

Josep Domingo-Ferrer  
Melek Önen (Eds.)

LNCS 14915

# Privacy in Statistical Databases

International Conference, PSD 2024  
Antibes Juan-les-Pins, France, September 25–27, 2024  
Proceedings



Springer

## Founding Editors

Gerhard Goos

Juris Hartmanis

## Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*

Wen Gao, *Peking University, Beijing, China*

Bernhard Steffen , *TU Dortmund University, Dortmund, Germany*

Moti Yung , *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Josep Domingo-Ferrer · Melek Önen  
Editors

# Privacy in Statistical Databases

International Conference, PSD 2024  
Antibes Juan-les-Pins, France, September 25–27, 2024  
Proceedings

*Editors*

Josep Domingo-Ferrer   
Rovira i Virgili University  
Tarragona, Catalonia, Spain

Melek Önen   
EURECOM  
Biot, France

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-031-69650-3

ISBN 978-3-031-69651-0 (eBook)

<https://doi.org/10.1007/978-3-031-69651-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

## Preface

Privacy in statistical databases is a discipline whose purpose is to provide solutions to the tension between social, political, economic and corporate demands for accurate information, and the legal and ethical obligation to protect the privacy of the various parties involved. In particular, the need to enforce privacy regulations in our data- and AI-driven world has made this tension all the more pressing. Stakeholders include the subjects, sometimes a.k.a. respondents (the individuals and enterprises to which the data contents refer), the data controllers (those organizations collecting, curating and to some extent sharing or releasing the data) and the users (the ones querying the database or the search engine, who would like their queries to stay confidential). Beyond law and ethics, there are also practical reasons for data controllers to invest in subject privacy: if individual subjects feel their privacy is guaranteed, they are likely to provide more accurate content. Data controller privacy is primarily motivated by practical considerations: if an enterprise collects data at its own expense and responsibility, it may wish to minimize leakage of those data contents to other enterprises (even to those with whom joint data exploitation is planned). Finally, user privacy results in increased user satisfaction, even if it may curtail the ability of the data controller to profile users.

There are at least two traditions in statistical database privacy, both of which started in the 1970s: the first one stems from official statistics, where the discipline is also known as statistical disclosure control (SDC) or statistical disclosure limitation (SDL), and the second one originates from computer science and database technology, and lately also from AI. In official statistics, the basic concern is subject privacy. In computer science, the initial motivation was also subject privacy but, from 2000 onwards, growing attention has been devoted to controller/owner privacy (privacy-preserving data mining) and user privacy (private information retrieval). In the last few years, the interest and the achievements of computer scientists in the topic have substantially increased, as reflected in the contents of this volume. At the same time, the widespread use of machine learning and AI is challenging privacy technologies in many ways: this volume also contains recent research aimed at tackling some of these challenges. In particular, synthetic data generation is gaining attention to reconcile data-hungry machine learning with data protection.

“Privacy in Statistical Databases 2024” (PSD 2024) was held in Antibes-Juan-Les-Pins, France, organized by Universitat Rovira i Virgili and EURECOM. The PSD series started twenty years ago, in 2004, and PSD 2024 is the eleventh conference in the series. Previous PSDs were held in various locations, mostly around the Mediterranean, and had their proceedings published in Springer LNCS: PSD 2022, Paris, LNCS 13463; PSD 2020, Tarragona, LNCS 12276; PSD 2018, Valencia, LNCS 11126; PSD 2016, Dubrovnik, LNCS 9867; PSD 2014, Eivissa, LNCS 8744; PSD 2012, Palermo, LNCS 7556; PSD 2010, Corfu, LNCS 6344; PSD 2008, Istanbul, LNCS 5262; PSD 2006, Rome, LNCS 4302; and PSD 2004, Barcelona, LNCS 3050. The PSD series took over from high-quality technical conferences on SDC which started twenty-six years ago

with “Statistical Data Protection-SDP” (Lisbon, 1998, OPOCE proceedings) and the AMRADS project SDC Workshop (Luxemburg, 2001, LNCS 2316).

The PSD 2024 Program Committee accepted for publication in this volume 28 papers out of 46 submissions. Nine more of the above submissions were accepted for short oral presentation at the conference. Papers came from 21 different countries on four different continents. Each submitted paper received at least two single-blind reviews. The revised versions of the 28 accepted papers in this volume are a fine blend of contributions from official statistics and computer science. Covered topics include privacy models, microdata protection, statistical table protection, synthetic data generation methods, synthetic data generation software, disclosure risk assessment, spatial and georeferenced data, machine learning and privacy, and case studies.

We are indebted to many people. First, to the Organization Committee for making the conference possible and especially to Jesús A. Manjón, who helped prepare these proceedings. In evaluating the papers we were assisted by the Program Committee and by Zhiyu Wan, Michel Reiffert, Martin Moehler and Chao Yan as external reviewers. We also wish to thank all the authors of submitted papers and we apologize for possible omissions.

July 2024

Josep Domingo-Ferrer  
Melek Önen

# Organization

## Program Chair

Josep Domingo-Ferrer

Universitat Rovira i Virgili, Catalonia, Spain

## General Chair

Melek Önen

EURECOM, France

## Program Committee

Jane Bambauer	University of Florida, USA
Bettina Berendt	Technical University of Berlin, Germany
Aleksandra Bujnowska	EUROSTAT, Luxembourg
Jordi Castro	Polytechnical University of Catalonia, Spain
Anne-Sophie Charest	Université Laval, Canada
Peter Christen	Australian National University, Australia
Graham Cormode	University of Warwick, UK
Josep Domingo-Ferrer	Universitat Rovira i Virgili, Spain
Jörg Drechsler	IAB, Germany
Mark Elliot	University of Manchester, UK
Sébastien Gambs	Université du Québec à Montréal, Canada
Sarah Giessing	Destatis, Germany
Hiroaki Kikuchi	Meiji University, Japan
Maryline Laurent	Télécom SudParis, France
Bradley Malin	Vanderbilt University, USA
Robin Mitra	University College London, UK
Nuno Moniz	University of Notre Dame, USA
Anna Monreale	Università di Pisa, Italy
Krishnamurty Muralidhar	University of Oklahoma, USA
Benjamin Nguyen	INSA Centre Val de Loire, France
Anna Oganyan	National Center for Health Statistics, USA
Melek Önen	EURECOM, France
Javier Parra-Arnau	Polytechnical University of Catalonia, Spain
Gillian Raab	University of Edinburgh, UK
Jerome Reiter	Duke University, USA

Yosef Rinott	Hebrew University, Israel
Felix Ritchie	University of the West of England, UK
Steven Ruggles	University of Minnesota, USA
Nicolas Ruiz	OECD; Universitat Rovira i Virgili, Spain
David Sánchez	Universitat Rovira i Virgili, Spain
Eric Schulte-Nordholt	Statistics Netherlands, The Netherlands
Natalie Shlomo	University of Manchester, UK
Aleksandra Slavković	Penn State University, UK
Tamir Tassa	Open University, Israel
Vicenç Torra	Umeå University, Sweden
Peter-Paul de Wolf	Statistics Netherlands, The Netherlands

## **Additional Reviewers**

Martin Moehler  
Michel Reiffert  
Zhiyu Wan  
Chao Yan

## **Organization Committee**

Joaquín García-Alfaro      Télécom SudParis, France  
Jesús A. Manjón      Universitat Rovira i Virgili, Spain

# Contents

## Privacy Models and Concepts

From Isolation to Identification .....	3
<i>Giuseppe D'Acquisto, Aloni Cohen, Maurizio Naldi, and Kobbi Nissim</i>	

Differentially Private Quantile Regression .....	18
<i>Tran Tran, Matthew Reimherr, and Aleksandra Slavkovic</i>	

Utility Analysis of Differentially Private Anonymized Data Based on Random Sampling .....	35
<i>Takumi Sugiyama, Hiroto Oosugi, Io Yamanaka, and Kazuhiro Minami</i>	

## Microdata Protection

Asymptotic Utility of Spectral Anonymization .....	51
<i>Katariina Perkonoja and Joni Virta</i>	

Robin Hood: A De-identification Method to Preserve Minority Representation for Disparities Research .....	67
<i>James Thomas Brown, Ellen W. Clayton, Michael Matheny,     Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A. Malin</i>	

## Statistical Table Protection

Secondary Cell Suppression by Gaussian Elimination: An Algorithm Suitable for Handling Issues with Zeros and Singletons .....	87
<i>Øyvind Langsrød</i>	

Obtaining $(\epsilon, \delta)$ -Differential Privacy Guarantees When Using a Poisson Mechanism to Synthesize Contingency Tables .....	102
<i>James Jackson, Robin Mitra, Brian Francis, and Iain Dove</i>	

## Synthetic Data Generation Methods

Generating Synthetic Data is Complicated: Know Your Data and Know Your Generator .....	115
<i>Jonathan Latner, Marcel Neunhoeffer, and Jörg Drechsler</i>	

Evaluating the Pseudo Likelihood Approach for Synthesizing Surveys Under Informative Sampling .....	129
<i>Anna Oganian, Jörg Drechsler, and Mehtab Iqbal</i>	
The Production of Bespoke Synthetic Teaching Datasets Without Access to the Original Data .....	144
<i>Mark Elliot, Claire Little, and Richard Allmendinger</i>	
<b>Synthetic Data Generation Software</b>	
A Comparison of SynDiffix Multi-table Versus Single-table Synthetic Data .....	161
<i>Paul Francis</i>	
An Evaluation of Synthetic Data Generators Implemented in the Python Library <i>Synthcity</i> .....	178
<i>Emma Fössing and Jörg Drechsler</i>	
Evaluation of Synthetic Data Generators on Complex Tabular Data .....	194
<i>Oscar Thees, Jiří Novák, and Matthias Templ</i>	
<b>Disclosure Risk Assessment</b>	
An Examination of the Alleged Privacy Threats of Confidence-Ranked Reconstruction of Census Microdata .....	213
<i>David Sánchez, Najeeb Jebreel, Krishnamurty Muralidhar,     Josep Domingo-Ferrer, and Alberto Blanco-Justicia</i>	
Synthetic Data: Comparing Utility and Risk in Microdata and Tables .....	225
<i>Simon Xi Ning Kolb, Jui Andreas Tang, and Sarah Giessing</i>	
Synthetic Data Outliers: Navigating Identity Disclosure .....	240
<i>Carolina Trindade, Luís Antunes, Tânia Carvalho, and Nuno Moniz</i>	
Privacy Risk from Synthetic Data: Practical Proposals .....	254
<i>Gillian M. Raab</i>	
Attribute Disclosure Risk in Smart Meter Data .....	274
<i>Guillermo Navarro-Arribas and Vicenç Torra</i>	
The statbarn: A New Model for Output Statistical Disclosure Control .....	284
<i>Elizabeth Green, Felix Ritche, and Paul White</i>	

**Spatial and Georeferenced Data**

- Masking Georeferenced Health Data - An Analysis Taking the Example  
of Partially Synthetic Data on Sleep Disorder ..... 297  
*Simon Cremer, Lydia Jehmlich, and Rainer Lenz*

- Privacy and Disclosure Risks in Spatial Dynamic Microsimulations ..... 310  
*Hanna Brenzel, Martin Palm, Jan Weymeirsch, and Ralf Münnich*

**Machine Learning and Privacy**

- Combinations of AI Models and XAI Metrics Vulnerable to Record  
Reconstruction Risk ..... 329  
*Ryotaro Toma and Hiroaki Kikuchi*

- DISCOLEAF: Personalized DIScretization of COntinuous Attributes  
for LEArning with Federated Decision Trees ..... 344  
*Saloni Kwatra and Vicenç Torra*

- Node Injection Link Stealing Attack ..... 358  
*Oualid Zari, Javier Parra-Arnau, Ayşe Ünsal, and Melek Önen*

- Assessing the Potentials of LLMs and GANs as State-of-the-Art Tabular  
Synthetic Data Generation Methods ..... 374  
*Marko Miletic and Murat Sariyar*

**Case Studies**

- Escalation of Commitment: A Case Study of the United States Census  
Bureau Efforts to Implement Differential Privacy for the 2020 Decennial  
Census ..... 393  
*Krishnamurty Muralidhar and Steven Ruggles*

- Relational Or Single: A Comparative Analysis of Data Synthesis  
Approaches for Privacy and Utility on a Use Case from Statistical Office ..... 403  
*Manel Slokom, Shruti Agrawal, Nynke C. Krol, and Peter-Paul de Wolf*

- A Case Study Exploring Data Synthesis Strategies on Tabular vs.  
Aggregated Data Sources for Official Statistics ..... 420  
*Mohamed Aghaddar, Liu Nuo Su, Manel Slokom, Lucas Barnhoorn,  
and Peter-Paul de Wolf*

- Author Index** ..... 437

# **Privacy Models and Concepts**



# From Isolation to Identification

Giuseppe D'Acquisto<sup>1</sup>, Aloni Cohen<sup>2(✉)</sup>, Maurizio Naldi<sup>3</sup>, and Kobbi Nissim<sup>4</sup>

<sup>1</sup> LUISS University, Rome, Italy

[gdacquisto@luiss.it](mailto:gdacquisto@luiss.it)

<sup>2</sup> University of Chicago, Chicago, USA

[aloni@g.uchicago.edu](mailto:aloni@g.uchicago.edu)

<sup>3</sup> LUMSA University, Rome, Italy

[m.naldi@lumsa.it](mailto:m.naldi@lumsa.it)

<sup>4</sup> Georgetown University, Washington, DC, USA

[kobbi.nissim@georgetown.edu](mailto:kobbi.nissim@georgetown.edu)

**Abstract.** We present a mathematical framework for understanding when successfully distinguishing a person from all other persons in a data set—a phenomenon which we call *isolation*—may enable *identification*, a notion which is central to deciding whether a release based on the data set is subject to data protection regulation. We show that a baseline degree of isolation is unavoidable in the sense that isolation can typically happen with high probability even before a release was made about the data set and hence identification is not enabled. We then describe settings where isolation resulting from a data release may enable identification.

**Keywords:** privacy · identification · isolation · data protection

## 1 Introduction

The notion of *identification* is central to privacy and data protection regulation. For example, the GDPR regulates the processing of personal data: “any information relating to an identified or identifiable natural person.”<sup>1</sup> But the GDPR leaves “identifiable” undefined. What constitutes identification?

It is an old idea that identification may be possible when attributes together uniquely distinguish a individual within a population. In 1986, Dalenius wrote that “it is well known” that “the data for some variables may, for some individuals, be unique and publicly known” and expose those individuals to record linkage attacks [6]. The U.S. National Institute of Standards and Technology even defines identification as “the process of using claimed or observed attributes of an entity to single out the entity among other entities in a set of identities.”<sup>2</sup> Sweeney carried out this process in her re-identification of MA Governor Weld in a dataset of state employee health records. Moreover, she showed that very few attributes are needed to uniquely distinguish most US residents: 5-digit ZIP, gender, and date of birth suffice for 87% of respondents in the 1990 Census [15].

