy holds several properties which allows algorithm architects to take advantage of them. The most important properties are:

1. *Sequential Composition Theorem* [71]
2. *Parallel Composition Theorem* [111]
3. *Post-Processing Theorem* [111]

Sequential Composition

2.1.10. Theorem. (Sequential Composition [40]) *The sequential composition theorem allows for the analysis of the privacy guarantees of a sequence of operations, each of which provides differential privacy guarantees individually.*

Let $\mathcal{M}_1 : \mathbb{N}^{|\mathcal{D}|} \to \mathcal{R}_1$ *be an* ε_1*-differentially private algorithm, and let* $\mathcal{M}_2 : \mathbb{N}^{|\mathcal{D}|} \to \mathcal{R}_2$ *be an* ε_2*-differentially private algorithm. Then their combination, defined to be* $\mathcal{M}_{1,2} : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_1 \times \mathcal{R}_2$ *by the mapping:* $\mathcal{M}_{1,2}(x) = (\mathcal{M}_1(x), \mathcal{M}_2(x))$ *is* $\varepsilon_1 + \varepsilon_2$*-differentially private.*

In other words, if we apply the two algorithms $\mathcal{M}_1$ and $\mathcal{M}_2$ in sequence to a dataset, the combined sequence algorithm $\mathcal{M}$ satisfies differential privacy with a privacy budget of $\varepsilon_1 + \varepsilon_2$.

Parallel Composition

The parallel composition theorem of differential privacy states that if two differentially private mechanisms are used independently on disjoint subsets of a data set, then their composition is also differentially private. In other words, if two mechanisms satisfy differential privacy individually, then running them in parallel on different parts of the dataset also satisfies differential privacy.

2.1.11. Theorem. (Parallel Composition [89]) *Let* $\mathcal{M}_i$ *each provide* ε*-differential privacy. Let* D^i *be arbitrary disjoint subsets of the input dataset D. The sequence of* $M_i\left(X \cap D^i\right)$ *provides* ε*-differential privacy.*

This theorem is useful in situations where the dataset is partitioned into disjoint subsets, and different mechanisms are applied to each subset. The parallel composition theorem allows us to analyze the privacy guarantees of the overall mechanism based on the privacy guarantees of each individual mechanism. However, it is important to note that the parallel composition theorem assumes that the subsets are disjoint, and applying the mechanisms to overlapping subsets requires additional analysis. If all the assumptions are satisfied this theorem allows the user to run in parallel privacy tasks with no privacy concerns.

Post-Processing

The Post-Processing Theorem describes how to compose differentially private mechanisms to obtain new differentially private mechanisms. In particular, the theorem states that any function computed on the output of a differentially private mechanism remains differentially private, provided that the function is independent of the input data.

2.1.12. Theorem. (Post-Processing Composition [89]) *Let* $\mathcal{M} : \mathbb{N}^{|\mathcal{D}|} \to R$ *be a randomized algorithm that is* (ε, δ)*-differentially private. Let* $f : R \to R'$ *be an arbitrary randomized mapping. Then* $f \circ \mathcal{M} : \mathbb{N}^{|\mathcal{D}|} \to R'$ *is* (ε, δ) *differentially private.*

This theorem provides a robust guarantee that users can handle the output of the privacy mechanism with complete disregard for subsequent post-processing or public release. Specifically, it ensures that the privacy assurances provided by the mechanism remain intact even if the output is subjected to further manipulation or made publicly available for later analysis. Consequently, users can confidently process or disseminate the data without compromising the privacy protections originally established, thereby maintaining the integrity of the privacy guarantees throughout the entire lifecycle of the data.

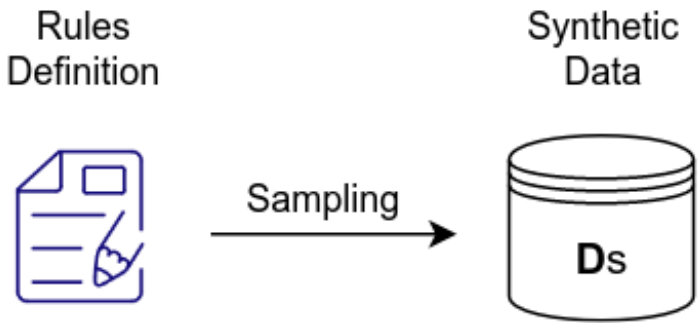
2.2 Synthetic Data Generation (SDG)

Synthetic data generators maintain a high level of scientific interest over time. Several scholarly publications, including [41, 87, 102], have extensively studied the development

and application of synthetic data generators. In general a *Synthetic Data Generator* (SDG) is a tool or software that creates datasets based on statistics [122] or rules [30].

Rule-based data generators follow a set of predefined rules or specifications to generate data sets. These rules may include specific criteria, such as minimum and maximum values for each data point or constraints that must be satisfied. Rule-based generators are often used to generate structured datasets that follow a particular schema or format and this kind of generators do not need to be fed with any dataset beyond the rule set. Figure 2.3 illustrates the typical rule-based SDG pipeline.

Fig. 2.3 Rule-based Synthetic Data Generation Diagram

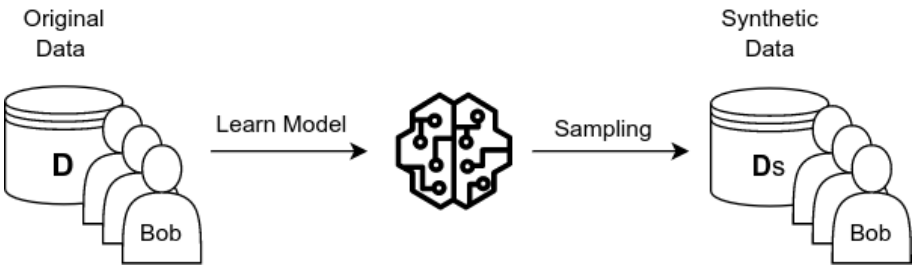


Algorithm-based or Model-based data generators, on the other hand, use complex mathematical algorithms and statistical techniques to generate datasets from preexisting data. These algorithms may be based on learning specific probability distributions or machine learning techniques, and they may incorporate complex statistical models. These generators often can also create datasets with higher levels of complexity than rule-based generators, as they can capture and simulate more complex relationships between data points. Algorithm based generators, on the other end do need to be fed with a dataset. Figure 2.4 illustrates the typical Model-based SDG pipeline.

These algorithms generate datasets which are representative of real-world data or designed to simulate specific experimental conditions [3].

Data generators are particularly useful in scientific research, where they are used to simulate experimental conditions or create synthetic datasets for statistical analysis. For

Fig. 2.4 Algorithm-based Synthetic Data Generation Diagram



example, a data generator might be used to create synthetic datasets to test the efficacy of new statistical models or to investigate the properties of complex systems.

In industry, data generators are widely used to create large volumes of realistic data for machine learning, deep learning, and other data-driven applications. These datasets are used to train and validate machine learning algorithms, optimize data analytics pipelines, and improve business processes beyond privacy and confidentiality.

The creation of datasets that accurately represent real-world data is a complex task. Therefore, data generators employ advanced statistical techniques to ensure that the generated data is both representative and diverse. For example, data generators may incorporate *probability distributions* and/or *correlation analysis* [85, 101] to ensure that the generated data accurately reflects the statistical properties of the underlying data.

2.3 Differentially Private Synthetic Data Generation (DP SDG)

Differentially Private Synthetic Data Generation refers to techniques generating synthetic data in a way that preserves the privacy of the individuals whose data is used to generate it thanks to the use of the previously introduced Differential Privacy approaches. Thus to achieve it there is the need of further steps if we do take into consideration non-private synthetic data generation (figures 2.4 and 2.5 provide an illustration of the additional steps). The goal is to produce a dataset that is similar to the original data in terms of statistical properties and data distribution, but does not contain any information that could be used

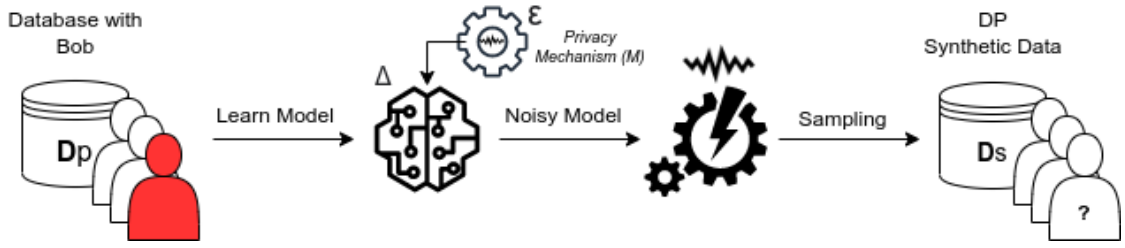
to identify the individuals whose data was used to generate it thus the sensitivity to be computed is the sensitivity of the learner model.

The process of generating differentially private synthetic data involves various steps:

1. Learner Model Selection
2. Learner sensitivity mathematical proof (Δ)
3. Choose the DP Mechanism
4. Learning the model on any D
5. Inject calibrated noise
6. Sampling Synthetic Data from the noisy model
7. Public release of DP Synthetically Generated Data

DP SDG is increasingly being used in situations where traditional methods of data sharing such as releasing the raw data, performing data masking [104] or weak data anonymization [50] would risk violating the privacy of individuals [48, 69]. By generating privacy proof differentially private synthetic data that is statistically similar to the original data, researchers and analysts can conduct analyses without compromising the privacy of individuals within a defined bound. Figure 2.5 illustrates the typical DP SDG Pipeline.

Fig. 2.5 Differentially Private Synthetic Data Generation Pipeline Schema



More formally we do define the task of DP SDG as:

2.3.1. Definiton. (General DP SDG Task) *The task of differentially private synthetic data generation (SDG) tackled in this dissertation is: given a private dataset* $D_p \subset X_1 \times X_2 \times \cdots \times X_m$ *of* $|D_p| = n_p$ *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* $|D_s| = n_s$ *of rows. The attributes (columns)* $X_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ *refer to n-dimensional vectors having numeric, ordinal or nominal values n being equal either to* n_p *or* n_s*. The task 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).*

A well-designed SDG solution generates data with similar statistical characteristics to the private data, e.g. maintains the correlations between inter-related attributes, keeps data utility high while guarantees a reasonable level of privacy.

2.4 Related Work

Differential Privacy preliminaries were firstly presented in 2006 by Dwork et al. [38]. This new mathematical approach suddenly took the attention of the scientific community [16, 28, 44, 59, 62, 70, 72, 83, 129, 130, 139–141] due mainly to two characteristics: there is a formal and quantitative level of privacy embedded in it as form of probability and it's immune to most of privacy breaches because it assumes that the attacker has access to external databases containing individual's data.

In [10] authors do present a technique to release contingency tables making use of the *Laplace Mechanism* and *Fourier Coefficients* transforming the data into the *Fourier Domain*. Authors propose to construct a synthetic database that corresponds to these noisy tables as those can be used to output a synthetic database preserving privacy and all low-order marginals with small error. To get a noisy approximation of the marginals (which are $\varepsilon-$differentially private and could be released as well) of the contingency table then the authors propose an optional post-processing step 2.1.12 using *linear programming* on the contingency tables to guarantee *consistency* (non-negative counts) and *integrality* of results. Worth to note is that the authors show that the case of not including the post-processing

step the integrity can not be guaranteed while maintaining non-negativity. In this case the authors use various metrics to quantify the error, such as the L_1 distance and the maximum absolute error. They also assess the consistency of the released marginals by checking whether there exists a valid contingency table that could have generated those marginals.

