Evaluating Privacy Techniques for Secure Data Publishing

By Laurine Owino

Supervisor: Prof. Pierre-Martin Tardif

Outline

- Introduction
- Background
- Methodology
- Results & Discussion
- Conclusion

Introduction

- Organizations (governments, businesses, researchers) collect and process large amounts of personal data.
- Data may need to be **shared or published** for:
  - Research and analysis
  - Public transparency
  - Collaboration between entities
- Pose privacy risks
  - Legal liabilities (GDPR, CCPA)
  - Fraud
  - Identity theft
  - Reputational damage

Privacy Threats [2]

- Identity Disclosure
- Attribute Disclosure
- Membership Disclosure

Objectives

To **explore and implement** various Privacy Enhancing Techniques (PETs) used in Privacy-Preserving Data Publishing (PPDP).

To **evaluate and compare the effectiveness** of these techniques in protecting individual privacy.

To **analyze the privacy-utility trade-off** associated with each of the techniques in practical data publishing scenarios.

Background

Privacy Enhancing Techniques (PETs)

- Anonymization
- Pseudonymization
- Differential Privacy
- Synthetic Data Generation
- Homomorphic Encryption

Data Attributes[1][2]

- **Direct Identifiers (DI)** – can identify user -> name, email.
  - *Deleted*
- **Quasi-Identifiers (QI)** – auxiliary info that can reveal identity -> age, ZIP code
  - *Generalized, suppressed*
- **Sensitive Attributes (SA)** – info that the user wants to be hidden -> crime, disease
  - *Retained*
- **Non-Sensitive Attributes** – other attributes -> height
  - *Not collected, removed, published as is*

	DI	NSA	Quasi Identifiers (QIs)			SA
ID	Name	Height	Age	Zip Code	Martial Status	Crime
1	Joe	5	29	32042	Separated	Murder
2	Jill	4	20	32021	Single	Theft
3	Sue	6	24	32024	Widowed	Traffic
4	Abe	5	28	32046	Separated	Assault
5	Bob	7	25	32045	Widowed	Piracy
6	Amy	6	23	32027	Single	Indecency

PPDP Techniques Used in this Study

- **Anonymization**
  - Still widely used
- **Synthetic Data Generation**
  - Especially good for ML tasks
- **Differential Privacy**
  - Good Privacy, de-facto method

Anonymization

Anonymization [2]

Generalization	• <25 or 25– 30
Suppression	• 2*
Permutation	• Records partitioned into groups and shuffled within those groups
Perturbation	• Replace values with synthetically generated
Anatomization	• Separate QIs and SAs

Anonymization Models

Model	**What is it?**	**Protects Against**	**Measures**	**Weaknesses**
k-anonymity	Ensures each record is indistinguishable from at least k-1 others based on quasi-identifiers.	Identity disclosure	Frequency of quasi-identifier values	Vulnerable to attribute disclosure
ℓ-diversity	Extends k-anonymity by ensuring that sensitive attributes have at least ℓ well-represented distinct values in each group.	Attribute disclosure	Count/entropy of sensitive attribute values	Cannot handle semantic similarity or skewed distributions
t-closeness	Further extends ℓ-diversity by ensuring the distribution of sensitive attributes in each group is close to the overall dataset distribution.	Stronger attribute disclosure	Distance between distributions	Harder to implement, may reduce data utility

Differential Privacy

Differential Privacy [4]

- Ensures that the output of a computation does not reveal too much about any single individual's data, even if attackers have access to auxiliary information.
- It introduces random noise to a query - prevents identification of a single user.
- Formally defined as:
  - $P(M(x) \in R) \leq P(M(x') \in R) . e\ \varepsilon$
- 2 main types:
  - Local
  - Statistical

Differential Privacy (Continued)

- Main concepts:
  - Privacy budget ε
  - Noise addition
  - Indistinguishability
- Used in:
  - Google Chrome RAPPOR
  - US Census 2020
  - Apple usage statistics
- Utility/accuracy vs privacy tradeoff

Synthetic Data Generation

Synthetic Data Generation [5]

- Used to **replicate the statistical properties of real data** without exposing any actual personal information.
- Techniques:
  - GAN-based
  - ML-based
  - Statistical based
  - Kernel based
- Metrics:
  - Privacy – no consensus
  - Performance - statistics, utility, distance & similarity
- Utility considered over resemblance
- DP is integrated in SDGs

Methodology

Methodology Overview

Dataset: UCI Adult Income	• Widely used for privacy and ML tasks • Moderate size & feature diversity
Technique Implementation	• Python libraries
Utility evaluation	• Performance on classification task
Privacy Evaluation	• Privacy parameter