<sup>1</sup> Regulation (EU) 2016/679 (General Data Protection Regulation), Article 4.

<sup>2</sup> For variations and sources, see <https://csrc.nist.gov/glossary/term/identification>.

But there is an important difference between distinguishing a person within a dataset or a sample—which this paper calls *isolation*—and distinguishing a person within a population [3,8]. Consider the study by De Montjoye et al. showing that four random time-location pairs were enough to isolate 95% of individuals’ records in a dataset on 1.5M people. Sánchez et al. reply: “With a nonexhaustive sample, an individual’s sample uniqueness/unicity does not imply population uniqueness and, hence, does not allow unequivocal reidentification” [14, citing [2]]. With this specific claim, we agree – absent other information, it’s not clear to what extent these isolations amount to identification. Of course, if one can also check whether a given person was in the sample, then isolation in the sample plus the fact that person was in the sample results in isolation in the population.

## 1.1 This Work’s Contributions

*Identification and Isolation.* While our goal is to provide a better understanding of identification to help design approaches for releasing information while protecting individual privacy, we do not attempt to define what identification means in a mathematically formal way. We see identification as an inherently-fuzzy legal concept for which a satisfactory mathematical treatment may not even exist. What we do is take a closer look at the related phenomenon of isolation.

In a little more detail, we consider a setting where an information holder produces a data release  $\mathbf{R}$  based on a table  $\mathbf{X}$  of PII. An adversarial information receiver tries to use the release  $\mathbf{R}$  to *identify* one or more of the data items in  $\mathbf{X}$ . We say that the information receiver *isolates* in  $\mathbf{X}$  if they succeed in producing a description<sup>3</sup> that matches exactly one person in  $\mathbf{X}$ . That isolation and identification are related follows from observing that isolation has been a major stepping stone towards identification in linkage attacks, where, typically, an entry of a presumably deidentified dataset is first isolated and then re-identified via linkage with an dataset containing identifying information (see, e.g., [15]).

We provide novel additions to the discussion of the relationship between identification and isolation:

1. We argue that some baseline degree of isolation is unavoidable and does not enable identification. In Sects. 3.1 and 3.2 we show that an information receiver having knowledge of the probability distribution underlying the data in  $\mathbf{X}$  can isolate individuals in it prior to seeing any data release. (In Appendix A we extend the results to the case of an information receiver having only partial knowledge of the distribution.)
2. In Sect. 4 we ask when a data release leads to identification. We identify *isolation gain*—a measure of how an information receiver’s confidence in an isolation attempt grows once they receive a release  $\mathbf{R}$  based on the dataset. In Sect. 4.2 we revisit examples from the re-identification literature, analyzing them through this lens.

---

<sup>3</sup> For example:  $(25 \leq \text{Age} \leq 28) \wedge (10,000 \leq \text{Salary} \leq 50,000)$ .

*Postulates About Identification.* As a surrogate for defining identification mathematically, our analysis proceeds from two postulates about identification which we believe are simple, intuitive, and uncontroversial.

**Postulate 1.** *If a release  $\mathbf{R}$  contains no information derived from the dataset  $\mathbf{X}$ , then  $\mathbf{R}$  itself cannot be used to identify any individual in  $\mathbf{X}$ .*

In particular, the baseline degree of isolation (item 1 above) is unavoidable does not constitute “singling out” as used in the GDPR.<sup>4</sup>

**Postulate 2.** *A description of an individual record in  $\mathbf{X}$  may enable identification if it is specific enough to uniquely distinguish the corresponding individual in the underlying population. A release  $\mathbf{R}$  from which such a description is derived may enable identification.*

We stress that these postulates *do not* characterize identification. The postulates describe extreme cases leaving a bulk of real-world data analyses somewhere in gray area in between. Even so, these postulates are useful as they allow us to describe the outer bounds of identification from data release, or a lack thereof.

## 1.2 Related Work

As discussed above and in Sect. 4.2 below, a long line of work studies re-identification from anonymized or de-identified data releases [2–4, 6, 7, 11, 12, 14, 15].<sup>5</sup> A recent work of particular relevance is that of Rocher, Hendricks, and De Montjoye [12], who address the gap between sample- and population-uniqueness by showing that it is often possible to empirically estimate the probability that an isolating set of attributes uniquely distinguishes an individual within the underlying population, (an estimation task with a long history [3, 8]).

Postulate 1 is implicit in Ruggles and Van Riper’s critique of the US Census Bureau’s reconstruction of the 2010 Decennial Census: empirically comparing those results to a baseline that “would be expected by chance” [13]. Jarmin et al. analyze disclosure risk assessment frameworks in part using as a sanity check that they should “deem releasing uninformative statistics not a disclosure risk” [10].

We use Postulate 1 to derive a baseline level of isolation that does not amount to disclosure of any sort. A line of work does a version of this by instead excluding an individual from the data analysis [9, and citations therein]. Most recently, Francis and Wagner put forward a “non-member framework”<sup>6</sup>, empirically apply it to prior attacks, and discuss its relevance to the concepts of identifiability and anonymization under the GDPR to past re-identification studies [9].

---

<sup>4</sup> General Data Protection Regulation, Recital 26. See also: Article 29 Working Party, *Opinion 05/2014 on Anonymisation Techniques*.

<sup>5</sup> See citations in [12] for more.

<sup>6</sup> “If an individual is not present in a dataset, and is independent of all other individuals in the dataset, then the release of that dataset does not violate that individual’s privacy.”

Our approach builds on prior work by Cohen and Nissim [5] and Altman et al. [1]. These works introduce an abstract framework for isolation and analyze its relation with the GDPR notion of singling out, based on what we call Postulate 1. The current paper connects isolation with identification by introducing Postulate 2, extends the prior work by considering multiple isolation attempts, and presents its findings in more concrete settings. We also consider heuristic isolation strategies for the setting where the information receiver does not have sufficient knowledge of the probability measure underlying the data.

## 2 The Isolation Problem

Consider a scenario where some personal data is stored by an information holder, and an information receiver wishes to uncover that information (or part of it). The information holder owns a table (the “ground truth” table) with a row for each individual,

$$\mathbf{X} = (\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n).$$

As an example, each row  $\underline{x}_i$  may take value in some space  $D \subset R^d$ . We assume that we can build a probability space on  $D$  by defining the space of elementary events and a probability measure  $P$ , so that we can assign a probability value to all events (i.e., subsets of  $D$ ) forming a Borel field. The table  $\mathbf{X}$  is assumed to contain the result of  $n$  i.i.d. random samples draw from  $P$ .

We define the following basic statistic, computed over  $\mathbf{X}$ :

**Definition 1 (counting function).** *For a generic subdomain  $B \subseteq D$ , the number of instances in  $\mathbf{X}$  falling in it is given by the counting function:*

$$H_{\mathbf{X}}(B) = \sum_{i=1}^n I(\underline{x}_i \in B).$$

The information holder releases aggregate data about  $\mathbf{X}$ . For concreteness, we assume that a data release  $\mathbf{R}$  consists of a collection of pairs  $\mathbf{R} = \{(A_i, m_i)\}$  where each of the pairs in  $\mathbf{R}$  declares a subdomain  $A_i \subseteq D$  and a number  $m_i$  where  $m_i = H_{\mathbf{X}}(A_i) + \epsilon_i$ . In the case of absence of noise,  $\epsilon_i = 0$  for all  $i$  and therefore  $m_i$  is guaranteed to be the exact number of individuals in  $\mathbf{X}$  that fall into the subdomain  $A_i$ . In the case of a noisy release, the noise variable  $\epsilon_i$  is assumed to be drawn from a known distribution.

Initially, the release is empty (i.e., initially  $\mathbf{R}_0 = \emptyset$ ) and the only information known to the information receiver is the data subspace  $D$ , the underlying probability measure  $P$ , and the number of entries  $n$  in  $\mathbf{X}$ .<sup>7</sup> The information receiver's initial knowledge  $(D, P, n, \mathbf{R}_0)$  about  $\mathbf{X}$  is updated with the aggregate counts  $\mathbf{R} = \{(A_i, m_i)\}$  once they are released. We assume the receiver also knows whether the release is noiseless or noisy, along with the noise distribution in the latter case.

---

<sup>7</sup> Our results are robust to a substantial relaxation of these assumptions, in particular, knowledge of the distribution  $P$  and the number of records  $n$  may be approximate. See Remark 3 and Appendix A.

## 2.1 Isolation

We define *isolation* to happen when the information receiver outputs a description matching exactly one row in  $\mathbf{X}$ , formally:

**Definition 2 (isolation).** We say that  $B \subseteq D$  isolates in  $\mathbf{X}$  if  $H_{\mathbf{X}}(B) = 1$ .

*Guessing an isolating  $B$ .* The information receiver may use their knowledge about  $\mathbf{X}$ —namely, the data domain  $D$ , probability measure  $P$ , number of entries  $n$ , and a release  $\mathbf{R}$ —to guess one or more sets that isolate in  $\mathbf{X}$ . That is, the receiver comes up with subsets  $B_1, B_2, \dots, B_k \subseteq D$  that the information receiver hopes isolate many of the entries in  $\mathbf{X}$ .

*Verifying that a Guess  $B$  Isolates.* If the information receiver has query access to a noiseless release mechanism then they can ask for the release  $(B, H_{\mathbf{X}}(B))$ , and hence check whether the guess  $B$  indeed consists an isolation.<sup>8,9</sup>

*Remark 1.* The information receiver may try to isolate in  $\mathbf{X}$  even before receiving any release (i.e., only given  $D$ ,  $P$ ,  $n$ , and  $\mathbf{R}_0 = \emptyset$ ). With a release  $\mathbf{R} \neq \emptyset$  made about  $\mathbf{X}$ , the receiver can improve its confidence in guessing an isolating  $B$ . The information receiver's rate of successful isolation given  $\mathbf{R}_0$  can be thought of as a baseline to which their isolation ability given a release  $\mathbf{R} \neq \emptyset$  can be compared.

## 3 Optimal Isolation Without Any Release

We now consider strategies an information receiver may use to isolate one or more individuals in the dataset  $\mathbf{X}$ . We analyze the information receiver's isolation ability prior to any release  $\mathbf{R}$ , i.e., with  $\mathbf{R}_0 = \emptyset$ . The analysis is done under the assumption that the information receiver has knowledge of the data domain  $D$ , the underlying probability measure  $P$ , and the number of elements  $n$  in  $\mathbf{X}$ .<sup>10</sup>

*Remark 2.* The same analysis carried out in this section holds for the case of a non-empty release  $\mathbf{R}$  by replacing the probability measure  $P$  with the probability measure resulting from conditioning  $P$  on the release  $\mathbf{R}$ .

Subsections 3.1 and 3.2 discuss the guessing strategies the information receiver may use assuming they have perfect knowledge of the underlying probability measure  $P$ . In Appendix A we discuss how these strategies can be used by an information receiver that does not have complete knowledge of  $P$ .

---

<sup>8</sup> Query access to the release mechanism can also be used in other ways, see Remark 4 below.

<sup>9</sup> Access to alternative sources of information may also be used for boosting the information receiver's confidence that a guess  $B$  isolates.

<sup>10</sup> The number of records  $n$  is included as part of the the receiver's prior knowledge since  $n$  can often be inferred (exactly or approximately) from public information. Examples include (i) where a survey design specifying  $n$  was made public prior to the collection of information, and (ii) where  $n$  was made public in previous surveys (e.g., a census). For a reader who considers  $n$  as part of the release, this section should be understood as demonstrating that the release of  $n$  alone suffices for isolation.

### 3.1 A Single Isolation Guess

We begin with the analysis of how the information receiver may make one guess  $B$  so as to maximize their probability of a successful isolation.

For  $B \subseteq A$ ,  $P(B)$  is the probability that an individual sampled according to  $P$  falls into the subdomain  $B$ . The probability that  $B$  contains exactly one individual from the  $n$  individuals in  $D$  (i.e., that some individual from the  $n$  individuals in  $D$  has been isolated) is given by the expression

$$p^{\text{iso}}(B) := \Pr_{\mathbf{X} \sim P^n}[H_{\mathbf{X}}(B) = 1] = n \cdot P(B) \cdot (1 - P(B))^{n-1}. \quad (1)$$

Observe that  $p^{\text{iso}}(B)$  depends only on  $P(B)$ . The information receiver can choose the subset  $B$  so that  $P(B)$  maximizes  $p^{\text{iso}}(B)$ . All that remains is to determine the value  $P(B)$  that achieves the maximum.

**Theorem 1.** *The probability of isolation  $p^{\text{iso}}(B)$  achieves its maximum value of  $(1 - \frac{1}{n})^{n-1} \approx \frac{1}{e} \approx 0.37$  when  $p(B) = \frac{1}{n}$ .*

*Proof.* To simplify notation, let  $p^{\text{iso}} = p^{\text{iso}}(B)$  and  $p = P(B)$ . We compute the first and second derivatives of Eq. (1):

$$\frac{\partial p^{\text{iso}}}{\partial p} = n(1-p)^{n-2}(1-np), \quad \text{and} \quad \frac{\partial^2 p^{\text{iso}}}{\partial p^2} = -n(n-1)(1-p)^{n-3}(2-np).$$

The first derivative is positive on  $p \in [0, \frac{1}{n}]$ , zero at  $p = \frac{1}{n}$ , and negative on  $p \in (\frac{1}{n}, 1]$ . The second derivative is negative for  $p < \frac{2}{n}$ . Hence,  $p = \frac{1}{n}$  maximizes  $p^{\text{iso}}$ , with maximum value  $n \cdot \frac{1}{n} \cdot (1 - \frac{1}{n})^{n-1} = (1 - \frac{1}{n})^{n-1}$ .

Theorem 1 may be interpreted as follows: if the information receiver can make a single guess  $B$ , the receiver maximizes the probability that  $B$  isolates in  $\mathbf{X}$  by choosing  $B$  such that  $p(B) = \frac{1}{n}$ , in which case the receiver's guess is successful with probability about 0.37.