In 2009 authors of [126] proposed *Privelet* and *Privelet*$^+$ (a variant able to properly model both attributes with small domain and attributes with large domain) an ε-differential privacy implementation based on wavelet transforms (invertible linear transformations) [114] before adding noise (using the *Laplace Mechanism*) while providing accurate answers for range-count queries on both ordinal and nominal data providing a theoretical and experimental analysis on both privacy and utility. The authors evaluate their results making use of *relative error* and *square error* for different values of ε and different range count query types.

In [64] the authors propose a three-step approach to improve the accuracy of differentially private histograms while maintaining privacy guarantees on a set of queries. The first step is to define a set of *Range Count Queries* Q. In the second step they use the *Laplace Mechanism* (2.1.9) to get noisy answers $\tilde{q}$ adding random independent noise to each original query answer q. Third they propose a post-processing step applied to $\tilde{q}$ to resolve inconsistencies using a novel approach called *constrained inference*, finding a new set of answers $\overline{q}$ that is the "closest" set to $\tilde{q}$ while satisfying consistency constraints. The authors demonstrate this technique on two histogram tasks: *unattributed histograms* and *universal histograms*. They show that there is an efficiently-computable closed-form expression for the consistent query answer closest to a private randomized output for each of the tasks. Authors also prove bounds on the error of the inferred output and demonstrate significant improvements in accuracy through experiments. The proposed method is evaluated applying on real datasets using the *Mean Squared Error* as evaluation metric on three datasets and different values of ε. On unattributed histograms authors claim they measured an errors at least an order of magnitude lower than previous techniques, while their approach to universal histograms can reduce the error for larger ranges by 45-98% and improve on all ranges in some cases.

The authors of [51] propose a discrete variant of the *Laplce Mechanism* called *Geometric Mechanism*. Thanks to this mechanism the authors try to maintain the utility of the data as high as possible for each record (referred as *user* in the contribution). The authors aim to create mechanisms which provide strong and general utility guarantees to every potential user, regardless of their preferences and side information. Preferences are modeled by a loss function that measures a user's loss when the query result is compared to the mechanism's output. The side information of a user is represented as a prior probability distribution over query results, which reflects their beliefs. The expected loss of a user with a prior and loss function, and an oblivious mechanism is calculated as a measure of the mechanism's overall *(dis-)utility* to the user.

To evaluate the results, the authors use a utility model that is based on the user's preferences and side information. They define a loss function (l) and a prior probability distribution (p) to model the user's preferences and side information, respectively. The user's expected loss is calculated based on these parameters, and the overall utility of the mechanism to the user is measured by the expected loss over the coin tosses of the mechanism and the prior. The optimal ε-differentially private mechanism for a user is the one that minimizes the user-specific objective function. To prove the main result, the authors show that for every user, there is an optimal mechanism that can be factored into a user-independent part (the ε-geometric mechanism) and a user-specific computation (optimal remap using 2.1.12) that can be delegated to the user. They demonstrate that the ε-geometric mechanism is optimal for every rational user when combined with an optimal remap.

The main result of the paper is proven by showing that the ε-geometric mechanism is optimal for every rational user when combined with an optimal remap tailored to the user's preferences and side information.

In [134, 133] authors introduce *PrivBayes* as a novel solution. *PrivBayes* constructs a Bayesian network to model attribute correlations and approximate the data distribution using low-dimensional marginals. The authors introduce a novel approach for privately constructing Bayesian networks. Calibrated Noise injection into the marginals ensures

differential privacy, and synthetic data is generated from the noisy marginals and the Bayesian network sampling along the DAG of marginals.

Unlike prior methods which optimize output for specific queries, *PrivBayes* aims to accurately approximate the high-dimensional data distribution, enabling diverse query evaluations on the same released data. Operating in low-dimensional spaces mitigates the signal-to-noise problem and avoids scalability issues. The algorithm comprises network learning, distribution learning, and data synthesis steps. Experimental evaluation demonstrates *PrivBayes* to be reliable, even without optimization for specific queries. The paper contributes significant theoretical insights and technical advancements, proposing a generic solution adaptable to different types of queries without prior knowledge of the data usage.

The evaluation of the results in the conducted experiments encompasses two primary tasks: assessing the accuracy of *PrivBayes* in constructing marginals of datasets and evaluating the performance of *PrivBayes* in training multiple Support Vector Machine (SVM) classifiers. For both tasks, *PrivBayes* is compared against several baseline methods to gauge its efficacy in preserving differential privacy while maintaining data utility.

In the first task, the accuracy of *PrivBayes* in constructing marginals is evaluated using the *Total Variation Distance (TVD)* [124] between the noisy marginals generated by *PrivBayes* and their noise-free counterparts obtained from the original dataset.

The second task involves training multiple SVM classifiers simultaneously on multiple datasets. *PrivBayes* generates synthetic datasets from the training data and authors used them to train SVM classifiers, with the quality of each classifier assessed based on its misclassification rate on a testing set. Authors compare their results with other methods such as PrivateERM [138], PrivGene [136]. Overall, the experimental results demonstrate overall good performance of *PrivBayes* compared to baseline methods across both tasks, highlighting its effectiveness in preserving privacy while maintaining data utility. However, it is noted that *PrivBayes* may not always outperform baseline methods optimized for specific tasks, such as SVM classification.

The authors of [21] propose a technique for producing privacy-preserving approximations of classifiers learned through *Empirical Risk Minimization* (ERM) in the context

of *privacy-preserving machine learning*. The goal is to protect sensitive personal information while still being able to analyze datasets containing such information. The paper presents two methods: *Output Perturbation* and *Objective Perturbation* as an extension of [20]. Output perturbation adds noise to the output of the standard *ERM* algorithm, while objective perturbation involves perturbing the objective function before optimizing over classifiers. Theoretical results are provided to show that these methods preserve privacy and provide generalization bounds for linear and nonlinear kernels for *SVM* classifiers [13]. The paper also addresses the issue of parameter tuning in machine learning algorithms. A privacy-preserving parameter tuning algorithm is developed using a randomized selection procedure, ensuring privacy during the learning process.

Experiments on real datasets demonstrate the effectiveness of the proposed methods. The results show that *objective perturbation outperforms output perturbation* in balancing the trade-off between privacy and learning performance.

The metrics used to evaluate synthetic data it's *Misclassification Rate* comparing *logistic regression* and *SVM* over two datasets on different values of ε.

In [92] introduce *DiffGen* a Differentially Private anonymization algorithm based on Generalization.

The authors present a pioneering generalization-based algorithm for differentially private data release that retains information for classification analysis. The proposed solution generates a generalized contingency table and adds noise to the counts. The algorithm can handle both categorical and numerical attributes without requiring pre-discretization of numerical attributes. Furthermore, *DiffGen* adaptively determines split points for numerical attributes and partitions data based on the workload, ensuring ε-differential privacy. The presented is demonstrated to be highly efficient and scalable (namely the complexity analysis results in $O(h \times |D| \log |D|)$), capable of handling large datasets. The authors address questions concerning the interactive versus non-interactive approaches to data mining and whether differentially private data offers less utility than k-anonymous [117] data.

Experimental results demonstrate that the proposed algorithm outperforms previously proposed differentially-private interactive algorithms for learning classifiers. The authors demonstrate that their anonymization algorithm can effectively maintain information for data mining tasks, particularly decision tree classifiers.

The Authors evaluate their results using the classifier *Accuracy Score* against the original data accuracy for different values of ε.

In [31] the authors introduce a framework based on cuboids. Cuboids are subsets of a data cube formed by aggregating data from a fact table along specific dimensions. In a data cube, the fact table contains raw data, and the cuboids represent different levels of aggregation of this data. Each cuboid corresponds to a unique combination of dimensions from the fact table, and its cells contain aggregated measures such as counts, sums, or averages. In this case a subset of cuboids is computed directly from the fact tables with added noise, while the remaining cuboids are derived from the noisy subset. Two publishing scenarios are considered: minimizing the maximum noise in published cuboids and maximizing the number of precise cuboids within a given noise threshold. The authors prove that selecting the initial subset of cuboids is NP-hard and provide efficient algorithms with approximation guarantees. The paper primarily evaluates the effectiveness of different publishing algorithms for differential privacy (DP) based on two key evaluation metrics:

Error Measurement: The paper measures error as the absolute difference between the real count measure computed directly from the fact table and the noisy measure released by each of the seven DP algorithms. The error is calculated at both the individual cell level and aggregated cuboid level.

Max Cuboid Error and Average Cuboid Error: These metrics quantify the level of error in the published cuboids. The "max cuboid error" represents the maximum error observed among all cuboids, while the "average cuboid error" computes the average error across all cuboids, unfortunately the authors do not compare their results with other DP techniques.

The authors of [60] propose Multiplicative Weights Exponential Mechanism (*MWEM*) a method based on a combination of the *Exponential Mechanism* [75] and the *Multiplicative Weights* approach [61, 57]. The authors present a scalable implementation of *MWEM*

capable of processing datasets with substantial complexity (high number of attributes). While producing synthetic data for the considered query classes is known to be computationally hard, their implementation can process datasets with thousands of attributes in a finite amount of time. This is achieved by integrating a *scalable parallel implementation* of Multiplicative Weights and a representation of the approximating distribution in a factored form that only exhibits complexity when required by the model.

The Authors evaluate their method with the *Squared Error* on *range query* answers against different values of ε. Furthermore the authors do evaluate the attributes probability distribution distance through *Relative Entropy* (Kullback-Leibler (KL) divergence).

In [2] the paper proposes two new techniques for generating differentially private histograms. These techniques exploit the redundancy of real-life datasets to improve the accuracy of the histograms while still maintaining privacy. The first technique is an enhanced version of the *Fourier Perturbation Algorithm* called *EFPA*, which compresses the data and then sanitizes it. The second technique, *P-HPartition*, is based on *hierarchical clustering* and exploits the redundancy between bins. The experimental results show that the proposed techniques outperform existing solutions and provide a better trade-off between reconstruction and perturbation errors. The techniques are well-suited to unattributed histograms, and their utility is high for many real-life datasets, which in real applications often are highly compressible. The paper concludes that these techniques exhibit great promise for outperforming the state-of-the-art solutions.

The authors evaluate the dataset making use of *Mean Squared Error* on range count queries and *KL-Divergence* over histogram's distributions. Authors compare their results against [64, 60, 126] for different values of ε.

In [25] the authors propose several variants of a mechanism based on the popular *KD-tree data structure* [11], which is used for organizing, indexing and partitioning spatial data. The *Private-KD* algorithm enables the construction of a differentially private spatial decomposition by adding *Laplacian Noise* during the process of splitting the space into partitions. The proposed method ensures that the released spatial decomposition satisfies differential privacy, protecting individuals' location information in the dataset from being

re-identified. Furthermore, the authors show that their approach can be applied to a wide range of spatial data analysis tasks, such as answering range queries using *Relative Error* as utility metric. The authors also explore the trade-off between privacy and utility, providing insights into how the algorithm's parameters can be adjusted to achieve the desired level of privacy while still preserving the usefulness of the released data for various spatial data analysis tasks. This approach has been later extended in [135] as *PrivTree*.

In [26] the authors propose a solution to address the scalability issue encountered in applying differential privacy to sparse datasets called *Filter Priority*. It addresses the inefficiency of traditional methods that add noise to contingency tables, leading to impractical increases in data size. Instead, the proposed framework generates compact summaries of noisy data directly from the input, circumventing the need for materializing the noisy data. This approach drastically reduces summary size, making them significantly smaller and computationally faster compared to conventional methods. Moreover, the framework accommodates various data transformations like wavelets or sketches (sketches denote a data structure which can be thought of as a linear projection of the input) [19]. In addition to data summarization, the paper explores data reduction techniques such as filters and sampling, with a focus on their application to noisy contingency tables. Filters selectively retain significant data elements, while sampling involves the random selection of a subset of elements to approximate queries over the dataset. The paper also discusses data transformations, particularly Fourier [15] or Wavelet transforms [126] and filtering techniques, to ensure privacy preservation and mitigate errors in large range queries. The final evaluation consist on accuracy expressed as relative error on queries over different privacy settings (ε and δ) and query sizes.