Methodology Summary

Technique	**Libraries/Tools Used**	**Parameter Tested**	**Privacy Measurement**	**Utility Measurement**
Anonymization	AnonyPy, Anjana	k = 2, 3, 5, 10, 50, 100	Privacy Parameter: **k (k-Anonymity)**	**Suppression Level** (percentage of records suppressed)
Synthetic Data Generation	SDV	GaussianCopula, TVAE, CTGAN	**Distance Metrics** (statistical similarity to real data – Euclidean distance)	**Classification Accuracy** (Logistic Regression: trained on synthetic, tested on real)
Differential Privacy	IBM Diffprivlib	ε = 0.1, 0.5, 1.0, 5.0, 10.0	Privacy Parameter: **ε (epsilon)**	**Classification Accuracy** (Logistic Regression with DP applied)

Results & Discussion

Anonymization

- As the level of privacy increases(k), more records and more QIs are suppressed
- Suppression & generalization reduce usability of data.
- Anjana preserved utility more compared to the AnonyPy library.

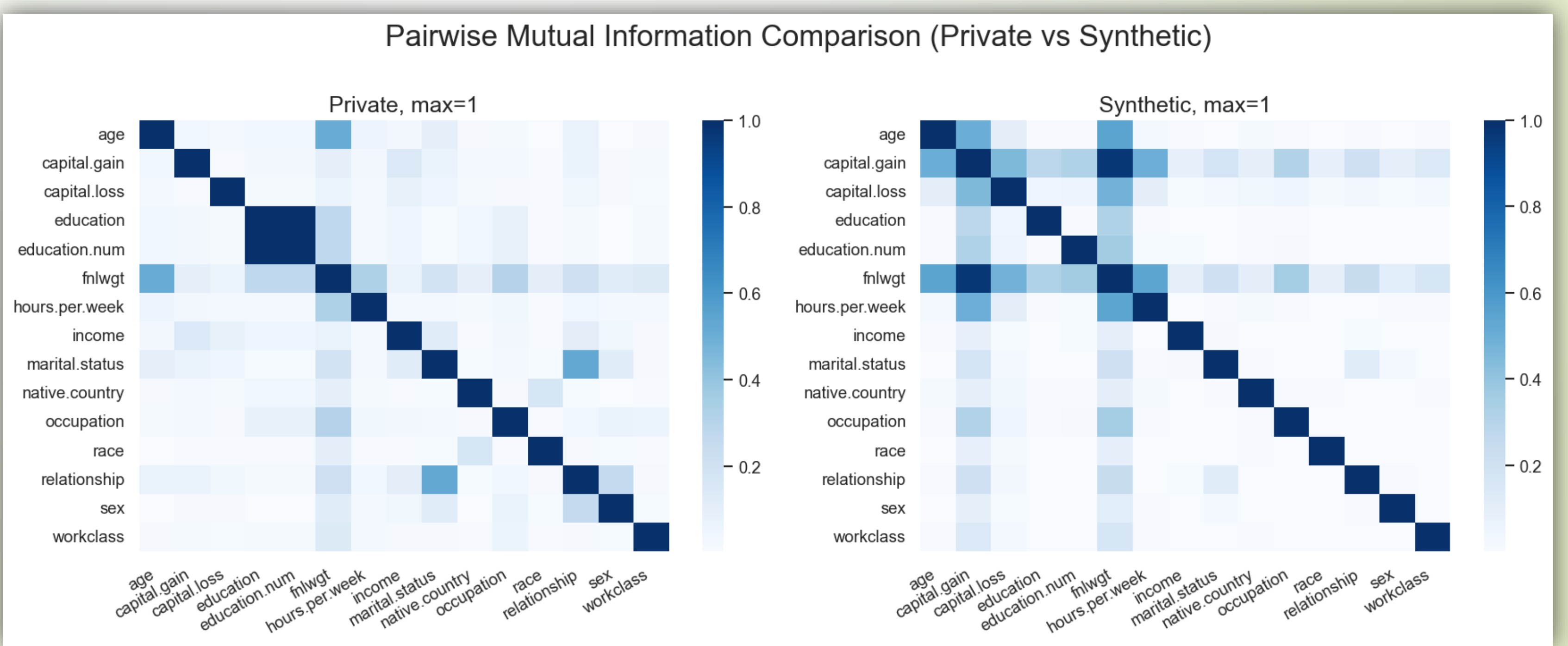
k	Suppressed Records (%)	Suppressed QIs
2	30.75	0
3	42.53	0
4	27.31	0
5	31.26	0
10	43.71	1
50	37.78	3
100	38.42	4

SDG

- GaussianCopula used due to its speed and modifiability despite it's lower quality.
- Using correlation map, synthetic data is seen to mostly follow the trends in the original data.