*Remark 3.* If instead of picking  $B$  such that  $p(B) = \frac{1}{n}$  the information receiver picks  $B$  such that  $p(B) = \frac{c}{n}$  then the isolation probability drops to

$$n \cdot \frac{c}{n} \cdot \left(1 - \frac{c}{n}\right)^{n-1} = \frac{c}{1 - \frac{c}{n}} \left(1 - \frac{c}{n}\right)^n \approx \frac{c}{e^c}.$$

- This implies that the result of Theorem 1 is only mildly sensitive to errors in the information receiver's knowledge of  $P$  and  $n$ . For example, if the receiver's knowledge is off by a factor of at most 2 (i.e.,  $\frac{1}{2n} \leq P(B) \leq \frac{2}{n}$  or, equivalently,  $\frac{1}{2} \leq c \leq 2$ ), then  $p^{\text{iso}}(B) \gtrsim 0.27$ .
- If, however, the information receiver chooses to make a guess with  $p(B) = \frac{c}{n}$  where  $c$  is very small, then we get  $e^c \approx 1$  and the isolation probability is  $p^{\text{iso}}(B) \approx c$ , i.e., also very small.

### 3.2 Multiple Isolation Guesses

We now consider an information receiver that makes  $k > 1$  guesses  $B_1, \dots, B_k \subseteq D$  trying to isolate multiple individuals in  $\mathbf{X}$ . We will require that the guesses  $B_i$  are mutually disjoint, so it is impossible for two of them to isolate the same element in  $\mathbf{X}$ . The receiver's goal is to maximize the total number isolated elements in  $\mathbf{X}$  (i.e., number of guesses in  $B_1, \dots, B_k$  which successfully isolate in  $\mathbf{X}$ ), i.e.,  $\sum_{i=1}^k I(H_{\mathbf{X}}(B_i) = 1)$ . As discussed in Sect. 2.1, if after choosing  $B_1, \dots, B_k$  the information receiver can issue them as queries then the receiver would learn with certainty which of them isolates.

*Remark 4.* Our analysis considers an *oblivious* information receiver which makes all the guesses  $B_1, \dots, B_k$  at once. Indeed, this is the information receiver's only choice if they cannot issue queries and obtain more releases. However, if the information receiver may issue queries, then an *adaptive* strategy would be more effective for multiple isolations. In such a strategy the receiver would choose guess  $B_{i+1}$  after seeing  $H_{\mathbf{X}}(B_i)$ . As an example, an adaptive adversary may choose to use a divide-and-conquer strategy to isolate elements in  $\mathbf{X}$ .

*Remark 5.* We ignore other, less direct, modes of isolation. For example, if  $B_i$  is a strict subset of  $B_j$  and  $H_{\mathbf{X}}(B_i) = H_{\mathbf{X}}(B_j) - 1$  then their difference  $B_j \setminus B_i$  isolates.

**Observation 1.** If the number of guesses  $k \leq n$  then we can use Theorem 1: the information receiver may choose  $k$  disjoint subsets where  $P(B_i) = \frac{1}{n}$  for all  $i = 1, \dots, k$ . In expectation about  $\frac{k}{e}$  of the guesses  $B_i$  would isolate.

The information receiver's optimal strategy in the case  $k > n$  is characterized by the following theorem:

**Theorem 2.** When  $k \geq n$ , the expected number of isolations achieves its maximum when  $P(B_i) = \frac{1}{k}$  for all  $i \in 1, \dots, k$ .

*Proof.* To simplify notation, let  $p_i = P(B_i)$ . For  $0 \leq p \leq 1$ , define  $f(p) = n \cdot p \cdot (1-p)^{n-1}$ . For  $\mathbf{p} = (p_1, \dots, p_k)$ , define  $F(\mathbf{p}) = \sum_{i=1}^k f(p_i)$ . By Eq. (1), we can rewrite the expected number of isolations as  $F(\mathbf{p})$ :

$$\mathbf{E} \left( \sum_{i=1}^k I(H_{\mathbf{X}}(B_i) = 1) \right) = \sum_{i=1}^k \mathbf{E}(p_i^{\text{iso}}(B_i)) = \sum_{i=1}^k f(p_i) = F(\mathbf{p}).$$

Fix  $\mathbf{p} = (p_1, \dots, p_k)$  arbitrarily and let  $\mathbf{p}^* = (\frac{1}{k}, \dots, \frac{1}{k})$ . We must show that  $F(\mathbf{p}) \leq F(\mathbf{p}^*)$ . To do so, we will give two intermediate variables  $\mathbf{p}'$  and  $\mathbf{p}''$  and show that  $F(\mathbf{p}) \leq F(\mathbf{p}') \leq F(\mathbf{p}'') \leq F(\mathbf{p}^*)$ . We use two facts about  $f$  already shown in the proof of Theorem 1. First,  $f$  is (strictly) concave in the interval  $0 \leq p \leq \frac{1}{n}$ . Second,  $f$  is (strictly) increasing as  $p \rightarrow \frac{1}{n}$  from either side.

Construct  $\mathbf{p}'$  by clamping each  $p_i$  to the interval  $[0, 1/n]$ . Namely,  $p'_i = \min(p_i, 1/n)$  for each  $i = 1, \dots, k$ . If  $p'_i = p_i$ , then  $f(p_i) = f(p'_i)$ . Otherwise,  $1/n \leq p'_i < p_i$  and hence  $f(p_i) < f(p'_i)$ . Therefore  $F(\mathbf{p}) \leq F(\mathbf{p}')$ .

Observe that  $0 \leq \sum p'_i \leq 1$ . Construct  $\mathbf{p}''$  arbitrarily such  $\sum p''_i = 1$  and  $p'_i \leq p''_i \leq 1/n$  for all  $i$ . This is possible because  $\sum_{i=1}^k \frac{1}{n} = \frac{k}{n} \geq 1$ . For all  $i$ ,  $f(p''_i) \geq f(p'_i)$ . Therefore  $F(\mathbf{p}') \leq F(\mathbf{p}'')$ .

By construction,  $0 \leq p''_i \leq 1/n$  for all  $i$ . Because  $f$  is concave on  $[0, 1/n]$ ,

$$F(\mathbf{p}) \leq \sum_{i=1}^k f(p''_i) \leq k \cdot f\left(\frac{p''_1 + \dots + p''_k}{k}\right) = F(\mathbf{p}^*).$$

*Putting it all Together.* By the combination of Observation 1 and Theorem 2, setting  $p(B_i) = \frac{1}{\max(k, n)}$  for all  $B_i$  maximizes the expected number of isolations. For  $k \leq n$ , the expected number of isolations is

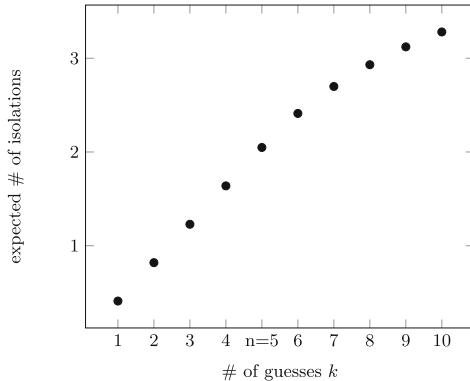
$$k \left(1 - \frac{1}{n}\right)^{n-1} \approx \frac{k}{e}.$$

For  $k > n$ , the expected number of isolations is

$$k \cdot n \cdot p \cdot (1-p)^{n-1} = n \cdot \left(1 - \frac{1}{k}\right)^{n-1} = n \cdot \left(1 - \frac{1}{k}\right)^{k \cdot \frac{n-1}{k}} \approx n \cdot e^{-\frac{n}{k}},$$

an expression that tends to  $n$  as  $k$  grows to infinity. More concretely, in the regime  $k \geq n$ , to achieve  $\alpha n$  isolations in expectation, it suffices for the information to make  $k \approx \frac{n}{\ln(1/\alpha)}$  guesses. In particular, approximately  $20n$  guesses suffice for  $\alpha = 0.95$  and approximately  $100n$  guesses suffice for  $\alpha = 0.99$ .

Figure 1 provides an example of how the expected number of isolations grows as a function of guesses  $k$  if  $n = 5$  and  $P(B_i)$  is taken based on Observation 1 and Theorem 2.



**Fig. 1.** Expected number of isolations as a function of  $k$ . ( $n = 5$ ).

## 4 From Isolation to Identification

*Isolation Alone is Insufficient for Identification.* Theorem 1 shows that even prior to any release (i.e., with  $\mathbf{R}_0 = \emptyset$ ) an information receiver making just a single isolation attempt would be successful in more than one in three trials. Likewise, Observation 1 and Theorem 2 show that by making many guesses  $B_1, \dots, B_k$ , the information receiver can drive the expected number of isolations up to  $n$  as  $k$  grows. By Postulate 1, none of these cases amount to identification. Some other criterion beyond mere isolation is needed.

*A Baseline.* We again look to Theorems 1 and 2. They characterize the optimum rate of isolation achievable without any data-specific knowledge, and what strategies achieve that optimum. Those results therefore describe a *baseline* against which an information receiver’s attempts at isolation may be measured. If the receiver can significantly outperform the baseline, it is indicative of some non-trivial disclosure.

We are now guided by Postulate 2. A guess  $B$  that isolates in  $\mathbf{X}$  describes an individual record in the dataset (e.g., SEX=male, ZIP=02138, DOB=07/31/1945). Such a description may enable identification if it is specific enough to uniquely distinguish that the corresponding individual in the underlying population.

*Isolation may enable identification when  $B$  is specific enough to uniquely distinguish an individual in a population, and also isolates in  $\mathbf{X}$  better than the baseline chance.* Two parameters are important:  $P(B)$ , the probability mass of  $B$  in the prior distribution; and  $\Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}]$ , the probability that  $B$  isolates in  $\mathbf{X}$  conditioned on the the release  $\mathbf{R}$ . Smaller  $P(B)$  and larger  $\Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}]$  are stronger evidence that identification may be possible.

Consider, e.g., a simple setting where a population  $\mathbf{X}^*$  of  $N$  individuals is drawn i.i.d. from  $P$ , and a subset  $\mathbf{X}$  of size  $n = rN$  is sampled uniformly at random, where  $0 < r \ll 1$ . The information receiver sees a release  $\mathbf{R}$  derived from  $\mathbf{X}$  and outputs a single guess  $B$ . The receiver’s goal is that  $B$  should both isolate in  $\mathbf{X}$  and uniquely distinguish an individual in the population  $\mathbf{X}^*$ . If  $B$  isolates in  $\mathbf{X}$ , then it uniquely distinguishes in  $\mathbf{X}^*$  if no element of  $\mathbf{X}^* \setminus \mathbf{X}$  is in  $B$ . Taking  $P(B) \leq \frac{1}{100N} = \frac{r}{100n}$ , we get that  $B$  uniquely distinguishes in  $\mathbf{X}^*$  with probability

$$(1 - P(B))^{N-n} = (1 - P(B))^{(1-r)N} \geq \left(1 - \frac{1}{100N}\right)^{100N \cdot \frac{1-r}{100}} \approx e^{-\frac{(1-r)}{100}} > 0.99.$$

On the other hand, by Remark 3, the baseline chance that the information receiver produces such a  $B$  that also isolates in  $\mathbf{X}$  after only receiving the empty release  $\mathbf{R}_0$  is:

$$p^{\text{iso}}(B) = \Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}_0] \lesssim \frac{r}{100} \ll 0.01.$$

In this example, the release  $\mathbf{R}$  may enable identification if—with probability much greater than 0.01—the information receiver produces a guess  $B$  such that both  $P(B) \leq \frac{1}{100N}$  and  $B$  isolates in  $\mathbf{X}$ .

We quantify the improvement in the probability of isolation conditioned on the release  $\mathbf{R}$  using a quantity we call *isolation gain*.

**Definition 3 (isolation gain).** Let  $B \subseteq D$ . Consider two datasets  $\mathbf{X}, \mathbf{X}' \sim P^n$  and let  $\mathbf{R}$  be a release derived from  $\mathbf{X}$ . The isolation gain for  $B$  is defined as:

$$G(B) = \frac{\Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}]}{\Pr[H_{\mathbf{X}'}(B) = 1]}.$$

Rephrasing the preceding discussion, the information receiver wants to make a guess  $B$  with a high isolation gain. Namely, the receiver attempts to maximize the numerator—the probability that  $B$  isolates given  $\mathbf{R}$ —while minimizing the denominator, which for  $P(B) < \frac{1}{n}$  is equivalent to minimizing  $P(B)$ .<sup>11</sup>

#### 4.1 How an Information Release may Enable Identification

We now consider how the information receiver might use a release  $\mathbf{R}$  about a dataset  $\mathbf{X}$  of size  $n$  in an attempt to identify an individual in a population of  $N \gg n$ . Throughout this section, we view a (noiseless) data release  $\mathbf{R}$  as consisting of a collection of pairs  $\mathbf{R} = \{(A_i, m_i)\}$ , each pair specifying a subdomain  $A_i \subseteq D$  and the count  $m_i = H_{\mathbf{X}}(A_i)$ . In this subsection, we assume that the  $A_i$  are chosen independently of  $\mathbf{X}$  (the worst case for the information receiver). We consider an information receiver who produces a single guess  $B$  seeking to minimize  $P(B)$  while maximizing the probability of isolation  $\Pr[H_{\mathbf{X}}(B) = 1 \mid \mathbf{R}]$ .

We use a simple observation: if  $B_i$  isolates a single record among the  $m_i$  records in  $A_i$ , then  $B = B_i \cap A_i$  isolates a single record in  $\mathbf{X}$ . To find such a  $B_i$ , we adapt the strategies analyzed in Sect. 3. In that section, we analyzed the optimum probability of isolation assuming only that  $\mathbf{X} \sim P^n$ , for any  $P$  and  $n$ . For any  $i$ , let  $\mathbf{X}_i = \{\underline{x} \in \mathbf{X} : \underline{x} \in A_i\}$  be those elements of  $\mathbf{X}$  contained in  $A_i$ , and  $P_i$  be the data distribution conditioned on  $\underline{x} \in A_i$ . The posterior distribution of  $\mathbf{X}_i$  conditioned on the release  $\mathbf{R}$  is  $m_i$  i.i.d. samples from  $P_i$ . (This uses the assumption that  $A_i$  is independent of  $\mathbf{X}$ .) Theorem 1 implies that  $B_i$  satisfying  $P_i(B_i) = P(B_i | A_i) = \frac{1}{m_i}$  will maximize the probability of isolating in  $\mathbf{X}_i$ , as desired. The isolation gain for  $B = B_i \cap A_i$  is

$$G(B) = \frac{m_i \cdot \frac{1}{m_i} \cdot (1 - \frac{1}{m_i})^{m_i-1}}{n \cdot P(B) \cdot (1 - P(B))^{n-1}}. \quad (2)$$

Whether the above strategy is good depends on which sets  $A_i$  are in the release  $\mathbf{R}$ . We analyze three examples based on the value of  $P(A_i)$  relative to  $\frac{1}{n}$ .