In [136] authors present *PrivGene*: a differentially private model fitting using genetic algorithms [53, 65], the authors propose a novel approach to fitting data models under the constraints of differential privacy utilizing the principles of genetic algorithms to ensure privacy preservation while maintaining data utility. PrivGene works by generating and evolving a population of candidate solutions. Each candidate solution is evaluated for its fitness, which reflects its accuracy in approximating the original data distribution.

The algorithm then selects the fittest candidates and applies genetic operations such as crossover and mutation to generate a new population of solutions. This process repeats until a termination condition is met, such as a set number of generations or a satisfactory level of fitness. The authors demonstrate that the PrivGene approach can be applied to various types of models, including *Gaussian Mixture Models*, and can be used for both discrete and continuous data making it very versatile. They compare PrivGene with other differentially private model fitting techniques [20, 93, 108, 137]. The evaluation metrics used are *miss-classification rate* measured on SVM classification tasks, logistic regression and *intra-cluster variance* on *k-means Clustering* technique.

In references [6, 49, 76] authors investigate and provide several frameworks using copula functions [95]. A copula function is a statistical tool used to describe and model the dependence structure between random variables. Formally, it is a multivariate distribution function that combines univariate marginal distributions into a joint distribution. Authors in general try to overcome the limits of existing techniques for generating differentially private histograms or synthetic data when applied to high-dimensional and large-domain data facing precision issues due to increased perturbation error and computational complexity. To address this gap, the authors propose different approaches for differentially private data synthesis utilizing Copula functions for high-dimensional data.

Specifically in the first publication in thFe Copula context the authors of present *DP-Copula* [76] as two differentially private Copula functions building techniques, which describe the dependence between multivariate random vectors, and then sample synthetic data from this function. The advantage of this approach is that Copula functions allow the construction of multivariate joint distributions using one-dimensional marginal distributions and a correlation matrix, making them suitable and computationally acceptable for modeling complex dependencies within the high-dimensional data context. In [76] two methods for estimating the parameters of Copula functions with differential privacy are presented: maximum likelihood estimation (MLE) [100] and Kendall's τ estimation [81]. The paper provides formal proofs for the privacy guarantee and convergence properties of the proposed methods. Experiments were conducted on both real and synthetic datasets.

In the experimental evaluation of DPCopula, the authors compare its performance with four state-of-the-art methods: Privelet+ [126], Private Spatial Decomposition (PSD) [25], KD-hybrid methods [25], Filter Priority (FP) with consistency checks [25], and P-HP [2]. The evaluation is conducted using both real and synthetic datasets. The utility of DPCopula is compared with other methods concerning various differential privacy budgets and datasets characteristics and is evaluated using relative error and absolute error metrics for random range-count queries. The computation time of DPCopula is evaluated concerning time against data cardinality and dimensionality to assess its scalability.

In [6] the authors adopted a similar approach but highlighting the following differences in the implementation:

1. [6] does not rely on arbitrary orders for nominal attributes, allowing it to produce synthetic datasets with pair-wise attribute correlations to be closer to the original dataset while respecting DP constraints; [76] imposes an artificial order on categorical attributes, affecting pair-wise correlations between them and potentially leading to inaccurate synthetic data generation;

2. [6] handles small domain attributes without computational feasibility issues while [76] relies on attributes combinations which might be expensive;

3. [6] constructs the differentially private correlation matrix by adding noise to margins (margins refer to the marginal distributions of a subset of variables in a joint distribution) before computing the Pearson correlation matrix while [76] computed a DP correlation matrix from original data and then perturbating the correlation matrix;

4. [6] Authors encode all the categorical variables using one-hot-encoding while [76] encodes categorical variables based on their cardinality to provide a numerical representation (as a rank).

In [6]the authors evaluate their result as Absolute Error on random counting queries as they quantify the error in the synthetic datasets by measuring the absolute difference between the true query results and the results obtained from the synthetic datasets. Another

evaluation technique is qualitatively checking the Cumulative Distribution Function (CDF) authors visualize the error distribution using CDF plots, which show the cumulative probability of observing errors up to a certain magnitude helping in understanding the overall performance of their method across different queries and datasets.

In [49] the authors present a novel approach for synthesizing data using *vine copulas*, which are widely used models known for their interpretability and robustness, particularly in actuarial applications. The proposed method, named COPULA-SHIRLEY, leverages differentially-private training of vine copulas to generate synthetic data across arbitrary dimensions while preserving utility. Finally the results are evaluated involving statistical tests and classification tasks. For statistical tests, the authors use the *Kolmogorov-Smirnov (KS)* distance to measure the fidelity of distributions between the original and synthetic datasets. Additionally, they assess the variation in pairwise correlations using Spearman's rank correlation coefficient. In classification tasks, the authors train classifiers on both original and synthetic datasets, then evaluate their performance on a test set. The author train gradient boosting classifiers, specifically XGBoost [22] implementation for such evaluation task. In terms of accuracy they employ the Matthews Correlation Coefficient (MCC) [24] to measure classification quality: results with higher scores indicate better performance. For further analysis, the authors conduct linear regression tasks and compute the root mean square error (RMSE) to evaluate the success of the synthetic data in replicating the original data's regression patterns.

2.4.1 Overview of Most Common Evaluation Metrics

In the literature review we've gathered together several evaluation metrics. Following are the most used ones categorized by their aspects:

Data Quality Metrics

- **Total Variation Distance (TVD)**: Measures the difference between two probability distributions.

- **Kolmogorov-Smirnov (KS) distance**: Measures the maximum difference between the cumulative distribution functions of two samples.
- **Spearman's rank correlation coefficient**: Measures the correlation between two ranked variables.
- **Correlation Analysis**: Spearman's Rank Correlation and Pearson's correlation coefficient to measure correlation based on the typo of data.

2.4.2 Data Utility Metrics

- **Misclassification Rate**: Measures the proportion of incorrect predictions in a classification task.
- **Accuracy Score**: Measures the proportion of correct predictions in a classification task.
- **Earth Mover's Distance**: Measures the dissimilarity between two frequency distributions.
- **Matthews Correlation Coefficient (MCC)**: Measures the quality of binary classification.
- **Mean Squared Error**: Measures the average squared error in range count queries.
- **Root Mean Square Error (RMSE)**: Measures the average magnitude of errors in a regression task.
- **Relative Error**: Measures the percentage difference between the estimated and actual values.
- **Absolute Error**: Measures the difference between the estimated and actual values.
- **Squared Error**: Measures the squared difference between the estimated and actual values.

- **Relative Entropy (Kullback-Leibler (KL) divergence)**: Measures the difference between two probability distributions.
- **Intra-cluster variance**: Measures the dispersion of data points within clusters.

In Chapter 2 we've investigated the literature on differentially private synthetic data generation it is possible to note that there is a notable absence of consistent and comprehensive evaluation metrics for the generated data and thus for the models. Evaluation methods vary widely across different publications, making it challenging to gain a general overview of the effectiveness of various generative techniques. In this chapter authors also identified two fundamental aspects of data evaluation for developing a proper and complete methodology: **data quality and data utility**.

One of the primary reasons for this lack of consistency is the diversity of use cases and objectives in privacy-preserving data synthesis. Different applications may prioritize different aspects of data utility and privacy protection, leading to the adoption of varied evaluation metrics tailored to specific contexts. As a result, there is no standard set of metrics that researchers universally apply to assess the performance of differentially private generative models.

Chapter 3

A Novel Evaluation Metric for Synthetic Data Evaluation

3.1 Motivation

The absence of standardized evaluation frameworks makes it difficult to compare the effectiveness of different methods across studies. Without consistent metrics and evaluation procedures, it becomes challenging for researchers to draw meaningful comparisons or identify best practices in the field of differentially private data synthesis.

Overall, the absence of consistent and comprehensive evaluation metrics underscores the need for further research to establish standardized evaluation methodologies. Such efforts would facilitate more robust comparisons between different generative techniques and advance the development of effective privacy-preserving data synthesis approaches with the privacy aspect covered by Differential Privacy while keeping the data into consideration.

3.2 Overview

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. We 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 DP-GroupFields 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.

3.3 The Proposed Framework

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, ordinal or nominal values. 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*.

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 correlationv 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.

Our metric G_ε is a composition of the above-listed characteristics.

Privacy Guarantee

The privacy guarantee is a parameter to quantify the privacy budget and it is the direct consequence of ε-*differentially private* mechanism definition introduced in [38] and well formalized in [39]. Namely epsilon is a guaranteed boundary of privacy loss. The term ε as in Definition 2.1.4 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 calibrated random noise when learning the probability distributions of the data-points based on the learning function's sensitivity; 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.[1]. Given a fixed value of ε different models produce different outputs due to several reasons: sampling error, sampling techniques, internal representation of the attributes distributions, sensitivity to noise, different handling of numerical and categorical variables.

[1]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

Macro-Statistics

As researching in the literature for generic correlation coefficients able to fairly capture correlations among heterogeneous data types, in reference [8], we've found a valuable and appropriate contribution in which 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 which is crucial when dealing with complex real-world datasets. Moreover, it reverts to the Pearson correlation coefficient in case of a bi-variate normal input distribution. These are fundamental aspects 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

$$\mu(D_s, D_p) = \frac{\|\phi_k(D_s) - \phi_k(D_p)\|_2}{m(m-1)/2} \tag{3.1}$$

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.

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). Here acc can be an arbitrary accuracy measure such that miss-classification rate, AUC for classification and R^2 for regression tasks or intra-cluster variance if the Machine Learning/Data Mining task is clustering 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

$$\delta(D_s, D_p) = \frac{1}{mKL} \sum_{i=1}^{m} \sum_{k=1}^{K} \sum_{l=1}^{L} \| acc^l(M^k_{X_i,D_s}) - acc^l(M^k_{X_i,D_p}) \|_2 \tag{3.2}$$

The Combined Metric

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

$$G_\varepsilon = \alpha\mu(D_s, D_p) + \beta\delta(D_s, D_p) \tag{3.3}$$

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.

3.4 Experimental Setup and Results

To evaluate our method, we've selected *PrivBayes* [134], which is based on *Bayesian Networks* (implemented in Python 3.7) and *DPGroupFields*. The latter was mentioned in [14] and was among the winners of NIST's *Differential Privacy Synthetic Data Challenge*

2018. In this competition, participants were tasked with developing a formally proven DP SDG model, which was then evaluated based on the quality of the data through count queries. DPGroupFields has been 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 we've selected two datasets. The selection of the datasets is contextual to the most probable fields of application, indeed, both our datasets 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 1 categorical-binary for classification for a total of 9 attributes and holding 768 records. The *adults* data-set composed by 14 numerical and categorical attributes (majority) and 1 categorical-binary (for classification). It is important to mention that at generation time the same number of records of the original datasets have been generated, namely: $n_s == n_p$ for all experiments. 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 0.22.2*. For a matter of brevity we're going to show the most relevant results.

For sake of simplicity, we used the following settings: $K = 1$ and $L = 1$ in equation (3.2), $\alpha = 1$ and $\beta = 1$ in equation (5.1). While for what concerns the values of the privacy budget ε (parameter of the SDG routines at generation time) the following values were 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 were 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 references [134] and [76]. 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). For what concerns L in our settings we set its value to 1 as the *miss-classification rate* over categorical attributes.

3.4.1 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 dataset both making use of PrivBayes [134] and DPFieldGroups [14], then we computed the average values of μ for each value of ε as plotted in Fig. 3.1. As expected the distance in terms of macro-statistics defined in equation 3.1 grows as the magnitude of the matrices of the difference among correlation coefficients defined as ϕ in 3.1 grows. Generally, as expected, we observes 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 deteriorate more and more between the original dataset and the synthetic ones. In Fig. 3.1 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 on the Adults dataset, this behavior is due to the binning of continuous values performed by PrivBayes.