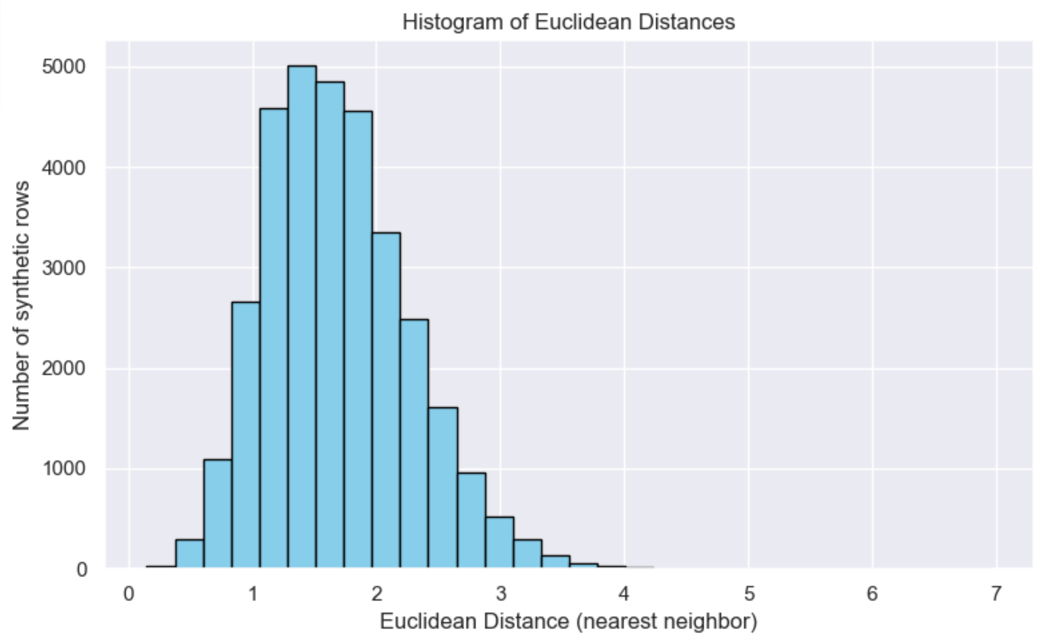
Synthesizer	Quality	Time Taken (s)
GaussianCopula	0.8442	4.3
CTGAN	0.8553	744.3
TVAE	0.8995	156.6

Synthesizers Comparison



SDG (Continued)

- Privacy
  - Euclidean distance ranges from 0-4
  - Datasets not similar - private
- Utility
  - Performance on classification task
  - % drop in all metrics



Metric	Original	Synthetic
Accuracy	0.86	0.79
Precision	0.85	0.80
Recall	0.86	0.79
F1-score	0.85	0.73

Differential Privacy

- Accuracy of classification model increases with increase in ε
- Clear privacy vs utility tradeoff
- Further increase beyond 5.0 seems to yield no better improvement to the performance of the classification task.

Epsilon (ε)	Classification Accuracy
0.1	0.7652
0.2	0.8145
0.3	0.8121
0.5	0.8239
1.0	0.8272
5.0	0.8270
10.0	0.8265

Conclusion

Key Findings

- **Anonymization** provided good privacy control via the **k-Anonymity parameter**, but utility was reduced as suppression increased.
- **Synthetic Data Generation (SDV)** resulted in data that preserved structure but suffered a **7% drop in classification accuracy**, showing a trade-off between privacy and utility.
- **Differential Privacy (Diffprivlib)** offered the **best trade-off**, maintaining acceptable utility with formal privacy guarantees.
- The **privacy-utility trade-off** is inevitable — stronger privacy generally reduces data usability.

Future Work

MORE DATASETS

UTILITY MEASURES

PRIVACY MEASURES & MODELLING

References

1. Carvalho, T., Moniz, N., Faria, P., & Antunes, L. (2022). Survey on privacy-preserving techniques for data publishing. *arXiv preprint arXiv:2201.08120*. (https://arxiv.org/abs/2201.08120)
2. A. Majeed and S. Lee. (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, vol. 9, pp. 8512-8545, 2021. (https://ieeexplore.ieee.org/document/9298747)
3. Latanya Sweeney. 2002. K-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 5 (October 2002), 557–570. (https://doi.org/10.1142/S0218488502001648)
4. Ying Zhao, Jia Tina Du, and Jinjun Chen. 2024. Scenario-based Adaptations of Differential Privacy: A Technical Survey. ACM Comput. Surv. 56, 8, Article 199. (https://doi.org/10.1145/3651153)
5. P. A. Osorio-Marulanda, G. Epelde, M. Hernandez, I. Isasa, N. M. Reyes and A. B. Iraola. 2024. Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review in *IEEE Access*, vol. 12, pp. 88048-88074. (https://ieeexplore.ieee.org/document/10568134)
6. Wagner, I., & Boiten, E. (2018). Privacy risk assessment: from art to science, by metrics. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2018 International Workshops, DPM 2018 and CBT 2018, Barcelona, Spain, September 6-7, 2018, Proceedings 13* (pp. 225-241). Springer International Publishing. (https://arxiv.org/abs/1709.03776)