If  $P(A_i) \ll \frac{1}{n}$ , the attack is very successful:  $G(B) \gg 1$ . Furthermore, if  $P(A_i) \ll \frac{1}{N}$ , then the release may enable identification. To see why, observe that many  $A_i$  will contain exactly 1 record:  $m_i = 1$ . Take  $B = A_i$  for any such

---

<sup>11</sup> We exclude  $P(B) \geq \frac{1}{n}$ , as the chance that  $B$  uniquely distinguishes an individual in a population of size  $N = n/r$  is extremely small. Namely,  $(1 - P(B))^{N-n} \leq e^{-(1-r)/r}$ . Taking  $r = 0.01$ , say, the probability is about  $10^{-43}$ .

$A_i$ . The numerator of  $G$  is  $\Pr[H_{\mathbf{X}}(B) = 1] = \Pr[m_i = 1] = 1$ . By Remark 3, the denominator of  $G$  is  $\Pr[H_{\mathbf{X}'}(B) = 1] \approx n \cdot P(B) \ll 1$ . If  $P(B) = P(A_i) \ll \frac{1}{N}$ , then  $B$  uniquely distinguishes the isolated individual in the population with high probability.

If  $P(A_i) = \frac{1}{n}$ , the attack slightly beats the baseline:  $G(B) \gtrsim \frac{\ln(n)}{e}$ . By a standard balls-in-bins analysis, there exists  $i$  such that  $m_i \approx \ln(n)$  with high probability. For this  $i$ , we take  $B = B_i \cap A_i$ , yielding  $P(B) \approx \frac{1}{n \ln(n)}$ . The numerator of  $G$  is  $\approx \frac{1}{e}$ . The denominator is  $< n \cdot P(B) \approx \frac{1}{\ln(n)}$ .

If  $P(A_i) \gg \frac{1}{n}$ , the attack doesn't beat the baseline:  $G \approx 1$  in expectation. In this case,  $m_i \gg 1$  with high probability, and thus the numerator of Eq. (2) is  $\approx \frac{1}{e}$ . Next, observe that

$$\mathbf{E}[P(B)] = \mathbf{E}\left[\frac{P(A_i)}{m_i}\right] = \mathbf{E}\left[\frac{\mathbf{E}(H_{\mathbf{X}}(A_i))}{n \cdot H_{\mathbf{X}}(A_i)}\right] = 1.$$

For  $P(B) \approx \mathbf{E}[P(B)]$ ,<sup>12</sup> the denominator of Eq. (2) is also  $\approx \frac{1}{e}$ .

## 4.2 Examples from the Re-identification Literature

We briefly consider few types of releases  $\mathbf{R}$  which capture re-identification attacks from prior work.

*Microdata Releases.* In a typical microdata release  $\mathbf{R}$ , the subdomain  $A_i$  describes the attributes of record  $\underline{x}_i \in \mathbf{X}$ , possibly with some attributes generalized or redacted. Typically, the attributes are very rich, hence  $P(A_i) \ll \frac{1}{n}$ . As described in the previous section, the information receiver can achieve a high isolation gain by taking  $B = A_i$ . Whether  $B$  uniquely distinguishes the isolated individual in the population depends on details of the release. But, since it does not take many attributes to uniquely distinguish somebody in the population [15], it is likely that, for a rich data domain,  $P(B) < \frac{1}{100N}$ .

A well-known example is the “Unique in the Crowd” study by De Montjoye et al. [7]. The  $A_i$  consisted of many time-location data points for each of 1.5M people. The study showed that taking  $B \supseteq A_i$  to contain just four of these points sufficed to isolate for 95% of the rows  $i$ . But no evidence was given that  $P(B)$  was small enough to uniquely distinguish an individual in the underlying population [14].

To get beyond mere isolation, Rocher et al. directly estimated the probability of population uniqueness for microdata releases [12]. Like us, they observed that the probability that  $B$  uniquely distinguishes the isolated individual in the population is at least  $(1 - P(B))^{N-1}$ . They showed that for real-world datasets, an information receiver can empirically estimate  $(1 - P(B))^{N-1}$  to within a few percent by using sample of the population to learn the distribution  $P$ .<sup>13</sup> Using

<sup>12</sup> Accounting for the variance of  $P(B)$  yields only an insignificant improvement.

<sup>13</sup> E.g., for Governor Weld's attributes used by Sweeney, they estimated  $(1 - P(B))^{N-1} \approx 0.58$ .

this estimate, an information receiver can choose  $B$  such that  $(1 - P(B))^{N-1} \geq 0.95$ , say, which implies that  $P(B) \ll \frac{1}{N}$ .

Other re-identification studies on microdata releases include Sweeney's re-identification of Governor Weld [15] (see below) and Narayanan and Shmatikov's re-identification using the Netflix Prize Dataset [11].

*k-anonymity.* A  $k$ -anonymous data release contains  $\ell$ -many counts  $(A_i, m_i)$  subject to the constraint that  $m_i \geq k$ . The parameter  $k > 1$  is a small constant (e.g.,  $k = 5, 10$ ). For simplicity, let us assume that  $m_i = k$  for all  $i$ .

Cohen and Nissim analyze the success of the following information receiver for *arbitrary k-anonymization* algorithms [5]. Guess  $B \subseteq A_i$  arbitrary subject to  $P(B|A_i) = \frac{1}{k}$ , for any  $i$ . They show that so long as the data distribution has a moderate amount of entropy,  $B$  isolates with probability about  $(1 - \frac{1}{k})^{k-1} > \frac{1}{e}$ , regardless of the  $k$ -anonymization algorithm.<sup>14</sup> It remains to analyze  $P(B) = \frac{P(A_i)}{k}$ . As for the microdata release, how small  $P(A_i)$  depends on the  $k$ -anonymization algorithm and data distribution. Most  $k$ -anonymization algorithms are designed to preserve as much richness of the input dataset  $\mathbf{X}$  as possible, i.e., minimizing  $P(A_i)$ . For rich-enough data, it is possible to provide  $k$ -anonymity while also guaranteeing that  $P(A_i) < \frac{1}{100N}$  with high probability.

Cohen gives a much more effective strategy called *downcoding*, but which requires some assumptions on the  $k$ -anonymization algorithm and data distribution [4]. The core observation is that, if the  $k$ -anonymization algorithm preserves as much of  $\mathbf{X}$  as possible, the sets  $A_i$  must depend on the data. Cohen shows that for some data distributions, the  $A_i$  enable the information receiver to recover a very detailed description  $B_j$  of some fraction of the rows  $x_j \in \mathbf{X}$  ( $\geq 3\%$  of the rows for  $k \leq 15$ ). These  $B_j$  isolate in  $\mathbf{X}$ , and  $P(B_j) < \frac{1}{100N}$  as long as  $\mathbf{X}$  contains at least  $3 \ln(100N)$  attributes.

*When Membership in  $\mathbf{X}$  is Known.* Often,  $\mathbf{X}$  is not a random sample of the population. Rather, membership in  $\mathbf{X}$  is correlated with some attribute of the data. This extra information can help the information receiver turn isolation into identification by excluding from  $B$  individuals not in  $\mathbf{X}$ . For example, Cohen's re-identification of EdX students required only a few attributes about the students. Cohen was able to exclude all individuals not in the EdX release using the certificates of completion posted by many EdX students on their LinkedIn profiles, thereby turning isolation into identification [4]. As another example, Sweeney's re-identification of Governor Weld made use of the fact that the dataset contained the hospital records for all state employees [15].

*Overlapping Contingency Tables.* An example that doesn't fit neatly into the above comes from Israel's Central Bureau of Statistics.<sup>15</sup> Very roughly, the release included a count  $m$  for subdomains  $A$  specified by any choice of up to 5 attributes. For example, there was exactly 1 male widower veteran with

<sup>14</sup> The proof of this fact is somewhat nuanced, as  $A_i$  can depend arbitrarily on the dataset  $\mathbf{X}$ .

<sup>15</sup> See <https://www.slideserve.com/ordell/razi-mukatren-golan-salman>, and <https://archive.is/W20kx>.

no children among the survey respondents. Alone, these subdomains had probability  $P(A) \approx \frac{1}{n}$ . By the analysis in the previous section, identification would seem impossible. But if only four attributes were needed to isolate a record—as for the widower above—it is easy to reconstruct the record entirely. For every additional attribute, exactly one possible value will be non-zero. In this way, the information receiver can bootstrap many isolations into possible identifications.

**Acknowledgments.** Work of K.N. was supported by NSF Grant No. CCF2217678 “DASS: Co-design of law and computer science for privacy in sociotechnical software systems” and a gift to Georgetown University. Work completed while K.N. visited Bocconi University, Milan.

## A Isolating with a Partial Knowledge of $P$

The analysis in Sects. 3.1 and 3.2 assumed that the information receiver has perfect knowledge of the underlying probability measure  $P$  (but not  $\mathbf{X}$  sampled from  $P$ ). We now discuss what the receiver may do when they do not know  $P$  in full.

Observation 1 and Theorem 2 teach that all the information receiver needs is a partition of the data space into sets of probability weight  $p^* = \frac{1}{\max(n,k)}$  and Remark 3 suggests that it suffices that the partition is close to the optimal weight for the information receiver to succeed in isolating. We now develop these ideas.

Let  $\mathcal{C} = \{C_i\}_{i=1}^\ell$  be a partition of  $D$  where  $\ell = \max(n, k)$ . The information receiver may choose the partition  $\mathcal{C}$  heuristically in combination with their partial knowledge about  $P$  and the data domain  $D$ . For example,  $\mathcal{C}$  may partition  $D$  into high-dimensional rectangles, each described as the conjunction of one or more attribute ranges (e.g., all combinations of 5-year Age by Sex by City). Denote by  $p_i = P(C_i)$  the probability of an individual falling into  $C_i$ . We show that if  $\underline{p} = (p_1, \dots, p_\ell)$  is close enough to  $p^* = (\frac{1}{\ell}, \dots, \frac{1}{\ell})$  then (even without knowing  $p_1, \dots, p_\ell$ ) the information receiver succeeds in isolating.

As  $\mathcal{C}$  is a partition of the data domain  $D$  we have that  $\sum_{i=1}^\ell p_i = 1$ . Hence, if we pick a partition element at random, then, in expectation, its probability weight would be exactly  $p^* = \frac{1}{\ell}$ :

$$\mathbf{E}_{i \sim U_\ell} [p_i] = \sum_{j=1}^\ell \Pr_{i \sim U_\ell} [i = j] \cdot p_j = \frac{1}{\ell} \sum_{j=1}^\ell p_j = \frac{1}{\ell},$$

where we use  $i \sim U_\ell$  to denote that the expectancy is over choosing an element of the partition  $i \in \{1, \dots, \ell\}$  uniformly at random.

An important parameter of the partition is its standard deviation  $\sigma$ :

$$\sigma^2 := \mathbf{Var}_{i \sim U_\ell} [p_i] = \mathbf{E}_{i \sim U_\ell} [p_i^2] - \left( \mathbf{E}_{i \sim U_\ell} [p_i] \right)^2 = \frac{1}{\ell} \sum_{i=1}^\ell p_i^2 - \frac{1}{\ell^2} = \frac{1}{\ell} \cdot \|\underline{p}\|_2^2 - \frac{1}{\ell^2}.$$

If the standard deviation  $\sigma$  is small compared to  $\frac{1}{\ell}$ , say  $\sigma \leq \frac{c}{\ell}$  for  $c \ll 1$ , then many of the elements of the partition have weight  $p_i \approx \frac{1}{\ell}$ . More precisely, by Chebyshev's inequality<sup>16</sup> we have:

$$\Pr \left[ p_i \notin \left[ \frac{1}{2\ell}, \frac{3}{2\ell} \right] \right] = \Pr \left[ |p(C_i) - \mathbf{E}[p(C_i)]| > \frac{1}{2\ell} \right] < 4c^2.$$

Hence, if the information receiver samples guesses  $B_1, \dots, B_k$  from the partition  $\mathcal{C}$  without replacement, then in expectation at least  $(1 - 4c^2)k$  of them would satisfy  $p(B_i) \in [\frac{1}{2\ell}, \frac{3}{2\ell}]$ . Using Eq. 1 each of these guesses would result in isolation probability  $p^{\text{iso}}(B_i) \geq n \cdot \min(\frac{1}{2\ell} \cdot (1 - \frac{1}{2\ell})^{n-1}, \frac{3}{2\ell} \cdot (1 - \frac{3}{2\ell})^{n-1})$ , and in expectation the number of isolating guesses would be at least

$$(1 - 4c^2) \cdot \frac{kn}{2\ell} \cdot \min \left( \left(1 - \frac{1}{2\ell}\right)^{n-1}, 3 \cdot \left(1 - \frac{3}{2\ell}\right)^{n-1} \right).$$

As an example, if  $c = \frac{1}{4}$  and  $k = n$  (hence  $\ell = k = n$ ) we get that in expectation the number successful isolations is at least

$$\begin{aligned} & \left(1 - 4 \cdot \left(\frac{1}{4}\right)^2\right) \cdot \frac{n^2}{2n} \cdot \min \left( \left(1 - \frac{1}{2n}\right)^{n-1}, 3 \cdot \left(1 - \frac{3}{2n}\right)^{n-1} \right) \\ & \approx \frac{3n}{8} \cdot \min \left( e^{-1/2}, 3e^{-3/2} \right) \approx 0.23n. \end{aligned}$$

I.e., in expectation almost a quarter of the guesses would consist successful isolations in spite of  $B_i$  not being chosen optimally.

*Remark 6.* Cohen and Nissim [5] used hashing to create a structure that is equivalent to a partition  $\mathcal{C}$  where  $p_i$  is very close to  $\frac{1}{\ell}$  (assuming  $P$  has sufficient min-entropy). The main qualitative difference between the hashing approach and the one described in this work is that hashing destroys the structure of the data domain and makes it harder for the information receiver to make effective use of isolation (e.g., as a step towards a linkage attack) whereas in the approach described herein the information receiver may choose partitions  $\mathcal{C}$  that are more suitable for their purposes.

## References

- Altman, M., Cohen, A., Nissim, K., Wood, A.: What a hybrid legal-technical analysis teaches us about privacy regulation: the case of singling out. *BUJ Sci. Tech. L.* **27**, 1 (2021)
- Barth-Jones, D.: The ‘re-identification’ of governor William Weld’s medical information: a critical re-examination of health data identification risks and privacy protections, then and now. *Then Now* (2012) (2012)
- Bethlehem, J.G., Keller, W.J., Pannekoek, J.: Disclosure control of microdata. *J. Am. Stat. Assoc.* **85**(409), 38–45 (1990)

---

<sup>16</sup> Chebyshev's inequality:  $\Pr [|X - \mathbf{E}[X]| \geq k] \leq \frac{\mathbf{Var}[X]}{k^2}$ .