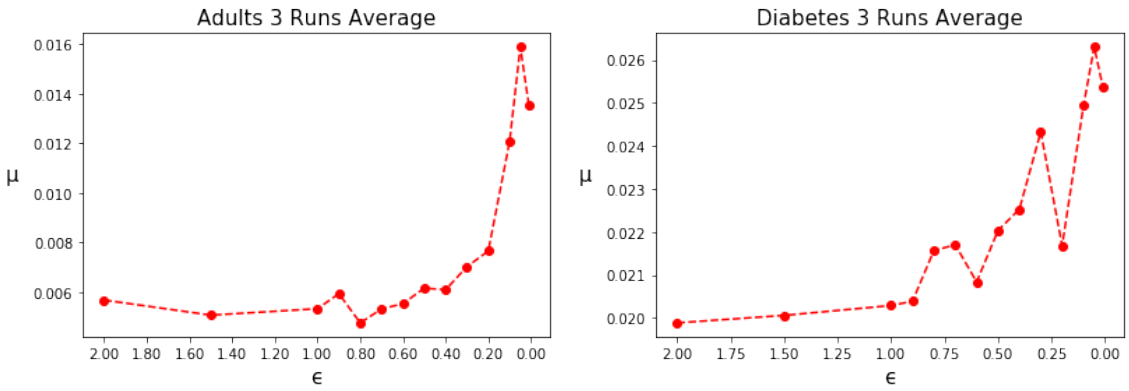


Fig. 3.1 Values of μ against the two datasets Adults and Diabetes using PrivBayes synthetic data generation method.

3.4.2 Data Utility δ

The δ factor exhibited a similar behavior as μ but with a slightly different magnitude. This characteristic justifies the presence of the two constants α and β in equation 5.1. Also in

this case over our experimental setup were 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 equation 3.2 tends to grow. Generally, it is possible to observe that values of δ tends to grow at 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. 3.2 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 binning 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.

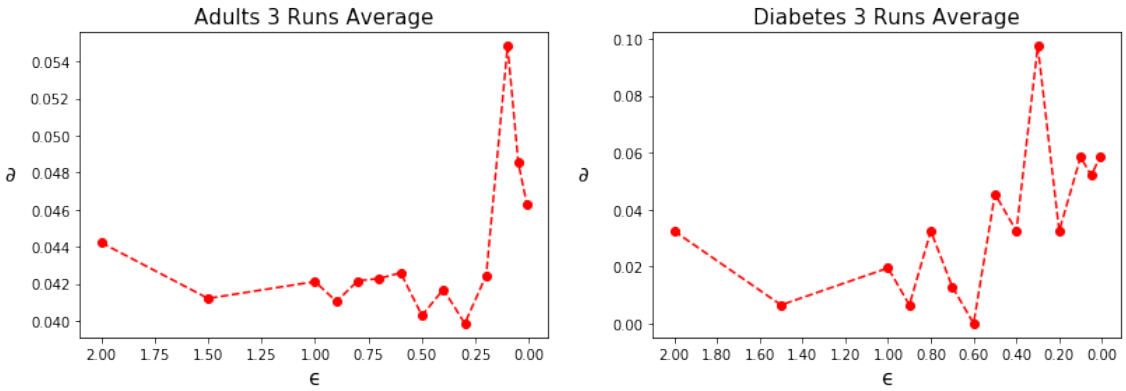


Fig. 3.2 Values of δ against the two datasets Adults and Diabetes using PrivBayes for both as synthetic data generation method. Values of δ tend to grow *"faster"* due to the nature of PrivBayes which splits continuous intervals.

3.4.3 Composite Measure G

Finally we calculated the values of composite measure G defined by equation 5.1 for $\alpha = 1$ and $\beta = 1$

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 datasets holding a double digit number of attributes. Thus, we state that our method reflects and confirms earlier literature results [134], [76], [77]. Figure 3.3, instead, shows results of G_ε regarding just PrivBayes since DPFieldGroups has not been designed to handle attributes holding floating numbers (namely not integers). Also in this case the effect of continuous attributes splitting has a noticeable impact on the stability of the score.

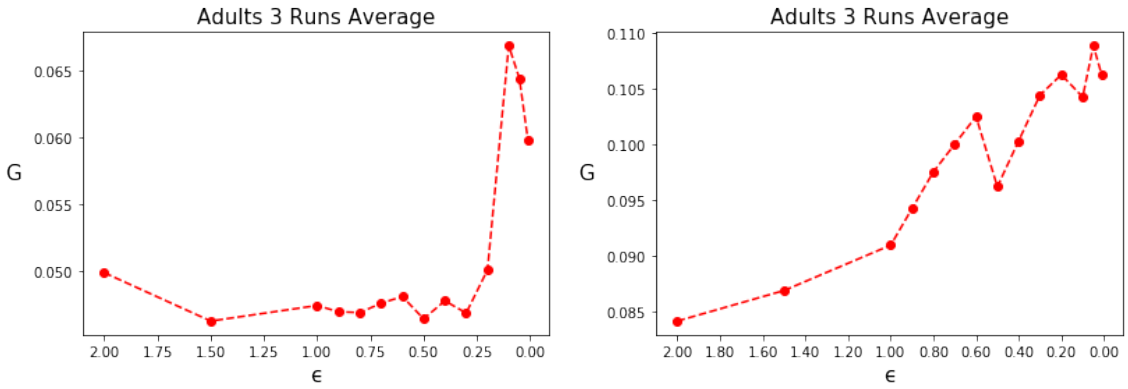


Fig. 3.3 Values of G_ε over the two algorithms PrivBayes (left) and DPFieldGroups (right) deployed on Adults dataset. Within this setup PrivBayes clearly keeps a better data utility over varying ε (lower G).

3.5 Conclusions and Summary

In this chapter we highlighted the literature gap regarding SDG evaluation metrics and methodologies identifying the lack of a unified evaluation methodology and framework. At the same time, as our contribution we proposed a novel composite and comprehensive evaluation metric G_ε for quantitatively measuring synthetic data generation solutions. The metric takes into account dataset statistical similarities and data utility, against privacy

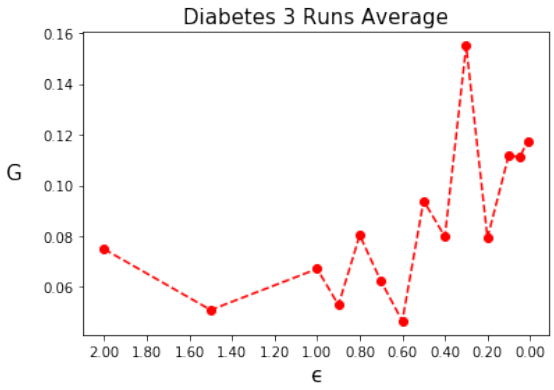


Fig. 3.4 Values of G_ε over Diabetes dataset using PrivBayes as generative algorithm.

budget ε, between synthetically generated data and the original data. We tested the newly introduced evaluation metric against two datasets comparing two different differentially private synthetic data generation algorithms. The results are consistent with literature and this contribution might 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.

Chapter 4

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

4.1 Motivation

Once introduced a comprehensive evaluation framework in Chapter 3, our aim was to compare different but related SDG techniques we performed a comparative evaluation of the statistical similarities and utility of a specific set of related algorithms in the realistic context of credit-risk and banking. More specifically we compared PrivBayes [134], Copula-Shirley [49], and DPCopula [76] algorithms and their variants using the previously proposed evaluation framework across three different datasets. The purpose of this study was 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. Finally the aim of our work is to cover the lack of comprehensiveness of two proposed evaluation frameworks available to date: DPBench [63] which uses only counting queries and [29] using attributes ranges, counts and macro-statistics such as Pearson correlation. Furthermore we aim to

contribute to the literature comparing a specific family of algorithms Copula Based, *DAG* (Directed Acyclic Graph) Based and Hybrid version of them.

Our evaluation framework G_ε introduced in [46] and described in Chapter 3 covers 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.

4.2 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 with shared properties. The selection of the privacy algorithms had 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 4.1 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 4.1 provides a summary of the key characteristics of each algorithm. All the selected algorithms are interrelated because all of them share at least one common aspect.

The interrelation of the selected generative methods can be further understood by examining their shared use of marginal probabilities. While each method employs slightly different approaches, they all rely on marginal distributions to represent the individual characteristics of attributes within the data but not only Copula Shirley shares also the use of bivariate correlation matrices with Gaussian copula and finally shares both marginals and a DAG srtucture with PrivBayes.

- **Gaussian Copula:** This method utilizes marginal distributions and a correlation matrix to capture the relationships between attributes. The correlation matrix provides information about the dependencies between variables, while the marginal distributions describe the individual characteristics of each attribute.

- **Copula-Shirley:** This method extends Gaussian Copula by incorporating bivariate correlation matrices and a tree structure (DAG). The DAG provides a more flexible representation of dependencies, allowing for complex relationships between

attributes. This hybrid approach combines the strengths of Gaussian Copula and DPBayes.

- **DPBayes:** This method uses marginal distributions and a DAG structure to model dependencies. The DAG represents the causal relationships between attributes, while marginal probabilities provide information about the distributions of the attributes.

4.2.1 PrivBayes

PrivBayes [134] 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.

4.2.2 DPCopula and Gaussian Copula

DPCopula [76] 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.

4.2.3 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 [49], 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.

4.3 Proposed Framework

Within this context, the evaluation framework is similar to the one introduced in [46], 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:

$$G_\varepsilon = \alpha\mu(D_s, D_p) + \beta\delta(D_s, D_p) \tag{4.1}$$

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.

4.3.1 Macro Statistics

The proposed macro-statistics measure μ is computed using ϕ_k introduced in [8], as it represents a practical correlation coefficient for heterogeneous datasets where m denotes the number of attributes of the dataset.

$$\mu(D_s, D_p) = \frac{\|\phi_k(D_s) - \phi_k(D_p)\|_2}{m(m-1)/2} \tag{4.2}$$

In equation 4.2 we compute the L_2 norm of the difference of the correlation matrices ϕ_k computed on 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.

4.3.2 Data Utility

The data utility measure δ is calculated as

$$\delta(D_s, D_p) = \frac{1}{mKL} \sum_{i=1}^{m} \sum_{k=1}^{K} \sum_{l=1}^{L} \|acc^l(M^k_{X_i,D_s}) - acc^l(M^k_{X_i,D_p})\|_2 \tag{4.3}$$

Where as described more in depth in [46] m is the number of Machine Learning Tasks (one per attribute in [46]), 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.

4.4 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) [67, 132] dataset mostly consists of numerical attributes. The *Adults Census* (Adults) [106] and *Credit Approval* (CRX) [103] datasets are mostly composed of categorical attributes, but differ in their sizes as shown in table 4.2. The *Financial Services* (Fin Services)

[OTP Bank] 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 [36] while *Financial Services* comes from a private (intended as not publicly available) and anonymized dataset gently donated by 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 (4.3), $\alpha = 1$ and $\beta = 1$ in equation (4.1). 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 were selected: (0.05, 0.1, 0.2, 0.4, 0.8, 1.6).

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

4.4.1 G Score

The G_ε score for varying ε can be observed in figures 4.1 and 4.2. While for the ML performances the data utility term δ in equation 4.3 over different values of ε can be observed in figures 4.3 and 4.4. 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.

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.

Observing the G_ε score *it is possible to observe that no algorithm clearly dominates* as in figures 4.1 and 4.2. 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 (4.1).

It is worth to note that all of the algorithms performed worst on Default Credit dataset which mostly consists on numerical attributes except for the target class. This outcome can be explained by the fact that correlations among attributes are not linear and most models used to validate the data utility can't really capture non-linear relations between features. Furthermore, 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 4.2 Properties of each dataset.