4. Cohen, A.: Attacks on deidentification’s defenses. In: Butler, K.R.B., Thomas, K. (eds.) 31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, 10–12 August 2022, pp. 1469–1486. USENIX Association (2022). <https://www.usenix.org/conference/usenixsecurity22/presentation/cohen>
5. Cohen, A., Nissim, K.: Towards formalizing the GDPR’s notion of singling out. Proc. Natl. Acad. Sci. USA **117**(15), 8344–8352 (2020). <https://doi.org/10.1073/pnas.1914598117>
6. Dalenius, T.: Finding a needle in a haystack or identifying anonymous census records. J. Official Stat. **2**(3), 329 (1986)
7. De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. **3**(1), 1–5 (2013)
8. Fienberg, S.E., Makov, U.E.: Confidentiality, uniqueness, and disclosure limitation for categorical data. J. Official stat. **14**(4), 385 (1998)
9. Francis, P., Wagner, D.: Towards more accurate and useful data anonymity vulnerability measures. arXiv preprint [arXiv:2403.06595](https://arxiv.org/abs/2403.06595) (2024)
10. Jarmin, R.S., et al.: An in-depth examination of requirements for disclosure risk assessment. Proc. Natl. Acad. Sci. **120**(43), e2220558120 (2023)
11. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (SP 2008), 18–21 May 2008, Oakland, California, USA, pp. 111–125. IEEE Computer Society (2008). <https://doi.org/10.1109/SP.2008.33>
12. Rocher, L., Hendrickx, J.M., De Montjoye, Y.A.: Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun. **10**(1), 1–9 (2019)
13. Ruggles, S., Van Riper, D.: The role of chance in the census bureau database reconstruction experiment. Popul. Res. Policy Rev. **41**, 781–788 (2022). <https://doi.org/10.1007/s11113-021-09674-3>
14. Sánchez, D., Martínez, S., Domingo-Ferrer, J.: Comment on “unique in the shopping mall: on the reidentifiability of credit card metadata”. Science **351**(6279), 1274–1274 (2016)
15. Sweeney, L.: Simple demographics often identify people uniquely. Health (San Francisco) **671**(2000), 1–34 (2000)



# Differentially Private Quantile Regression

Tran Tran<sup>(✉)</sup>, Matthew Reimherr, and Aleksandra Slavkovic

Department of Statistics, The Pennsylvania State University, University Park,  
State College, PA, USA  
`{ttran,mreimherr,sesa}@psu.edu`

**Abstract.** Quantile regression (QR) is a powerful and robust statistical modeling method broadly used in many fields such as economics, ecology, and healthcare. However, it has not been well-explored in differential privacy (DP) since its loss function lacks strong convexity and twice differentiability, often required by many DP mechanisms. We implement the smoothed QR loss via convolution within the K-Norm Gradient mechanism (KNG) and prove the resulting estimate converges to the non-private one asymptotically. Additionally, our work is the first to extensively investigate the empirical performance of DP smoothing QR under pure-, approximate- and concentrated-DP and four mechanisms, and cases commonly encountered in practice such as heavy-tailed and heteroscedastic data. We find that the Objective Perturbation Mechanism and KNG are the top performers across the simulated settings.

**Keywords:** Differential privacy · Quantile regression · Smoothed loss

## 1 Introduction

Government regulations, such as the Federal Policy for the Protection of Human Subjects and Title 13 of the U.S. Code, require researchers and agencies to safeguard personally identifiable information and limit individuals' privacy loss. Over the past decade, the commonly used data anonymization and traditional statistical disclosure control (SDC) techniques that satisfy the regulations have been shown to not provide adequate privacy protection. For example, in 2018, the U.S. Census Bureau conducted an internal study examining a database reconstruction attack on the Census 2010, which had undergone rigorous SDC before data were released to the public. The Census claimed they could reconstruct the microdata for 144 million people, which is 46% of the U.S. population, and reidentify another 52 million by linking them with commercial data [22]. SDC methods can also fall short of formally quantifying the amount of privacy loss in the presence of publicly available data or when the same data are used to publish multiple statistics [3]. Furthermore, there is often limited practical guidance on what criteria should be met to maintain good privacy-utility trade-off when publishing aggregate statistics and data analysis results. This lack of guidance creates potential vulnerabilities, where individuals' sensitive information can be inferred from the aggregate statistics with the aid of public data sources [19, 46].

Differential privacy (DP), originally proposed by [18], has emerged as a leading framework aiming to address the shortcomings of some SDC methods and their susceptibility to privacy attacks; for a survey of those attacks, see [19]. There are many variants of the originally proposed  $\epsilon$ -DP; e.g., see [35]. The DP framework proposes infusing a well-designed random noise into the data or statistical outputs to obscure the data attributable to any specific individual. Methods that satisfy DP can guarantee provable protection against large groups of attacks, in some cases arbitrary. Importantly, data and statistics published with DP guarantees allow researchers to gain insight into the population while limiting the amount of information revealed about any individual [46]. As a result, DP has gained attention from governmental agencies (e.g., U.S. Census 2008 OnTheMap product [32], the 2020 U.S. Census [2]) and leading technology companies, such as Apple [4], Google [20], Microsoft [15], and LinkedIn [38]. Researchers from various fields have also shown interest in using DP data, but some have expressed concerns about the potential biases and usability of the data and statistical outputs due to the random noise introduced by DP [34, 39]. This has led to a rapidly growing literature on new DP definitions and algorithms to solve specific statistical problems with an eye on an improved privacy-utility trade-off. In this paper, we focus on DP quantile regression (DPQR).

Quantile regression (QR) is a powerful statistical tool for modeling the conditional distribution of the response variable given its predictors. As it does not make any distributional assumptions and is robust to outliers, QR has been applied in many areas, such as economics [47], health science [40], and machine learning [14]. However, there has been a limited consideration of QR within DP because its loss function lacks strong convexity and twice differentiability, requirements of many DP mechanisms. Most of the existing work in this area centers around leveraging the favorable statistical properties of the median to perform private estimation and inference [9, 13, 16, 17, 36]. There also exists another stream of work focusing solely on private quantile estimation (without predictors) [23, 28, 42], leaving a gap in the DPQR problem.

In the non-private literature, making inferences or implementing gradient-based optimization methods, such as gradient descent, with quantile regression is not straightforward, again due to its loss function lacking strong convexity and twice differentiability. To address this problem, multiple solutions have been proposed, e.g., [21, 24, 25, 27, 45, 48]. Among them, Fernandes et al. [21] introduced a convolution-typed smoothing technique to ensure the loss has these two properties. They proved this approach can incur negligible bias while inheriting many favorable asymptotic properties. More recently, Chen and Chua [12] adapted this strategy for DPQR. Specifically, they applied the smoothed loss to two  $(\epsilon, \delta)$ -DP mechanisms, the Objective Perturbation Mechanism (OPM) [29] and the DP Stochastic Gradient Descent (DP-SGD) [7], and proved several DP theoretical results, such as (near) optimal excess generalization risks and the parameter error upper bounds.

In this paper, building upon the work by Chen and Chua [12], we propose and implement the convolution-smoothed loss for DPQR in two prominent  $\epsilon$ -DP

and Concentrated-DP (CDP) mechanisms, respectively: the K-Norm Gradient mechanism (KNG) [37] and the DP Gradient Descent algorithm with adaptive per-iteration privacy-loss budget (DP-AGD) [31]. DP-AGD, with its unique capability to dynamically determine the optimal privacy budget and step size at each iteration, can significantly enhance the efficiency and accuracy of traditional DP gradient descent methods. KNG has been demonstrated to add an asymptotically negligible (compared to the estimation error) amount of noise to satisfy DP when the loss function is strongly convex and twice differentiable. Specifically, we derive an additional asymptotic property for KNG where we quantify the magnitude of the noise due to privacy as a function of the Hessian of the loss function. This becomes especially useful for DPQR where we utilize a smooth loss that asymptotically approximates the traditional QR loss function; we can then show that the DP estimate using the smooth loss converges to the original non-private QR estimate asymptotically. We expand the simulation study in [12], which only explored one quantile and data with normally distributed noise, by evaluating the mechanisms' performance at other commonly used quantiles (5th, 25th, 50th, 75th, 95th, and 99th) at varying privacy-loss budgets, sample sizes, and different data structures, such as heavy-tailed and heteroscedastic data. Our work is the first to compare DPQR across three different DP definitions (e.g.,  $\epsilon$ -DP,  $(\epsilon, \delta)$ -DP and CDP) and four mechanism (e.g., KNG, OPM, DP-SGD, and DP-AGD), and thus, can provide a valuable benchmark for researchers and practitioners when determining which mechanism to use in their studies. Our simulations reveal that OPM and KNG consistently outperform other methods in terms of accuracy and runtime across multiple quantiles and data settings.

The remainder of the paper is structured as follows. Section 2 briefly introduces quantile regression, including the smoothed quantile loss via convolution smoothing, and DP. Section 3 compiles the details of DP quantile regression under KNG, OPM, DP-SGD, and DP-AGD, and provides a theoretical utility result for KNG. In Sect. 4, we evaluate the performance of the four mechanisms in different simulated settings before concluding the paper.

## 2 Background

### 2.1 Quantile Regression

Quantile regression [30] focuses on estimating the conditional quantiles of the response variable given its predictors. Unlike ordinary least squares, which can only estimate the conditional mean, QR models the whole conditional distribution of the outcome of interest. Let  $y_i \in \mathbb{R}$ ,  $\theta_\tau \in \mathbb{R}^{p \times 1}$  and  $x_i \in \mathbb{R}^{p \times 1}$  ( $i = 1, \dots, n$ ), then quantile  $\tau \in (0, 1)$  of  $y_i$  given covariate  $x_i$  within linear QR model is of the form  $q_\tau(y_i|x_i) = x_i^\top \theta_\tau$ .

We assume that we observe data  $\{(y_i, x_i) : i = 1, 2, \dots, n\}$  and our goal is to estimate the true coefficients  $\theta_\tau$  through the minimizer of the loss function  $\hat{\theta}_\tau$ :

$$\hat{\theta}_\tau = \arg \min_{\theta} \sum_{i=1}^n \rho_\tau(y_i - x_i^\top \theta), \quad (1)$$

where  $\rho_\tau(u)$  is the check loss or check function, with the following form:

$$\rho_\tau(u) = (\tau - 1)u\mathbb{1}(u < 0) + \tau u\mathbb{1}(u \geq 0) = \frac{|u|}{2} + \left(\tau - \frac{1}{2}\right)u. \quad (2)$$

Quantile regression provides both flexibility in distributional assumptions and the robustness of estimators to outliers. However, its loss function not being twice differentiable and strongly convex makes inference procedures challenging. In particular, its asymptotic variance depends on the conditional density evaluated at specific  $x$ , and its convergence rates to asymptotic normality are slow [21]. The loss function also makes QR not computationally efficient because it cannot be implemented using fast gradient-based optimization algorithms. This issue is exacerbated in the large-scale setting, where both  $p$  and  $n$  are large [24]. To address this problem, multiple solutions have been proposed, e.g., [24, 25, 27, 44, 45, 48]. Among them, Horowitz [25] proposed smoothing the indicator of the check loss to make it differentiable, but this method cannot solve the convexity problem. Zheng [48] suggested approximating the check loss with a smooth alternative and successfully demonstrated its use in the setting of gradient-based optimization methods. Most recently, Fernandes et al. [21] generalized the smooth loss from [48] by applying the convolution smoothing method to the loss function to make it strongly convex and twice differentiable. The authors showed that their proposed method enjoys many desirable theoretical properties, such as a reduction in estimators' variance, asymptotically negligible bias, and asymptotic mean squared error that is lower than the original estimator proposed in [30]. He et al. [24] further extended this method and showed that it could yield favorable results even in high-dimension and large-scale settings. The formal definition of smoothed quantile loss via convolution smoothing is provided in Definition 1.

**Definition 1. Smoothed Quantile Regression Loss via Convolution [21]:** Let  $k$  be a smooth kernel that integrates to 1 ( $\int k(u)du = 1$ ), is symmetric and monotonic ( $k(u) = k(-u) \forall u \in \mathbb{R}$  and  $k(u) \leq k(v) \forall |u| \geq |v|$ ), and has finite absolute moments. The Smoothed Quantile Regression Loss is defined as

$$\rho_{\tau,h}(u) = \int_{-\infty}^{\infty} \rho_\tau(v)k_h(u-v)dv = \frac{h}{2} \int_{-\infty}^{\infty} \left| \frac{u}{h} + v \right| k(v)dv + \left( \tau - \frac{1}{2} \right)u, \quad (3)$$

where  $k_h(u) = \frac{1}{h}k\left(\frac{u}{h}\right)$ . Denote  $K(u) = \int_{-\infty}^u k(v)dv$  and  $K_h(u) = K\left(\frac{u}{h}\right)$ , then the first and second derivatives of  $\rho_{\tau,h}(u)$  have the form  $\rho'_{\tau,h}(u) = K_h(u) + \tau - 1$  and  $\rho''_{\tau,h}(u) = k_h(u)$ .

We denote  $\hat{\theta}_{\tau,h}$  as the minimizer of smoothed loss in Eq. 3. Additionally, in Table 1, we compiled some of the commonly used kernel functions satisfying the assumptions in Definition 1 that can be used in conjunction with Eq. 3 to obtain the corresponding smoothed loss.

**Table 1.** List of Commonly Used Kernels for smoothed quantile regression [12, 24, 44]

Kernels	$k(u)$	$\int_{-\infty}^{\infty} \frac{u}{h} + v  k(v)dv$
Gaussian	$(2\pi)^{-1/2} e^{-u^2/2}$	$(2/\pi)^{1/2} e^{-(u/h)^2/2} + (u/h)(1 - 2\Phi(-u/h))$
Logistic	$e^{-u}/(1 + e^{-u})^2$	$u/h + 2 \log(1 + e^{-u/h})$
Uniform	$(1/2)\mathbb{1}\{ u  \leq 1\}$	$((u/h)^2/2 + 1/2)\mathbb{1}\{ u/h  \leq 1\} +  u/h \mathbb{1}\{ u/h  > 1\}$
Laplacian	$e^{- u }/2$	$e^{- u/h } -  u/h $
Epanechnikov	$(3/4)(1 - u^2)\mathbb{1}\{ u  \leq 1\}$	$((3/4)(u/h)^2 - (u/h)^4/8 + 3/8)\mathbb{1}\{ u/h  \leq 1\} +  u/h \mathbb{1}\{ u/h  > 1\}$

## 2.2 Differential Privacy

Differential privacy (DP) [18] is one of the leading frameworks for providing mathematically provable privacy guarantees. The core objective of DP is to allow researchers to analyze confidential data to gain population-level insights without revealing any additional individual-specific information [46]. To achieve this objective, DP involves a randomized algorithm that yields similar outputs for datasets differing only by a single record. One of the key advantages of DP is that it allows us to go beyond the binary definition of privacy, private or not, by quantifying privacy loss via parameter  $\epsilon$ , also known as the *privacy-loss budget* [18]; see Definition 2. A lower privacy-loss budget  $\epsilon$  results in better privacy guarantees for individuals in the data, but often lowers the utility of analyses done by researchers.

**Definition 2.**  *$(\epsilon, \delta)$ -Differential Privacy* [18]: A randomized algorithm  $\mathcal{M}$  is  $(\epsilon, \delta)$ -differentially private if for all  $S \subseteq \text{Range}(\mathcal{M})$  and for all adjacent datasets  $D, D'$  differing by one row ( $d(D, D') = 1$ ):

$$\Pr(\mathcal{M}(D) \in S) \leq \exp(\epsilon) \Pr(\mathcal{M}(D') \in S) + \delta, \quad (4)$$

where  $\epsilon > 0$  and  $\delta \in [0, 1]$ .

When  $\delta = 0$ ,  $(\epsilon, \delta)$ -DP (approximate DP) becomes  $\epsilon$ -DP (pure DP), the strictest form of DP. Intuitively,  $\epsilon$  sets a bound on which the outputs of two datasets differing by one record can deviate from one another, while  $\delta$  accounts for the small probability where this bound can be violated, or in other words, there can be a privacy leakage. Aside from the approximate DP, there also exist other relaxations of pure DP, such as Concentrated-DP [10] and probabilistic DP [32]. While KNG provides  $\epsilon$ -DP guarantees, OPM can satisfy either  $\epsilon$ - or  $(\epsilon, \delta)$ -DP, DP-SGD satisfies  $(\epsilon, \delta)$ -DP, and DP-AGD satisfies CDP.

**Definition 3.**  *$\rho$ -Zero-Concentrated Differential Privacy (zCDP)* [10]: A randomized algorithm  $\mathcal{M}$  satisfies  $\rho$ -zCDP if for an output  $o \in \text{Range}(\mathcal{M})$  and  $\alpha \in (1, \infty)$ , its privacy loss random variable

$$Z = \log \left[ \frac{\Pr(\mathcal{M}(D) = o)}{\Pr(\mathcal{M}(D') = o)} \right] \quad (5)$$

satisfies  $\exp\{D_\alpha(\mathcal{M}(D)||\mathcal{M}(D'))\} = E[\exp\{(\alpha - 1)Z\}] \leq \exp\{(\alpha - 1)\alpha\rho\}$ , where  $D_\alpha(\cdot)$  denotes the  $\alpha$ -Rényi divergence.

**Lemma 1. Equivalence between  $\rho$ -zCDP and  $(\epsilon, \delta)$ -DP [10]:** If mechanism  $\mathcal{M}$  satisfies  $\rho$ -zCDP, then  $\mathcal{M}$  is  $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$  for any  $\delta > 0$ .

Numerous mechanisms that satisfy the above definitions, beyond the original Laplace [18] and the Exponential [33] ones, have been introduced in DP literature. The objective perturbation mechanism (OPM), initially proposed by [11] and further extended by [29], works by releasing the minimizer of the perturbed objective function. Lee and Kifer [31] extended previous works on DP gradient descent and introduced a new version with adaptive per-iteration budget (DP-AGD). More recently, Reimherr and Awan [37] proposed KNG, an  $\epsilon$ -DP mechanism that uses the gradient of the objective function to sanitize a statistic.

Another area within DP that has garnered significant scholarly attention is evaluating the utility of private outputs, e.g., [5, 26, 43]. Outputs from DP methods at the same privacy budget are often compared using certain metrics to demonstrate how well these methods work. The choice of metrics depends on the type of output being produced, i.e., whether it is DP synthetic data or DP summary statistics. Even within the same output category, there is no one-size-fits-all metric, and it is often recommended to use multiple utility measures to obtain a comprehensive assessment of DP methods [5, 8].

### 3 Differentially Private Quantile Regression

#### 3.1 Differentially Private Quantile Regression with KNG

The K-Norm Gradient mechanism (KNG) [37] sanitizes estimates by promoting solutions that lead the objective function gradient to be close to 0. This idea is similar to that of the Exponential mechanism [33], but KNG works with the gradient instead of the objective function. One advantage of KNG over other mechanisms is that it can offer the DP guarantee even when the objective function is not strongly convex and twice differentiable. In particular, KNG states that if the objective function  $\ell_n(\theta, D)$  is differentiable almost everywhere and  $\int \exp\left(\frac{-1}{\Delta(\theta)}\|\nabla\ell_n(\theta, D)\|_K\right)d\theta < \infty \forall D \in \mathcal{D}$ , then  $\tilde{\theta}^{KNG}$  sampled from  $\exp\left(\frac{-\epsilon}{2\Delta}\|\nabla\ell_n(\theta, D)\|_K\right)$  satisfies  $\epsilon$ -DP. Here  $\|\cdot\|_K$  denotes a norm on  $\mathbb{R}^p$  and  $\Delta$  is the sensitivity such that  $\|\nabla\ell_n(\theta, D) - \nabla\ell_n(\theta, D')\|_K \leq \Delta < \infty$  for all adjacent datasets  $D, D' \in \mathcal{D}$ . However, its utility result, which states that it introduces an asymptotically negligible amount of noise to satisfy DP, only applies when the objective function is convex and twice differentiable. In other words, using KNG with the classical check loss for DPQR is possible but may not yield optimal utility [37]. Thus, we propose using convolution-typed smoothed quantile regression loss to improve DP output utility from KNG. In this case, KNG samples from:

$$f_n(\theta) \propto \exp\left\{\frac{-\epsilon}{2\Delta}\left\|\sum_{i=1}^n \left[K_h(y_i - x_i^\top \theta) + \tau - 1\right]x_i\right\|_\infty\right\}. \quad (6)$$

We bound the sensitivity as  $\Delta = 2 \max(1 - \tau, \tau) C_X$ , where  $\max_i \|x_i\|_\infty \leq C_X$ , since  $K_h(u) = K(u/h)$  denotes the CDF of the kernel  $k(u/h)$  and is bounded between 0 and 1 (See Table 2). We choose  $\|\cdot\|_\infty$  as it generally yields good utility [6]. We now formally state the utility results of KNG when implemented with smoothed quantile regression loss and defer proofs to the Appendix.

**Theorem 1. KNG Utility Result:** Assume that the loss function  $\ell_n(\theta)$  is strongly convex, twice differentiable, and that its Hessian  $H(\theta) = \nabla^2 \ell_n(\theta)$  satisfies the matrix inequality  $\alpha_n I \leq H(\theta) \leq \beta_n I$ , where  $\alpha_n \rightarrow \infty$  and  $\beta_n \rightarrow \infty$  as  $n \rightarrow \infty$ . If  $\alpha_n^{1-\gamma} / \log(\beta_n) \rightarrow \infty$  as  $n \rightarrow \infty$  for some  $\gamma$  such that  $0 \leq \gamma < 1$ , then  $\|\tilde{\theta}_{\tau,h}^{KNG} - \hat{\theta}_{\tau,h}\| = o_p(\alpha_n^{-\gamma})$ .

**Corollary 1. KNG Utility Result with Convolution-Typed Smoothed Loss:** A sanitized estimate produced by KNG with a convolution-smoothed loss using Logistic, Gaussian, or Laplacian kernels has  $\alpha_n = n$  and  $\beta_n = n \log(1/h)$ . Assume the bandwidth  $h$  satisfies  $h \in [h_n^l, h_n^u]$  with  $1/h_n^l = o(\sqrt{n/\log(n)})$  and  $h_n^u = o(1)$ , and that  $k^{(s)}(\cdot)$  is uniformly continuous for some  $s \geq 1$ , then  $\|\tilde{\theta}_{\tau,h}^{KNG} - \theta_\tau\| = O_p(n^{-\gamma} + n^{-1/2} + h^{s+1})$  for some  $\gamma$ ,  $0 \leq \gamma < 1$ .

*Proof.* For quantile regression loss smoothed by the three listed kernels, we derive  $\alpha_n = n$  and  $\beta_n = n \log(1/h)$  and show that it satisfies the assumptions in Theorem 1 in the Appendix. Additionally, when  $h$  satisfies the stated assumptions,  $\|\hat{\theta}_{\tau,h} - \theta_\tau\| = O_p(n^{-1/2} + h^{s+1})$  (See [21]). Then,  $\|\tilde{\theta}_{\tau,h}^{KNG} - \theta_\tau\| \leq \|\tilde{\theta}_{\tau,h}^{KNG} - \hat{\theta}_{\tau,h}\| + \|\hat{\theta}_{\tau,h} - \theta_\tau\| = o_p(n^{-\gamma}) + O_p(n^{-1/2} + h^{s+1}) = O_p(n^{-\gamma} + n^{-1/2} + h^{s+1})$ .

*Remark 1.* Based on [21], when  $h = O(n^{-1/(2(s+1))})$ ,  $\|\hat{\theta}_{\tau,h} - \theta_\tau\| = O_p(n^{-1/2})$ , and if  $\gamma$  is chosen such that  $\alpha_n^{-\gamma} \leq n^{-1/2}$ , then,  $\|\tilde{\theta}_{\tau,h}^{KNG} - \theta_\tau\| = O_p(n^{-1/2})$ . This result also holds if the bandwidth is chosen based on the rule of thumb [41],  $h_{ROT} = 1.06 \min(\hat{\sigma}, \frac{IQR}{1.389}) n^{-1/5}$ , where  $\hat{\sigma}$  and  $IQR$  denote the standard deviation and the interquartile range. In the DP setting,  $\hat{\sigma}$  and  $IQR$  can be approximated using public data or other assumptions (i.e., the data is preprocessed to be between -1 and 1).

### 3.2 Differentially Private Quantile Regression with OPM

OPM, unlike KNG, requires the objective function to be strongly convex and twice differentiable almost everywhere to satisfy  $(\epsilon, \delta)$ -DP [29, 37]. Thus, it cannot be implemented with the check loss as-is without further smoothing. Chen and Chua [12] recently proposed implementing the smoothed loss with the  $(\epsilon, \delta)$ -DP version of OPM and derived its theoretical properties, including the excess generalization risk and estimation error bound. Here, we demonstrate how this applies to DPQR via OPM and refer to [12] for their theoretical results.

OPM [29] involves minimizing the objective function:

$$\ell_n(\theta, D) + \frac{(\Lambda - \gamma)^+}{2n} \|\theta\|_2^2 + \frac{b^\top \theta}{n} = \sum_{i=1}^n \rho_{\tau,h}(y_i - x_i^\top \theta) + \frac{(\Lambda - \gamma)^+}{2n} \|\theta\|_2^2 + \frac{b^\top \theta}{n}, \quad (7)$$

where  $\gamma$  denotes the strong convexity parameter,  $\Lambda \geq 2\lambda/\epsilon$  ( $\lambda$  is set as the upper bound on the eigenvalues of  $\nabla^2 \ell_n(\theta, D)$ ), and  $\zeta \geq \|\nabla \ell_n(\theta, D)\|$ . Specifically, in this setting, we set the values of  $\lambda$  and  $\zeta$  as:  $\|\nabla^2 \ell_n(\theta, D)\|_2 = \|\sum_{i=1}^n k_h(y_i - x_i^\top \theta) x_i x_i^\top\|_2 \leq \frac{n}{h} \bar{k} B_X^2 = \lambda$  and  $\|\nabla \ell_n(\theta, D)\|_2 = \|\sum_{i=1}^n [K_h(y_i - x_i^\top \theta) + \tau - 1] x_i\|_2 \leq n \max(\tau, 1 - \tau) B_X = \zeta$ , where  $B_X = \max_i \|x_i\|_2$  and  $\bar{k} = \sup_u k(u)$  (see Table 2). OPM satisfies  $\epsilon$ -DP if  $b$  is sampled from a Gamma distribution with density  $\exp(-\epsilon\|b\|_2/(2\zeta))$ , or  $(\epsilon, \delta)$ -DP if it is sampled from a Normal distribution with mean 0 and variance  $(\zeta^2(8\log(2/\delta) + 4\epsilon)/\epsilon^2)I_{p \times p}$ . Although [12] implemented the  $(\epsilon, \delta)$ -DP version of OPM, we use the  $\epsilon$ -DP version in this paper for a fair comparison with KNG.

**Table 2.** Characteristics of Commonly Used Kernels [12, 24, 44]

Kernels	$k(u)$	$K_h(u) = K\left(\frac{u}{h}\right) = \int_{-\infty}^{u/h} k(v)dv$	$\bar{k} = \sup k$
Gaussian	$(2\pi)^{-1/2} e^{-u^2/2}$	$\Phi(u/h)$	$(2\pi)^{-1/2}$
Logistic	$e^{-u}/(1 + e^{-u})^2$	$(1 + e^{-u/h})^{-1}$	$1/4$
Uniform	$(1/2)\mathbb{1}\{ u  \leq 1\}$	$\min((u/h + 1)/2, 1)\mathbb{1}\{u/h \geq -1\}$	$1/2$
Laplacian	$e^{- u }/2$	$1/2 + 1/2 \text{sign}(x)(1 - \exp(- x ))$	$1/2$
Epanechnikov	$(3/4)(1 - u^2)\mathbb{1}\{ u  \leq 1\}$	$\begin{aligned} & [(3/4)(u/h) - (1/4)(u/h)^3 + 1/2]\mathbb{1}\{ u/h  \leq 1\} \\ & + 1\{ u/h  > 1\} \end{aligned}$	$3/4$

### 3.3 Differentially Private Quantile Regression with DP-SGD

Chen and Chua [12] also proposed performing DPQR via the DP-SGD version introduced by [7] and the convolution-smoothed loss. After initializing the starting point, the algorithm will uniformly sample observation  $(x_{(t)}, y_{(t)})$  from the data and perform the following step update  $n^2 - 1$  times:

$$\hat{\theta}_{\tau,h,t+1} = \hat{\theta}_{\tau,h,t} - \eta(\nabla \ell_n(\hat{\theta}_{\tau,h,t}, x_{(t)}, y_{(t)}) + Z_t), \quad (8)$$

where  $Z_t$  is the noise randomly sampled from the multivariate Normal distribution with mean 0 and variance  $(8L^2 \log(1/\delta)/\epsilon^2)I_{p \times p}$  ( $L$  is the Lipschitz parameter of the loss function). The final sanitized output  $\tilde{\theta}_{\tau,h}^{SGD}$  is obtained by averaging over  $n^2$  steps. The private estimate  $\tilde{\theta}_{\tau,h}^{SGD}$  satisfies  $(\epsilon, \delta)$ -DP based on results of [7]. We refer to [12] for further theoretical results regarding the asymptotic optimality of this algorithm.