4.4.2 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 4.3 and 4.4). 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.

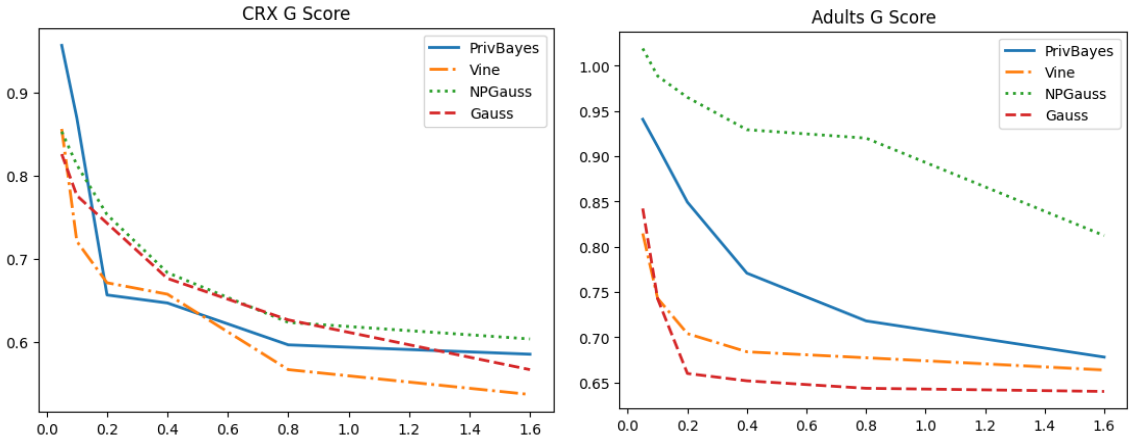


Fig. 4.1 Values of G_ε over the four algorithms deployed on CRX and Adults datasets (a lower value G_ε is better as it means that the synthetic dataset is similar to the original private one). The x axes represent values of ε while the y axes represent G_ε.

4.4.3 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 (4.2) 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 4.5 gives an illustration of this outcome. In Fig. 4.5 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.

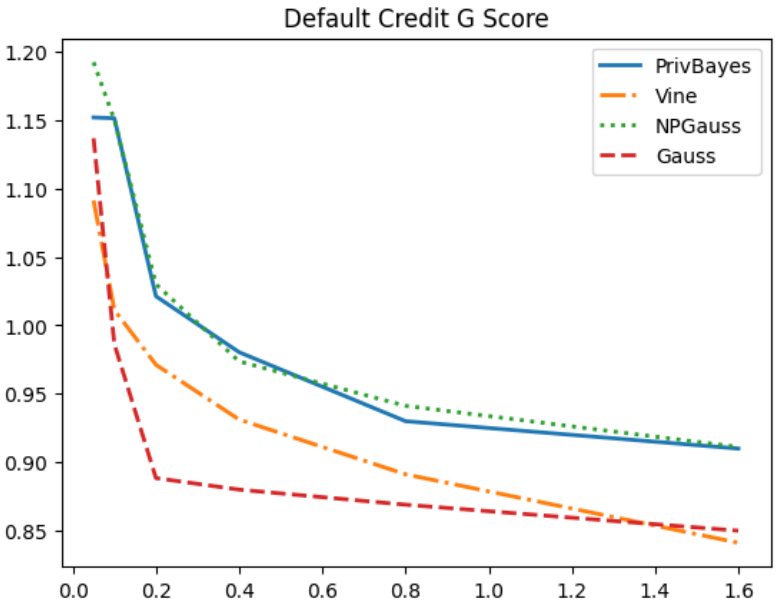


Fig. 4.2 Values of G_ε over Default Credit dataset (a lower value G_ε is better as it means that the synthetic dataset is similar to the original one). The x axis represents values of ε while the y axis represents G_ε.

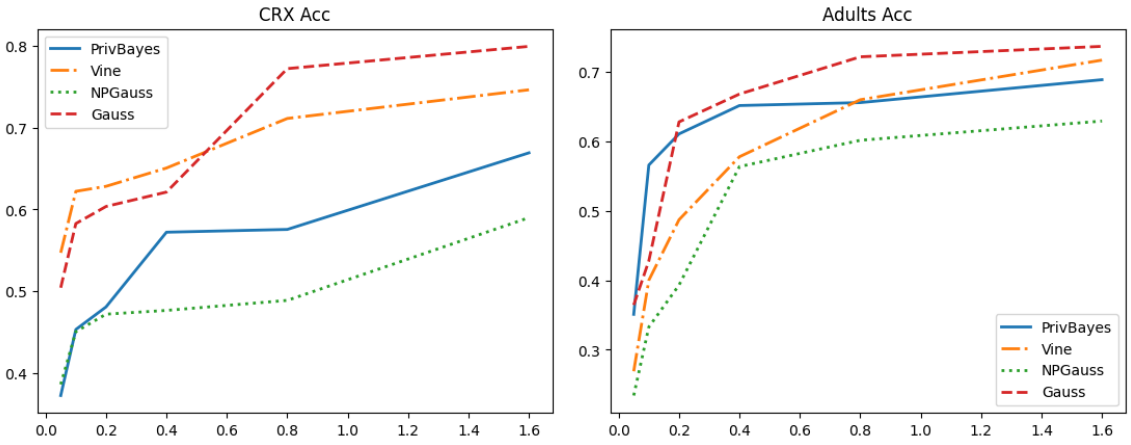


Fig. 4.3 Average values of Acc_ε over the four algorithms deployed on CRX and Adults datasets (higher Acc is better as its values get closer to the values achieved on the original private datasets). The x axes represent values of ε while the y axes represent the average Acc_ε.

4.5 Summary and Conclusions

We performed a benchmark of differentially-private synthetic data generation (DP SDG) algorithms for tabular data with heterogeneous attributes in the specific field of finance

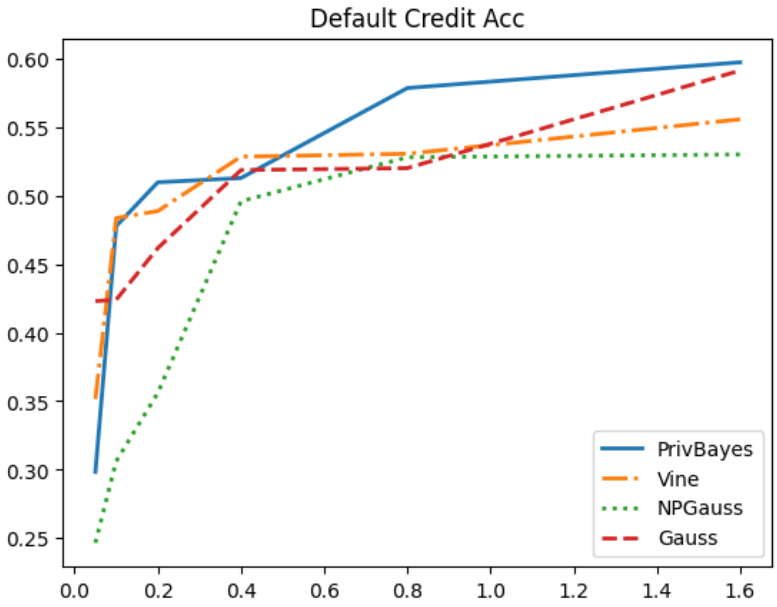


Fig. 4.4 Average values of Acc_ε over Default Credit dataset (higher Acc is better as its values get closer to the values achieved on the original private datasets). The x axes represent values of ε while the y axes represent the average Acc_ε.

and credit on a specific family of DP SDG techniques. 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.

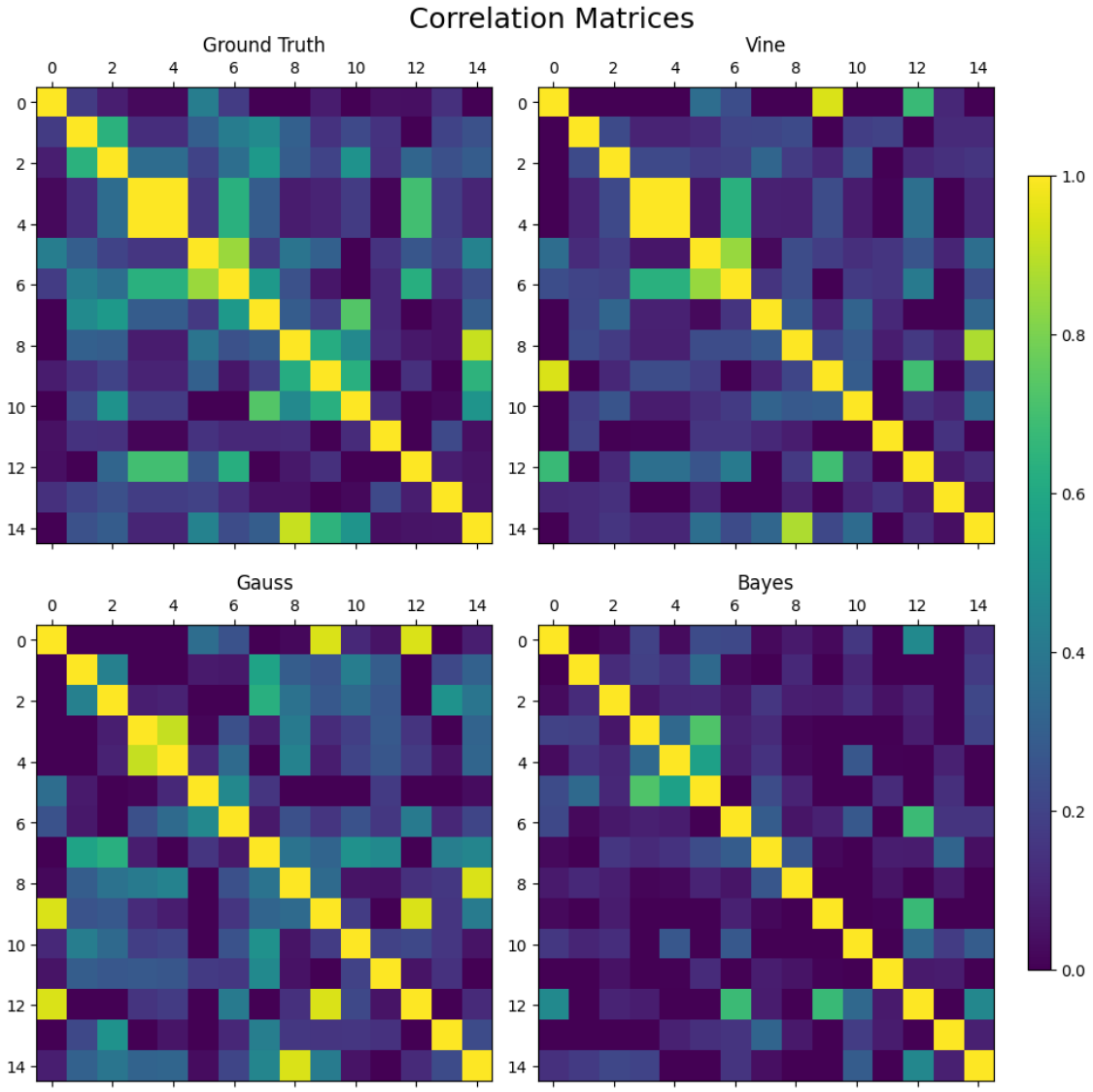


Fig. 4.5 Correlation matrices over Fin Services dataset given a fixed value of $\varepsilon = 0.8$, x and y axes are the attributes while the values of ϕ_k between each attribute are represented by the colored cells of the matrix.

Chapter 5

Extending Synthetic Data Evaluation Metrics

5.1 Motivation

Once introduced a comprehensive evaluation framework in Chapter 3, and after comparing different interrelated SDG techniques in Chapter 4, this Chapter presents an improvement of the evaluation metric G_ε called G_ε^+. Specifically, in this chapter we extend our previously introduced evaluation frameworks including also the evaluation of the distances among the distributions of the attributes.

5.2 The Proposed Framework and Improvements

As previously stated, our contribution presented in this chapeter is an extension of G_ε [46], 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 $n_p = n_s$).

5.2.1 G_ε – An Almost Comprehensive Evaluation Metric

A comprehensive method G_ε is defined in [46] as:

$$G_\varepsilon = \alpha\mu(D_s, D_p) + \beta\delta(D_s, D_p) \tag{5.1}$$

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 [38]. 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.

Macro-statistics (μ)

The macro-statistics term is mainly based on [8] in which a practical and versatile correlation coefficient ϕ_k was introduced as an improvement over different correlation analysis metrics of two variables, which can be applied consistently across categorical, ordinal and interval variables. The authors of [8] demonstrated 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 also took into account the importance of properly evaluating the statistical significance of correlations, especially when dealing with low statistics samples (small datasets). 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:

$$\mu(D_s, D_p) = \frac{\|\phi_k(D_s) - \phi_k(D_p)\|_2}{m(m-1)/2} \tag{5.2}$$

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

Data Utility (δ)

We assume the dataset D_s is primarily intended for analytical purposes and can be used for various machine learning tasks. As the authors of [46] stated, 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:

$$\delta(D_s, D_p) = \frac{1}{mKL} \sum_{i=1}^{m} \sum_{k=1}^{K} \sum_{l=1}^{L} \|acc^l(M^k_{X_i,D_s}) - acc^l(M^k_{X_i,D_p})\|_2 \tag{5.3}$$

5.2.2 Improved Metric G_{ε}^{+}

In this chapter we propose an enhanced version of the generative mechanism, denoted as G_{ε}^{+}. In order to provide a contextualization of our contribution, it is necessary to first elucidate the limitations of G_{ε}. Specifically, authors of [46] 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. Specifically, we advocate for the inclusion of attribute distribution distance metrics, such as the Total Variation Distance (TVD), in the evaluation process. This will provide a more comprehensive and robust evaluation of the synthetic data, covering all its relevant aspects. Authors selected TVD over KS distance due to several reasons, indeed TVD often offers several advantages over the Kolmogorov-Smirnov (KS) Test in comparing probability distributions:

1. *Sensitivity to Local Differences:* TVD is more sensitive to local differences in the probability mass function. This means it can better detect differences that might be missed by KS, especially when the overall shape of the distributions is similar but there are significant differences in specific regions.

2. *Direct Measure of Difference:* TVD provides a direct measure of the difference between two distributions, making it easier to interpret and compare results. KS, on the other hand, is a statistical test that provides a p-value, which can be less intuitive to understand.

3. *Robustness to Censored/Missing Data:* TVD can be more robust to censored data, where some observations are missing. This is because TVD focuses on the entire probability mass function, while KS is based on the cumulative distribution function, which can be affected by missing data.

4. *Applications in Machine Learning:* TVD is frequently used in machine learning tasks like evaluating generative models and comparing datasets. This is because it provides a direct measure of the distance between two distributions, which is often useful in these applications.

In summary TVD is often a more versatile, easy to interpret, reliable and informative metric for comparing probability distributions if compared to KS, especially when there is the need to detect local differences and it is more reliable when data contains outliers [68, 12].

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 [46] 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 [90]. Thus, Pearson's correlation and ϕ_k [8] are not significantly affected by linear transformations, leaving room for improvements.

Furthermore, the authors of [86] (*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 [46] 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 [86], 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 [86]. 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-1* μ such as correlation analysis;
2. *Data Utility* δ in terms of ML performances if applicable;
3. *Single-Attribute Distribution Distance* (1-TVD) ν.

5.2.3 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.

1. Definition. *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 $TVD(P,Q) = \sup_{A\in\mathcal{A}} |P(A) - Q(A)| = \sup_{A\in\mathcal{A}} |\int_A (p-q) d\nu|$ 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 [124].

When dealing with discrete Probability Density Functions (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.

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

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^s_i), computing this metric against the original (X^p_i), such as

$$\nu(D_p, D_s) = \frac{1}{m}\sum_{i=1}^{m} \text{TVD}(P(X^p_i), Q(X^s_i)) \tag{5.4}$$

5.2.4 The Combined Metric G^+_ε

As in [46], we do aim to construct a combined overall metric, such that

$$G^+_\varepsilon = \alpha\mu(D_s, D_p) + \beta\delta(D_s, D_p) + \gamma\nu(D_p, D_s) \tag{5.5}$$

where α, β and γ are weights allowing to designate the importance of macro-statistics similarity (μ), data utility similarity (δ) and total variation distance (ν), while ε is the privacy budget as defined in the DP framework [38].

5.3 Experimental Results

Due to space constraints, we present only the most relevant findings. All the algorithms were run using a computer equipped with an Intel *CPU i7-7500U@2.70GHz* and *16GB RAM DDR4*. All the ML tasks were deployed using Python's *Scikit-Learn 0.24.1* and *NumPy 1.20.1*.

We carried out two different kind of experiments: First, we compared 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 compared 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 set $\alpha = \beta = \gamma = 1$ in formula 5.5.

5.3.1 Synthetic Data

In order to emphasize the need of a probability distribution distance to create a more comprehensive metric, we carried out the following specific transformations (Table 5.1 and Figure 5.1): two kind of rotations, $\pi/2$ and π, a translation of the vector space (we added a constant to the whole dataset, namely the sum of the averages of the two variables) and, finally, we applied a nonlinear transformation using the algorithm presented in [86]. For the sake of completeness we included 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^+ converges to 0. While, for the π rotation, both coefficients 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 out 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 5.1 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 [86], 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.

5.3.2 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 (5.3). 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 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, this latter observation it is showed on the bold columns in table 5.1.

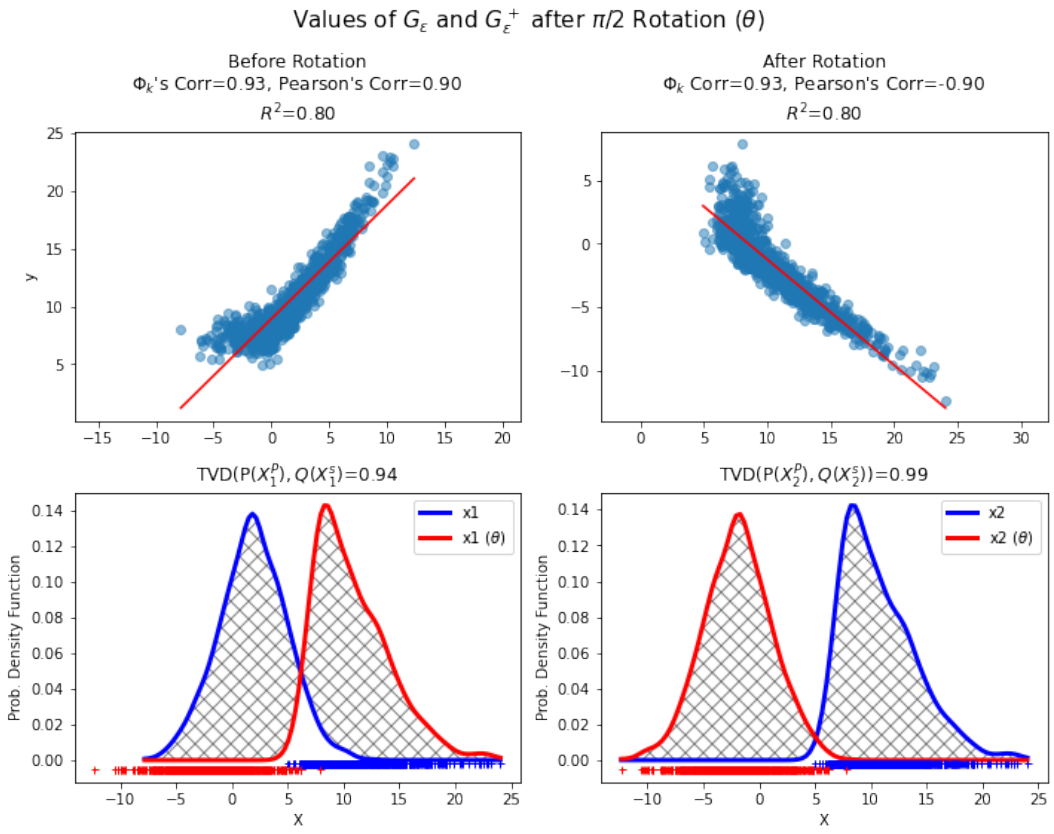


Fig. 5.1 Illustrative example of dataset rotation and the TVD distance.

Table 5.1 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^+

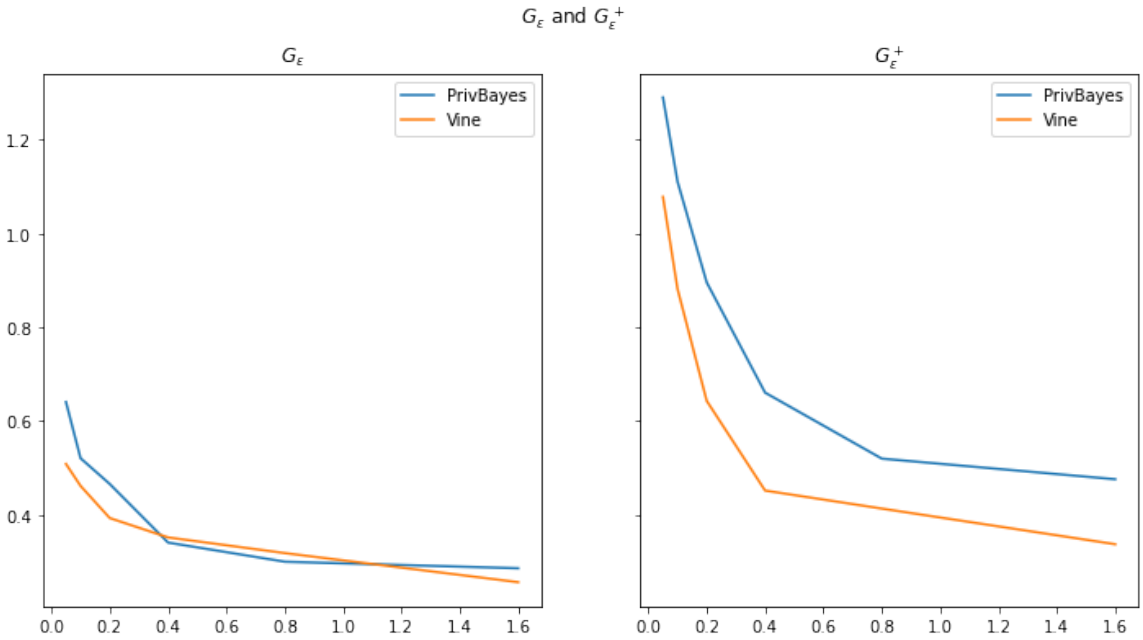


Fig. 5.2 The two scores over the same datasets

5.4 Conclusions and Summary

We proposed an improved version of G_ε [46], 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 tested the introduced evaluation metric against two datasets and by comparing two distinct differentially private synthetic

data generation algorithms. The results are consistent with previous work and represent an improvement in reliability and comprehensiveness, 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.

Chapter 6

Conclusion and Recommendation

6.1 Summary and Contributions

6.1.1 Contributions

Chapter 1 Introduction and Motivation: The introductory chapter outlines the motivation for this dissertation, emphasizing the importance of synthetic data generation and its evaluation, especially within the framework of differential privacy. The chapter delineates the core research tasks:

1. Identifying essential metrics for evaluating synthetic data quality and utility.
2. Establishing a standardized methodology for evaluating synthetic data and generative mechanisms.
3. Developing a comprehensive and heterogeneous scoring system for synthetic data evaluation.