### 3.4 Differentially Private Quantile Regression with DP-AGD

In this section, we implement DPQR with  $\rho$ -zCDP guarantee through the use of smoothed quantile regression loss within DP-AGD; to the best of our knowledge, first such implementation. To satisfy DP, DP-AGD requires the loss function to be twice differentiable and strongly convex, a requirement that previously hindered its application in quantile regression. Similar to previous works on DP

gradient descent [1], DP-AGD uses the noisy gradient to iteratively update steps  $\hat{\theta}_{\tau,h,t+1} = \hat{\theta}_{\tau,h,t} - \eta_t(\nabla \ell_n(\theta_t, D) + Z_t)$ , where  $\eta_t$  denotes the learning rate and  $Z_t \in \mathbb{R}^p$  is the random noise from the Normal distribution  $N(0, \Delta^2/(2\rho_t))$  and  $\rho_t$  is the privacy-loss budget at iteration  $t$ . The sanitized output from this algorithm will satisfy  $\rho$ -zCDP with  $\rho = \sum_t \rho_t$ , instead of the  $\epsilon$ - or  $(\epsilon, \delta)$ -DP framework typically found in other DP gradient descent algorithms.

One distinguishing feature of DP-AGD is its ability to dynamically and privately set the privacy budget at each iteration. This removes the need to divide the privacy budget based on a predetermined number of iterations, thereby avoiding setting the iteration count too high leading to excessive noise, or too low resulting in the algorithm not reaching the optimum. Thus, it can lead to a more efficient privacy-loss budget usage. Initially, in the optimization process, the gradient is large, thus requiring less privacy budget to move in the minimizing direction. As the process advances, the gradient decreases, necessitating more budget to continue in the right direction. If a proposed update falls below a prespecified utility threshold, the algorithm prompts an increase in the privacy budget to  $\rho_{t+1}$ . However, the gradient computed with  $\rho_t$  is not discarded. Instead, another gradient is computed with the difference in the privacy budget ( $S'_t = \nabla \ell_n(\theta_t, D) + N(0, \Delta^2/(2(\rho_{t+1} - \rho_t)))$ ) and the weighted average of the two gradients is used to improve the update quality and salvage the previously spent privacy budget. This approach incurs the same privacy loss and variance as if we had known in advance that using  $\rho_t$  was insufficient and used  $\rho_{t+1}$  instead.

Another innovative aspect of DP-AGD is the dynamic step-size tuning using a portion of the privacy-loss budget, instead of simply using a rule-of-thumb step size (i.e., varying with the iteration  $1/t$ ). It allows DP-AGD to converge to the optimum quickly. We refer to [31] for further details about the algorithm.

## 4 Simulation Study

In each study, we simulate 500 different datasets using different seeds and record the average L2 distance to the true coefficients at six quantiles  $\tau = 0.05, 0.25, 0.5, 0.75, 0.95$ , and  $0.99$ . We replicate this procedure for each combination of  $n = 2500, 250$  and  $\epsilon = 1, 0.5$ . For DP-AGD, we plug  $\epsilon$  and  $\delta = 1e-8$  into Eq. 1 and solve for  $\rho$ . As we do not observe significant differences in runtime across quantiles, we record one mean and one standard error of the runtime in seconds for each method under each scenario (when run on a 2.2 GHz CPU with 2GB memory). For a straightforward comparison, we report the results of the Logistic kernel at bandwidth  $h = 0.5$  for the four mechanisms, as this kernel-bandwidth combination generally yields good accuracy across methods. The code and extended simulation are available at [github.com/tranntran/dpqr\\_rep](https://github.com/tranntran/dpqr_rep).

### 4.1 Simulation Study 1: Normal Noise Data

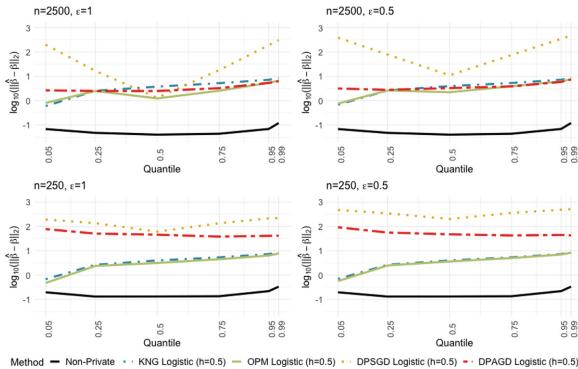
We first explore how the four methods perform when applied to data with random noise from the normal distribution. In particular, the data generating pro-

cess is as follows:  $x_1 \sim N(0, 3)$ ,  $x_2 \sim N(0, 1)$ , and  $y = 3 + 2x_1 - x_2 + \zeta$ , where  $\zeta \sim N(0, 2)$ . The upper bound of  $x$  is constrained to  $C_X = 10$ .

Our simulation reveals that OPM, KNG, and DP-AGD perform comparably under both privacy-loss budget settings when  $n$  is large (See Fig. 1). DP-SGD, however, works well only at the median,  $\tau = 0.5$ , and deteriorates in accuracy at extreme quantiles. As the sample size  $n$  decreases, OPM and KNG maintain their close approximation to the truth, while DP-AGD suffers a significant loss of accuracy. Overall, we observe that OPM and KNG can offer formal privacy protection at a minimal cost of accuracy compared to the non-private estimates.

In terms of runtime (see Table 3), OPM consistently outperforms other methods across all simulation setups. Compared to DP-SGD, DP-AGD greatly reduces runtime by 97% (from approximately 430 s to 12 s) when  $n = 2500$  and by 60% (from roughly 4.4 to 1.8 s) when  $n = 250$ . KNG comes third with large sample sizes but takes the longest as they get smaller.

Overall, considering the privacy-utility trade-off and runtime, with the normal noise and when the sample size is large, we recommend using OPM and DP-AGD. However, when the sample size gets smaller, OPM and KNG are better options; KNG's slightly longer runtime is justified by the better accuracy.



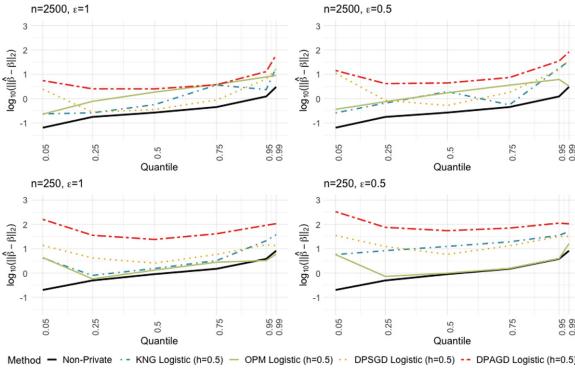
**Fig. 1.** L2 distance to the true coefficients when the data has normal distributed noise under different sample sizes and privacy-loss budget

## 4.2 Simulation Study 2: Heavy-Tailed Data

In this simulation, we focus on heavy-tailed data, which often appears in economic data (e.g., income). We generate data using the following process  $y = 3 + 2x_1 + 0.5x_2 + \zeta$ , where  $x_1$  and  $x_2 \sim U(0, 1)$  and  $\zeta \sim Exp(0.1)$ . Comparisons of accuracy and runtime across methods are in Fig. 2 and Table 3, respectively.

OPM outperforms all other methods across quantiles, sample sizes, and privacy-loss budgets in accuracy and runtime. KNG is the next best performer in utility, followed by DP-SGD and DP-AGD. Particularly, KNG shows high utility in comparison to non-private estimates and OPM across all quantiles and

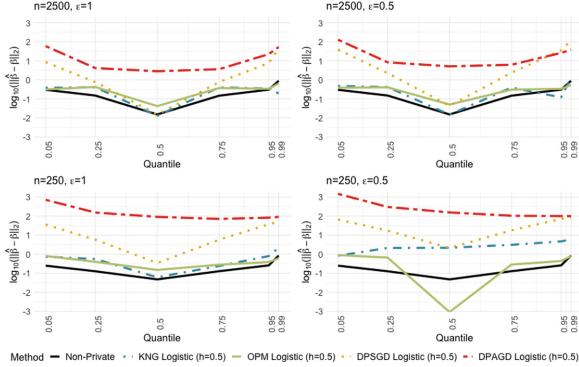
scenarios, barring the extreme case at  $\tau = 0.99$ . DP-SGD shows a decline in utility at both  $\tau = 0.05$  and  $0.99$  in all settings. DP-AGD performs well with larger sample sizes, but falls short as they decrease. Among the gradient-based methods, KNG outperforms DP-SGD when sample sizes are large, as it can maintain similar levels of utility at one-fourth of the runtime. However, they yield comparable accuracy and runtime with smaller sample sizes. Between DP-SGD and DP-AGD, when  $n = 2500$ , DP-AGD cuts the runtime from 400 to 25 s at the cost of some utility, but the trade-off is no longer worth it when  $n = 250$ .



**Fig. 2.** L2 distance to the true coefficients when the data is heavy-tailed under different sample sizes and privacy-loss budget

### 4.3 Simulation Study 3: Heteroscedastic Data

In the final simulation study, we test the four methods on a harder problem, heteroscedastic data. We generate  $x \sim U(0, 2)$ ,  $y = 1 + 2x + x * \zeta$  with  $\zeta \sim N(0, 1)$ , and naturally bound  $C_X = 2$ . Based on the average distance to the true coefficients shown in Fig. 3, estimates produced by OPM and KNG closely follow the non-private one in all different setups. DP-SGD achieves comparable utility with OPM and KNG at the median, but its performance worsens when moving toward the two extremes. Although DP-AGD cuts down the runtime a lot compared to DP-SGD (17 vs 400 s) with  $n = 2500$ , it is not worth the accuracy trade-off as DP-AGD does not work well with heteroscedastic data. We also observe the same pattern with DP-SGD's runtime in the previous two simulation studies. When taking both accuracy and runtime into account, we recommend using OPM and KNG when dealing with heteroscedastic data.



**Fig. 3.** L2 distance to the true coefficients when the data is heteroscedastic under different sample sizes and privacy-loss budget

**Table 3.** Runtime of each method to compute one quantile in seconds (Mean  $\pm$  SE)

	Method	$n = 2500, \epsilon = 1$	$n = 2500, \epsilon = 0.5$	$n = 250, \epsilon = 1$	$n = 2500, \epsilon = 0.5$
Simulation 1	Non-Private	$0.006 \pm 0.00$	$0.006 \pm 0.00$	$0.002 \pm 0.00$	$0.002 \pm 0.00$
	KNG Logistic (h=0.5)	$79.3 \pm 1.15$	$75.9 \pm 0.97$	$8.4 \pm 0.11$	$8.2 \pm 0.10$
	OPM Logistic (h=0.5)	$4.2 \pm 0.04$	$4.2 \pm 0.04$	$0.4 \pm 0.00$	$0.4 \pm 0.00$
	DP-SGD Logistic (h=0.5)	$435.1 \pm 0.94$	$440.1 \pm 1.66$	$4.4 \pm 0.01$	$4.5 \pm 0.02$
	DP-AGD Logistic (h=0.5)	$12.9 \pm 0.23$	$12.4 \pm 0.22$	$1.8 \pm 0.04$	$1.5 \pm 0.03$
Simulation 2	Non-Private	$0.005 \pm 0.00$	$0.005 \pm 0.00$	$0.002 \pm 0.00$	$0.002 \pm 0.00$
	KNG Logistic (h=0.5)	$135.1 \pm 1.20$	$129.5 \pm 1.43$	$12.5 \pm 0.08$	$11.3 \pm 0.13$
	OPM Logistic (h=0.5)	$2.8 \pm 0.04$	$2.8 \pm 0.05$	$0.2 \pm 0.00$	$0.2 \pm 0.01$
	DP-SGD Logistic (h=0.5)	$420.5 \pm 0.75$	$438.6 \pm 2.00$	$4.4 \pm 0.01$	$4.2 \pm 0.01$
	DP-AGD Logistic (h=0.5)	$27.5 \pm 0.52$	$25.1 \pm 0.49$	$11.8 \pm 0.38$	$8.0 \pm 0.20$
Simulation 3	Non-Private	$0.005 \pm 0.00$	$0.005 \pm 0.00$	$0.002 \pm 0.00$	$0.002 \pm 0.00$
	KNG Logistic (h=0.5)	$121.5 \pm 1.24$	$103.9 \pm 1.50$	$10.9 \pm 0.16$	$8.9 \pm 0.13$
	OPM Logistic (h=0.5)	$3.2 \pm 0.02$	$3.2 \pm 0.03$	$0.3 \pm 0.00$	$0.3 \pm 0.01$
	DP-SGD Logistic (h=0.5)	$423.0 \pm 0.90$	$417.9 \pm 0.87$	$4.5 \pm 0.04$	$4.3 \pm 0.01$
	DP-AGD Logistic (h=0.5)	$17.9 \pm 0.40$	$16.7 \pm 0.37$	$4.0 \pm 0.08$	$3.2 \pm 0.07$

## 5 Conclusion

Quantile regression is a powerful statistical tool with applications in many disciplines. However, it remains understudied within the DP literature, where most of the research has focused on median regression or quantile estimation. In this paper, we consider the general QR setting, and propose smoothing the quantile loss via convolution before implementing it within two DP mechanisms, KNG and DP-AGD. We evaluate their performance against two previously proposed DPQR with OPM and DP-SGD in different data settings. We note the previous implementations focused only on QR for  $\tau = 0.7$ . Our results demonstrate that OPM performs the best in accuracy and runtime. Within gradient-based methods, KNG is the most consistent in performance across different data types, sample sizes, and privacy-loss budgets. We also prove the theoretical utility result

that KNG adds an asymptotically negligible amount of noise in exchange for the DP guarantee. While OPM outperforms KNG in general scenarios, it is worth noting that KNG can be more beneficial under certain conditions. For instance, when linear constraints are implemented within the parameter space, KNG inherently satisfies DP, which is not guaranteed with OPM. Finally, our extensive simulation studies serve as a starting guide for applying DP quantile regression on diverse data types and estimates of interest.

**Acknowledgments.** This research was supported in part by NSF Grant #1853209 and #1702760 to The Pennsylvania State University, and the Huck Award funds.