Chapter 2 Differential Privacy and Synthetic Data Generation: This chapter delves into the evolution of data anonymization techniques, with a specific focus on differential privacy (DP). It explains DP notation, theoretical concepts, privacy mechanisms, and their properties. The chapter also explores synthetic data generation (SDG) and the integration of DP with SDG, termed Differentially Private Synthetic Data Generation (DP SDG). A

review of related work highlights the current state of research and identifies gaps addressed by this dissertation.

Chapter 3 A Novel Evaluation Metric for Synthetic Data Evaluation: This chapter introduces a new evaluation framework for synthetic data, designed to address the limitations of existing methods. The proposed framework incorporates macro-statistics, and machine learning performance metrics to provide a comprehensive assessment of synthetic data quality and utility. The chapter details the experimental setup and results, demonstrating the effectiveness of the evaluation metric G_ε through various case studies and comparisons.

Chapter 4 Differentially Private Copulas, DAG, and Hybrid Methods: A Comprehensive Data Utility Study In this chapter, the dissertation evaluates several differentially private generative models, including copulas, Directed Acyclic Graphs (DAGs), and hybrid methods. Using the developed evaluation framework, the chapter assesses these models' strengths and limitations across different datasets and privacy requirements. The analysis provides insights into the practical application of these models, highlighting scenarios where each method excels or falls short.

Chapter 5 Extending Synthetic Data Evaluation Metrics This Chapter focuses on refining the previously introduced evaluation metric. It proposes enhancements to make it more resilient and comprehensive by introducing advanced metrics such as the probability distribution function distance. The updated framework, referred to as G_ε^+, improves the robustness and applicability of synthetic data evaluation, ensuring more accurate and reliable assessments.

6.1.2 Contributions

In this section we do summarize the Scientific and Engineering contributions, for a more detailed overview of the contributions refer to section 1.3.

1. **Contribution on Data Science: Literature Review on Data Evaluation and Analysis Fundamental Criteria:** the authors conducted a comprehensive literature

review to identify essential aspects for a thorough synthetic data evaluation process. They identified two primary criteria:

(a) **Data Quality**

(b) **Data Utility**

The authors initially released a version of their synthetic data evaluation method, G_ε, which included metrics to measure both aspects. The method for evaluating data quality includes correlation analysis between the attributes of the private and synthetic datasets using the correlation coefficient (ϕ_k). It also encompasses machine learning performance 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. The results of this research validated the reliability of the method.

2. **Contribution on Data Science: Methodology Design, Development, Validation and Extension:** the authors conducted further research to identify limitations in their previously presented evaluation method, G_ε. They found that G_ε was insensitive to linear transformations and had limited consideration of attribute ranges. Additionally, it was insensitive to certain data generation techniques that preserve pairwise correlation but alter statistical distribution values.

   To address these limitations, the authors extended their research and identified two properties for a more comprehensive evaluation of the data quality aspect:

   (a) **Macro-Statistical Properties**

   (b) **Statistical Distribution Properties**

   The authors introduced a definitive methodology and framework for evaluating synthetic data and their related generators, G_ε^+. This framework includes the Total Variation Distance (TVD) to address the limitations of G_ε in capturing the impact of linear transformations or changes in attribute distributions. The results of this research validated the reliability of the improved methodology.

3. **Contribution on Software Engineering and Data Science: Experiment Design and Validation:** the authors conducted an extensive evaluation of a set of Differentially Private generative mechanisms. They used 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.

The findings of this dissertation have significant implications for the field of synthetic data generation. The proposed evaluation metrics offer a standardized and comprehensive approach for assessing the quality and utility of synthetic data, enabling researchers and practitioners to make informed decisions about the most suitable generative techniques for their specific use cases. Moreover, the insights gained from the comparative evaluation of different differentially private generative mechanisms contribute to the ongoing development and refinement of privacy-preserving synthetic data generation techniques.

In conclusion, this dissertation advanced the understanding and evaluation of synthetic data generators, particularly in the context of differential privacy. The proposed evaluation metrics and the comparative analysis of different generative mechanisms provide valuable tools and insights for researchers and practitioners in this field. The research presented here has the potential to significantly impact the development and adoption of privacy-preserving synthetic data generation techniques, ultimately contributing to the broader goal of responsible and ethical data utilization.

6.2 Future Research Directions

As future research directions we do propose the following:

Refinement of Evaluation Metrics Weights: Future research should focus on refining the proposed evaluation by finding optimal weights for the components of G_ε^+ and define a standardized but diverse set of machine learning algorithms based on data characteristics.

This could be depending on the type of data, the context and the type of use of the synthetic data.

Application to Diverse Data Domains: The framework should be applied to a broader range of data domains: telecommunication, healthcare, finance, and social sciences. This would help in understanding the generalizability of the evaluation metrics and in identifying domain-specific challenges and solutions.

Adaptive and Automatic Privacy Budget Selection: Research should investigate adaptive privacy mechanisms that dynamically adjust privacy parameters based on the sensitivity of the data but also taking into consideration the requirements of the application. This can help in achieving a better balance between privacy and data utility in a fully automated process.

Extending the developed scores to evaluate any generative algorithms: As the evaluation framework tends to be generic and comprehensive. Researchers should investigate the robustness of the developed frameworks against non differentially private generative techniques in order to demonstrate its effectiveness and adaptability beyond the privacy domains: e.g.: data augmentation.

References

[1] Abowd, J. M. and Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202.

[2] Acs, G., Castelluccia, C., and Chen, R. (2012). Differentially private histogram publishing through lossy compression. In *2012 IEEE 12th International Conference on Data Mining*, pages 1–10. IEEE.

[3] Ahmed, M. A. and Hermadi, I. (2008). Ga-based multiple paths test data generator. *Computers & Operations Research*, 35(10):3107–3124.

[4] Appenzeller, A., Leitner, M., Philipp, P., Krempel, E., and Beyerer, J. (2022). Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, 12(23):12320.

[5] Arachchige, P. C. M., Bertok, P., Khalil, I., Liu, D., Camtepe, S., and Atiquzzaman, M. (2019). Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7):5827–5842.

[6] Asghar, H. J., Ding, M., Rakotoarivelo, T., Mrabet, S., and Kaafar, D. (2020). Differentially private release of datasets using gaussian copula. *Journal of Privacy and Confidentiality*, 10(2).

[7] Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., and Veloso, M. (2020). Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8.

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

[9] Balle, B. and Wang, Y.-X. (2018). Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR.

[10] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). 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*, pages 273–282.

[11] Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

[12] Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, pages 445–463.

[13] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

[14] Bowen, C. M. and Snoke, J. (2019). Comparative study of differentially private synthetic data algorithms and evaluation standards. *arXiv preprint arXiv:1911.12704*.

[15] Bracewell, R. N. and Bracewell, R. N. (1986). *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York.

[16] Bu, Z., Dong, J., Long, Q., and Su, W. J. (2020). Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23):10–1162.

[17] Cai, J., Liu, X., and Wu, Y. (2020). Svm learning for default prediction of credit card under differential privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 51–53.

[18] Chamikara, M. A. P., Bertok, P., Khalil, I., Liu, D., and Camtepe, S. (2020). Privacy preserving face recognition utilizing differential privacy. *Computers & Security*, 97:101951.

[19] Charikar, M., Chen, K., and Farach-Colton, M. (2002). Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer.

[20] Chaudhuri, K. and Monteleoni, C. (2008). Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21.

[21] Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3).

[22] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[23] Cheu, A., Smith, A., Ullman, J., Zeber, D., and Zhilyaev, M. (2019). Distributed differential privacy via shuffling. In *Advances in Cryptology–EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part I 38*, pages 375–403. Springer.

[24] Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13.

[25] Cormode, G., Procopiuc, C., Srivastava, D., Shen, E., and Yu, T. (2012a). Differentially private spatial decompositions. In *2012 IEEE 28th International Conference on Data Engineering*, pages 20–31. IEEE.

[26] Cormode, G., Procopiuc, C., Srivastava, D., and Tran, T. T. (2012b). Differentially private summaries for sparse data. In *Proceedings of the 15th International Conference on Database Theory*, pages 299–311.

[27] Cummings, R. and Desai, D. (2018). The role of differential privacy in gdpr compliance. In *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*, volume 20.

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

[29] DataCebo (2019). Sdgym. https://github.com/sdv-dev/SDGym.

[30] Deason, W. H., Brown, D. B., Chang, K.-H., and Cross, J. H. (1991). A rule-based software test data generator. *IEEE transactions on Knowledge and Data Engineering*, 3(1):108–117.

[31] Ding, B., Winslett, M., Han, J., and Li, Z. (2011). Differentially private data cubes: optimizing noise sources and consistency. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 217–228.

[32] Dinur, I., Stemmer, U., Woodruff, D. P., and Zhou, S. (2023). On differential privacy and adaptive data analysis with bounded space. In *Advances in Cryptology–EUROCRYPT 2023: 42nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Lyon, France, April 23-27, 2023, Proceedings, Part III*, pages 35–65. Springer.

[33] Domingo-Ferrer, J. and Soria-Comas, J. (2015). From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, 74:151–158.

[34] Dong, J., Roth, A., and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37.

[35] D'Orazio, V., Honaker, J., and King, G. (2015). Differential privacy for social science inference. *Sloan Foundation Economics Research Paper*, (2676160).

[36] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[37] Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer.

[38] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). 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.

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

[40] Dwork, C., Roth, A., et al. (2014b). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

[41] Edvardsson, J. (1999). A survey on automatic test data generation. In *Proceedings of the 2nd Conference on Computer Science and Engineering*, pages 21–28.

[42] El Emam, K. and Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637.

[43] Erlingsson, Ú., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067.

[44] Fan, L. (2020). A survey of differentially private generative adversarial networks. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence*, page 8.

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

[46] Galloni, A., Lendák, I., and Horváth, T. (2020). 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.

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

[48] Gambs, S., Killijian, M.-O., and del Prado Cortez, M. N. (2014). De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614.

[49] Gambs, S., Ladouceur, F., Laurent, A., and Roy-Gaumond, A. (2021). Growing synthetic data through differentially-private vine copulas. *Proceedings on Privacy Enhancing Technologies*, 2021(3):122–141.

[50] Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. (2007). Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*, pages 758–769.

[51] Ghosh, A., Roughgarden, T., and Sundararajan, M. (2009). Universally utility-maximizing privacy mechanisms. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 351–360.

[52] Girgis, A., Data, D., Diggavi, S., Kairouz, P., and Suresh, A. T. (2021). Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR.

[53] Goldberg, D. E. (1989). Genetic algorithms in search, optimization and machine learning addison welssey publishing company. *Reading, MA*.

[54] Gonzalo, M. G., Xiaoyuan, L., Floria, M., and Dawn, S. (2023). Lessons learned: Surveying the practicality of differential privacy in the industry. *PoPETs Proceedings*.

[55] Gov., U. (2020). 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.

[56] Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C.-A., Goronflot, T., Karakachoff, M., Limou, S., Vince, N., Wargny, M., et al. (2022). Patient-centric synthetic data generation, no reason to risk re-identification in the analysis of biomedical pseudonymised data.

[57] Gupta, A., Hardt, M., Roth, A., and Ullman, J. (2011). Privately releasing conjunctions and the statistical query barrier. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 803–812.

[58] Gursoy, M. E., Tamersoy, A., Truex, S., Wei, W., and Liu, L. (2019). Secure and utility-aware data collection with condensed local differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2365–2378.

[59] Ha, T., Dang, T. K., Dang, T. T., Truong, T. A., and Nguyen, M. T. (2019). Differential privacy in deep learning: An overview. In *2019 International Conference on Advanced Computing and Applications (ACOMP)*, pages 97–102.

[60] Hardt, M., Ligett, K., and McSherry, F. (2012). A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*, 25.

[61] Hardt, M. and Rothblum, G. N. (2010). A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st annual symposium on foundations of computer science*, pages 61–70. IEEE.

[62] Hassan, M. U., Rehmani, M. H., and Chen, J. (2019). Differential privacy techniques for cyber physical systems: a survey. *IEEE Communications Surveys & Tutorials*, 22(1):746–789.

[63] Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y., and Zhang, D. (2016). Principled evaluation of differentially private algorithms using dpbench. In *Proceedings of the 2016 International Conference on Management of Data*, pages 139–154.

[64] Hay, M., Rastogi, V., Miklau, G., and Suciu, D. (2009). Boosting the accuracy of differentially-private histograms through consistency. *arXiv preprint arXiv:0904.0942*.

[65] Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

[66] Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., and Roth, A. (2014). Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410. IEEE.

[67] I-Cheng Yeh (2009). Default of credit card clients data. [Online; accessed 31-March-2023].

[68] Ji, H., Ke, P., Hu, Z., Zhang, R., and Huang, M. (2023). Tailoring language generation models under total variation distance. *arXiv preprint arXiv:2302.13344*.

[69] Ji, S., Mittal, P., and Beyah, R. (2016). Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys & Tutorials*, 19(2):1305–1326.

[70] Ji, Z., Lipton, Z. C., and Elkan, C. (2014). Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*.

[71] Kairouz, P., Oh, S., and Viswanath, P. (2015). The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR.

[72] Kim, J. W., Edemacu, K., Kim, J. S., Chung, Y. D., and Jang, B. (2021). A survey of differential privacy-based techniques and their applicability to location-based services. *Computers & Security*, 111:102464.

[73] Kurapati, S. and Gilli, L. (2023). Synthetic data: A convergence between innovation and gdpr. *J. Open Access L.*, 11:1.

[74] Leoni, D. (2012). Non-interactive differential privacy: a survey. In *Proceedings of the First International Workshop on Open Data*, pages 40–52.

[75] Li, C., Miklau, G., Hay, M., McGregor, A., and Rastogi, V. (2015). The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB journal*, 24:757–781.

[76] Li, H., Xiong, L., and Jiang, X. (2014a). 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.

[77] Li, H., Xiong, L., Zhang, L., and Jiang, X. (2014b). Dpsynthesizer: differentially private data synthesizer for privacy preserving data sharing. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, volume 7, page 1677. NIH Public Access.

[78] Li, H., Yu, L., and He, W. (2019). The impact of gdpr on global technology development.

[79] Li, N., Li, T., and Venkatasubramanian, S. (2006). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE.

[80] Li, N., Qardaji, W., Su, D., and Cao, J. (2012). Privbasis: Frequent itemset mining with differential privacy. *arXiv preprint arXiv:1208.0093*.

[81] Lovric, M. (2011). International encyclopedia of statistical science. *(No Title)*.

[82] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es.

[83] Majeed, A. and Lee, S. (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE access*, 9:8512–8545.

[84] Maniar, T., Akkinepally, A., and Sharma, A. (2021). Differential privacy for credit risk model. *arXiv preprint arXiv:2106.15343*.

[85] Mateo-Sanz, J. M., Martínez-Ballesté, A., and Domingo-Ferrer, J. (2004a). Fast generation of accurate synthetic microdata. In *Privacy in Statistical Databases*, volume 3050, pages 298–306. Springer.

[86] Mateo-Sanz, J. M., Martínez-Ballesté, A., and Domingo-Ferrer, J. (2004b). Fast generation of accurate synthetic microdata. In *Privacy in Statistical Databases*, volume 3050, pages 298–306. Springer.

[87] McMinn, P. (2004). Search-based software test data generation: a survey. *Software testing, Verification and reliability*, 14(2):105–156.

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

[89] McSherry, F. D. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30.

[90] Merener, M. M. (2012). Theoretical results on de-anonymization via linkage attacks. *Trans. Data Privacy*, 5(2):377–402.

[91] Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275.

[92] Mohammed, N., Chen, R., Fung, B. C., and Yu, P. S. (2011). Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–501.

[93] Mohan, P., Thakurta, A., Shi, E., Song, D., and Culler, D. (2012). Gupt: privacy preserving data analysis made easy. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 349–360.

[94] Near, J. (2018). Differential privacy at scale: Uber and berkeley collaboration. In *Enigma 2018 (Enigma 2018)*, Santa Clara, CA. USENIX Association.

[95] Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

[96] Nissim, K., Orlandi, C., and Smorodinsky, R. (2012). Privacy-aware mechanism design. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 774–789.

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

[OTP Bank] OTP Bank. Financial services. Proprietary Data Set.

[99] Pai, M. M. and Roth, A. (2013). Privacy and mechanism design. *ACM SIGecom Exchanges*, 12(1):8–29.

[100] Pan, J.-X., Fang, K.-T., Pan, J.-X., and Fang, K.-T. (2002). Maximum likelihood estimation. *Growth curve models and statistical diagnostics*, pages 77–158.

[101] Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410. IEEE.

[102] Popić, S., Pavković, B., Velikić, I., and Teslić, N. (2019). Data generators: a short survey of techniques and use cases with focus on testing. In *2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin)*, pages 189–194. IEEE.

[103] Quinlan (2007). Credit approval data set. [Online; accessed 31-March-2023].

[104] Ravikumar, G., Manjunath, T., Hegadi, R. S., and Umesh, I. (2011). A survey on recent trends, process and development in data masking for testing. *International Journal of Computer Science Issues (IJCSI)*, 8(2):535.

[105] Research, F. (2020). Protecting privacy in facebook mobility data during the covid-19 response. https://research.facebook.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/. Accessed: 2023-04-01.

[106] Ronny Kohavi and Barry Becker (2019). Adult data set. [Online; accessed 31-March-2023].

[107] Roth, A. and Roughgarden, T. (2010). Interactive privacy via the median mechanism. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 765–774.

[108] Rubinstein, B. I., Bartlett, P. L., Huang, L., and Taft, N. (2009). Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*.

[109] Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121.

[110] Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

[111] Smith, J., Asghar, H. J., Gioiosa, G., Mrabet, S., Gaspers, S., and Tyler, P. (2021). Making the most of parallel composition in differential privacy. *arXiv preprint arXiv:2109.09078*.

[112] Song, S., Chaudhuri, K., and Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE.

[113] Soria-Comas, J. and Domingo-Ferrert, J. (2013). Differential privacy via t-closeness in data publishing. In *2013 Eleventh Annual Conference on Privacy, Security and Trust*, pages 27–35. IEEE.

[114] Stollnitz, E. J., DeRose, T. D., DeRose, A. D., and Salesin, D. H. (1996). *Wavelets for computer graphics: theory and applications*. Morgan Kaufmann.

[115] Sun, H., Xiao, X., Khalil, I., Yang, Y., Qin, Z., Wang, H., and Yu, T. (2019). Analyzing subgraph statistics from extended local views with decentralized differential privacy. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 703–717.

[116] Sun, Y., Yin, L., Liu, L., and Xin, S. (2014). Toward inference attacks for k-anonymity. *Personal and ubiquitous computing*, 18:1871–1880.

[117] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

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

[119] Thakurta, A., Vyrros, A. H., Vaishampayan, U. S., Kapoor, G., Freudiger, J., Sridhar, V. R., and Davidson, D. (2017b). Learning new words. https://www.google.com/patents/US9645998. US Patent 9,645,998.

[120] Torfi, A., Fox, E. A., and Reddy, C. K. (2022). Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500.

[121] Torkzadehmahani, R., Kairouz, P., and Paten, B. (2019). Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

[122] Tran, K.-N., Vatsalan, D., and Christen, P. (2013). Geco: an online personal data generator and corruptor. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2473–2476.

[123] Truta, T. M. and Vinay, B. (2006). Privacy protection: p-sensitive k-anonymity property. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 94–94. IEEE.

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

[125] Xiao, X. and Tao, Y. (2006). Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 229–240.

[126] Xiao, X., Wang, G., and Gehrke, J. (2010). Differential privacy via wavelet transforms. *IEEE Transactions on knowledge and data engineering*, 23(8):1200–1214.

[127] Xiao, Y. and Xiong, L. (2015). Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309.

[128] Xiong, P., Zhu, T., and Wang, X.-F. (2014). A survey on differential privacy and applications.

[129] Xiong, X., Liu, S., Li, D., Cai, Z., and Niu, X. (2020). A comprehensive survey on local differential privacy. *Security and Communication Networks*, 2020:1–29.

[130] Yang, M., Lyu, L., Zhao, J., Zhu, T., and Lam, K.-Y. (2020). Local differential privacy and its applications: A comprehensive survey. *arXiv preprint arXiv:2008.03686*.

[131] Yao, X., Zhou, X., and Ma, J. (2016). Differential privacy of big data: an overview. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pages 7–12. IEEE.

[132] Yeh, I.-C. and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480.

[133] Young, J., Graham, P., and Penny, R. (2009). Using bayesian networks to create synthetic data. *Journal of Official Statistics*, 25(4):549.

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

[135] Zhang, J., Xiao, X., and Xie, X. (2016). Privtree: A differentially private algorithm for hierarchical decompositions. In *Proceedings of the 2016 international conference on management of data*, pages 155–170.

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

[137] Zhang, J., Zhang, Z., Xiao, X., Yang, Y., and Winslett, M. (2012). Functional mechanism: regression analysis under differential privacy. *arXiv preprint arXiv:1208.0219*.

[138] Zhang, J., Zheng, K., Mou, W., and Wang, L. (2017b). Efficient private erm for smooth objectives. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 3922–3928. AAAI Press.

[139] Zhao, J., Chen, Y., and Zhang, W. (2019a). Differential privacy preservation in deep learning: Challenges, opportunities and solutions. *IEEE Access*, 7:48901–48911.

[140] Zhao, J., Chen, Y., and Zhang, W. (2019b). Differential privacy preservation in deep learning: Challenges, opportunities and solutions. *IEEE Access*, 7:48901–48911.

[141] Zhao, Y. and Chen, J. (2022). A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s):1–28.

[142] Zhu, T., Li, G., Zhou, W., and Philip, S. Y. (2017). Differentially private data publishing and analysis: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1619–1638.

[143] Zhu, T., Ye, D., Wang, W., Zhou, W., and Philip, S. Y. (2020). More than privacy: Applying differential privacy in key areas of artificial intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2824–2843.

Summary

This dissertation addresses the critical need for robust evaluation methodology and framework in differentially private synthetic data generation (DP SDG). Synthetic data, which preserves privacy while enabling data analysis, is evaluated for its utility and quality through a novel framework proposed in this research. The dissertation begins by contextualizing the importance of synthetic data and differential privacy, identifying the gaps in current evaluation methods. The core contribution is the development of a comprehensive evaluation framework, initially introduced and later refined. This framework integrates macro-statistics, machine learning performance metrics, and probability distribution distance measures to provide a holistic assessment of synthetic data. Extensive empirical studies are conducted using various differentially private generative models, including copulas, Directed Acyclic Graphs (DAGs), and hybrid methods. These studies reveal the strengths and weaknesses of each model, offering valuable insights into their practical applications. Furthermore, the dissertation extends the evaluation framework to enhance its robustness and applicability. The new metrics introduced provide a more detailed understanding of the trade-offs between data utility and privacy. The findings and methodologies proposed in this research pave the way for more standardized and effective assessments of synthetic data, promoting its use in privacy-preserving data analysis. In conclusion, this work significantly advances the field of synthetic data evaluation, offering a comprehensive tool for researchers and practitioners to balance privacy and utility in data generation. Future research directions include refining evaluation metrics, applyisng the framework to diverse data domains.

DOI:
10.15476/ELTE.2024.310