## Appendix: Proofs

### Proofs for Theorem 1

Assume that the loss function  $\ell_n(\theta)$  is strongly convex and twice differentiable and that its Hessian satisfies  $\alpha_n I \leq H_n(\theta) \leq \beta_n I$ , where  $\alpha_n \rightarrow \infty$ ,  $\beta_n \rightarrow \infty$ . We further assume that  $\alpha_n$  and  $\beta_n$  satisfy  $\alpha_n^{1-\gamma} / \log(\beta_n) \rightarrow \infty$  as  $n \rightarrow \infty$ , for some  $0 \leq \gamma < 1$ . We wish to show that  $\alpha_n^\gamma \|\tilde{\theta} - \hat{\theta}\| = o_P(1)$ , or equivalently,  $\|\tilde{\theta} - \hat{\theta}\| = o_P(\alpha_n^{-\gamma})$ . By definition, this means we need to show that, for any fixed  $c > 0$  we have that  $P(\tilde{\theta} \in A) \rightarrow 0$ , where  $A = \{\theta : \|\theta - \hat{\theta}\| \geq c\alpha_n^{-\gamma}\}$ .

Consider the density of KNG  $f_n(\theta) = c_n^{-1} \exp\left\{-\frac{\epsilon}{2\Delta} \|\nabla \ell_n(\theta)\|_K\right\}$ , with  $c_n$  as the normalizing constant. Then,  $c_n = \int \exp\left\{-\frac{\epsilon}{2\Delta} \|\nabla \ell_n(\theta)\|\right\}$ . By the mean value theorem we have that  $\|\nabla \ell_n(\theta)\| = \|\nabla \ell_n(\theta) - \nabla \ell_n(\hat{\theta})\| = \|H_n(\xi)(\theta - \hat{\theta})\|$ , where  $\xi$  is some point on the line connecting  $\theta$  and  $\hat{\theta}$ . Note that  $\nabla \ell_n(\hat{\theta}) = 0$  by definition. This implies that  $\alpha_n \|\theta - \hat{\theta}\| \leq \|\nabla \ell_n(\theta)\| \leq \beta_n \|\theta - \hat{\theta}\|$ . Thus, the integration of constant  $c_n$  is then bounded by

$$\int \exp\left\{-\frac{\epsilon \beta_n}{2\Delta} \|\theta - \hat{\theta}\|\right\} \leq \int \exp\left\{-\frac{\epsilon}{2\Delta} \|\nabla \ell_n(\theta)\|\right\} \leq \int \exp\left\{-\frac{\epsilon \alpha_n}{2\Delta} \|\theta - \hat{\theta}\|\right\}.$$

Assuming we are integrating over all of  $\mathbb{R}^d$ , we have that

$$\int \exp\left\{-\frac{\epsilon \alpha_n}{2\Delta} \|\theta - \hat{\theta}\|\right\} = \int \exp\left\{-\frac{\epsilon \alpha_n}{2\Delta} \|\theta\|\right\} = \frac{2\Delta}{\epsilon \alpha_n} C,$$

where  $C > 0$  is some constant. Consequently, we have  $\frac{2\Delta}{\epsilon \beta_n} C \leq c_n \leq \frac{2\Delta}{\epsilon \alpha_n} C$  and

$$\begin{aligned} f_n(\theta) &\leq \frac{\epsilon \beta_n}{2\Delta C} \exp\left\{-\frac{\epsilon \alpha_n}{2\Delta} \|\theta - \hat{\theta}\|\right\} \leq \frac{\epsilon \beta_n}{2\Delta C} \exp\left\{-\frac{\epsilon c \alpha_n^{1-\gamma}}{2\Delta}\right\} \\ &= \exp\left\{-\frac{\epsilon c \alpha_n^{1-\gamma}}{2\Delta} + \log(\beta_n) + \log(\epsilon/(2\Delta C))\right\}. \end{aligned}$$

Since we assume that  $\alpha_n^{1-\gamma}/\log(\beta_n) \rightarrow \infty$ , we have  $f_n(\theta) \rightarrow 0$  on  $A$ . We now move on to prove that  $f_n(\theta)1_{\theta \in A} \leq g(\theta)$  where  $g(\theta)$  is integrable. We have:

$$\begin{aligned} f_n(\theta) &\leq \frac{\epsilon}{2\Delta C} \exp \left\{ -\frac{\epsilon\alpha_n}{2\Delta} \|\theta - \hat{\theta}\| + \log(\beta_n) \right\} \\ &= \frac{\epsilon}{2\Delta C} \exp \left\{ -\frac{\epsilon\alpha_n}{2\Delta} \|\theta - \hat{\theta}\| \left[ 1 - \frac{2\Delta \log(\beta_n)}{\epsilon\alpha_n \|\theta - \hat{\theta}\|} \right] \right\}. \end{aligned} \quad (9)$$

First, since  $\alpha_n \rightarrow \infty$ , eventually, for  $n$  large we have  $\alpha_n \geq 1$ . Next, since  $\|\theta - \hat{\theta}\| \geq c\alpha_n^{-\gamma}$  and  $\log(\beta_n)/\alpha_n^{1-\gamma} \rightarrow 0$  we have that

$$\frac{2\Delta \log(\beta_n)}{\epsilon\alpha_n \|\theta - \hat{\theta}\|} \leq \frac{2\Delta \log(\beta_n)}{\epsilon\alpha_n^{1-\gamma} c} \rightarrow 0, \text{ meaning for large } n, \frac{2\Delta \log(\beta_n)}{\epsilon\alpha_n^{1-\gamma} c} \leq 1/2.$$

Putting things together, for  $n$  large,  $f_n(\theta)1_{\theta \in A} \leq \frac{\epsilon}{2\Delta C} \exp \left\{ -\frac{\epsilon}{4\Delta} \|\theta - \hat{\theta}\| \right\}$ , which is integrable. Applying the Dominated Convergence Theorem, we get  $\lim_{n \rightarrow \infty} \int_A f_n(\theta) = 0$ . We can conclude that  $P(\tilde{\theta} \in A) \rightarrow 0 \forall c > 0$ .

## Proofs for Corollary 1

(Gaussian and Laplacian kernels follow similarly)

**Lower bound:** When the kernel is Logistic, we have  $k(x) = e^{-x}(1+e^{-x})^{-2} \geq e^{-x}$ . Let  $|y_i - x_i^\top \theta| \leq \delta$  ( $\delta > 0$ ) and  $A = \{i : |y_i - x_i^\top \theta| \leq \delta\}$ . Then,

$$\nabla^2 \ell(\theta) = \sum_{i=1}^n \frac{1}{h} k \left( \frac{|y_i - x_i^\top \theta|}{h} \right) x_i x_i^\top \geq \sum_{i \in A^c} \frac{1}{h} k \left( \frac{\delta}{h} \right) x_i x_i^\top \geq \sum_{i \in A^c} \frac{1}{h} e^{-\frac{\delta}{h}} x_i x_i^\top$$

If we set  $\delta = h \log(\frac{1}{h})$ , then as  $n \rightarrow \infty$  ( $h \rightarrow 0$  and  $\delta \rightarrow 0$ ),

$$\nabla^2 \ell(\theta) \geq \sum_{i \in A^c} \frac{1}{h} \exp \left( -\frac{h \log(\frac{1}{h})}{h} \right) x_i x_i^\top = \sum_{i \in A^c} x_i x_i^\top \approx \sum_{i=1}^n x_i x_i^\top.$$

**Upper bound:**

Case 1:  $|y_i - x_i^\top \theta| \leq \delta$ . Assume the size of  $A \approx \delta f_\theta(0)n$ , then

$$\sum_{i \in A} \frac{1}{h} k \left( \frac{|y_i - x_i^\top \theta|}{h} \right) x_i x_i^\top \leq \sum_{i \in A} \frac{1}{h} x_i x_i^\top \approx \frac{n\delta}{h} E[x_i x_i^\top]$$

Case 2:  $|y_i - x_i^\top \theta| \geq \delta$

$$\sum_{i \in A^c} \frac{1}{h} k \left( \frac{|y_i - x_i^\top \theta|}{h} \right) x_i x_i^\top \leq \frac{1}{h} k \left( \frac{\delta}{h} \right) \sum_{i=1}^n x_i x_i^\top \leq \frac{n}{h} E[x_i x_i^\top].$$

Since  $h \in [h^l, h^u]$ ,  $1/h^l = o(\sqrt{n \log(n)})$ ,  $h^u = o(1)$ , and  $\delta = h \log(\frac{1}{h})$ , the lower growth rate of the upper bound of the Hessian matrix is  $n \log(1/h) \leq$

$n \log(\sqrt{n \log(n)})$  ( $= \beta_n$ ). Since  $\alpha_n = n$  from above, the logistic kernel satisfies assumptions in Theorem 1:  $\alpha_n^{1-\gamma}/\log(\beta_n) = n^{1-\gamma}/\log(n \log(\sqrt{n \log(n)})) \rightarrow \infty$  as  $n \rightarrow \infty$  for some  $\gamma$  and, for large enough  $n$ ,  $(2\Delta \log(\beta_n))/(c\alpha_n^{1-\gamma}/c) = (2\Delta \log(n \log(\sqrt{n \log(n)})))/(c\epsilon n^{1-\gamma}/c) \leq 1/2$ . Thus, it follows that KNG's sanitized estimate converges to the unsanitized one asymptotically.

## References

1. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016)
2. Abowd, J., et al.: The 2020 Census Disclosure Avoidance System TopDown Algorithm. Harvard Data Science Review (2022)
3. Abowd, J.M., et al.: The modernization of statistical disclosure limitation at the U.S. Census Bureau (2020). <https://www.census.gov/library/working-papers/2020/adrm/CED-WP-2020-009.html>. Accessed 6 Mar 2024
4. Apple: Learning with privacy at scale (2017). <https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf>. Accessed 10 Sept 2023
5. Arnold, C., Neunhoeffer, M.: Really useful synthetic data—a framework to evaluate the quality of differentially private synthetic data. [arXiv:2004.07740](https://arxiv.org/abs/2004.07740) (2020)
6. Awan, J., Slavković, A.: Structure and sensitivity in differential privacy: comparing k-norm mechanisms. J. Am. Stat. Assoc. **116**(534), 935–954 (2021)
7. Bassily, R., Feldman, V., Talwar, K., Thakurta, A.: Private stochastic convex optimization with optimal rates. In: NeurIPS (2019)
8. Bowen, C.M.: Utility for differential privacy: no one size fits all (2021). <https://www.nist.gov/blogs/cybersecurity-insights/utility-metrics-differential-privacy-no-one-size-fits-all>. Accessed 9 Jan 2024
9. Brunel, V.E., Avella-Medina, M.: Propose, test, release: differentially private estimation with high probability. arXiv preprint [arXiv:2002.08774](https://arxiv.org/abs/2002.08774) (2020)
10. Bun, M., Steinke, T.: Concentrated differential privacy: simplifications, extensions, and lower bounds. In: Hirt, M., Smith, A. (eds.) TCC 2016. LNCS, vol. 9985, pp. 635–658. Springer, Heidelberg (2016). [https://doi.org/10.1007/978-3-662-53641-4\\_24](https://doi.org/10.1007/978-3-662-53641-4_24)
11. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. J. Mach. Learn. Res. **12**(3) (2011)
12. Chen, D., Chua, G.: Differentially private stochastic convex optimization under quantile loss. In: International Conference on Machine Learning. PMLR (2023)
13. Chen, E., Miao, Y., Tang, Y.: Median regression with DP. [arXiv:2006.02983](https://arxiv.org/abs/2006.02983) (2020)
14. Dabney, W., Rowland, M., Bellemare, M., Munos, R.: Distributional reinforcement learning with quantile regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
15. Ding, B., Kulkarni, J., Yekhanin, S.: Collecting telemetry data privately. arXiv preprint [arXiv:1712.01524](https://arxiv.org/abs/1712.01524) (2017)
16. Drechsler, J., Globus-Harris, I., Mcmillan, A., Sarathy, J., Smith, A.: Nonparametric differentially private confidence intervals for the median. J. Survey Stat. Methodol. **10**(3), 804–829 (2022)
17. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Minimax optimal procedures for locally private estimation. J. Am. Stat. Assoc. (2018)

18. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). [https://doi.org/10.1007/11681878\\_14](https://doi.org/10.1007/11681878_14)
19. Dwork, C., Smith, A., Steinke, T., Ullman, J.: Exposed! a survey of attacks on private data. *Ann. Rev. Stat. Appl.* **4**, 61–84 (2017)
20. Erlingsson, Ú., Pihur, V., Korolova, A.: RAPPOR: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 1054–1067 (2014)
21. Fernandes, M., Guerre, E., Horta, E.: Smoothing quantile regressions. *J. Bus. Econ. Stat.* **39**(1), 338–357 (2021)
22. Garfinkel, S.: Differential Privacy and the 2020 US Census. MIT Case Studies in Social and Ethical Responsibilities of Computing (2022)
23. Gillenwater, J., Joseph, M., Kulesza, A.: Differentially private quantiles. In: International Conference on Machine Learning, pp. 3713–3722. PMLR (2021)
24. He, X., Pan, X., Tan, K.M., Zhou, W.X.: Smoothed quantile regression with large-scale inference. *J. Econ.* **232**(2), 367–388 (2023)
25. Horowitz, J.: Bootstrap methods for median regression. *Econometrica* (1998)
26. Jarin, I., Eshete, B.: DP-UTIL: comprehensive utility analysis of differential privacy in machine learning. In: Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy, pp. 41–52 (2022)
27. Kaplan, D.M., Sun, Y.: Smoothed estimating equations for instrumental variables quantile regression. *Economet. Theor.* **33**(1), 105–157 (2017)
28. Kaplan, H., Schnapp, S., Stemmer, U.: Differentially private approximate quantiles. In: International Conference on Machine Learning, pp. 10751–10761. PMLR (2022)
29. Kifer, D., Smith, A., Thakurta, A.: Private convex empirical risk minimization and high-dimensional regression. In: Conference on Learning Theory. JMLR (2012)
30. Koenker, R., Bassett Jr, G.: Regression quantiles. *Econometrica* (1978)
31. Lee, J., Kifer, D.: Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018)
32. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: Conference on Data Engineering. IEEE (2008)
33. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science. IEEE (2007)
34. Oberski, D.L., Kreuter, F.: Differential privacy and social science: An urgent puzzle. *Harvard Data Sci. Rev.* **2**(1), 1–21 (2020)
35. Pejó, B., Desfontaines, D.: Guide To Differential Privacy Modifications: A Taxonomy of Variants and Extensions. Springer, Berlin (2022)
36. Ramsay, K., Jagannath, A., Chenouri, S.: Concentration of the exponential mechanism and differentially private multivariate medians. [arXiv:2210.06459](https://arxiv.org/abs/2210.06459) (2022)
37. Reimherr, M., Awan, J.: KNG: the k-norm gradient mechanism. In: NeuRIPS, vol. 32 (2019)
38. Rogers, R., et al.: LinkedIn’s audience engagements API: a privacy preserving data analytics system at scale. [arXiv:2002.05839](https://arxiv.org/abs/2002.05839) (2020)
39. Ruggles, S., Fitch, C., Magnuson, D., Schroeder, J.: DP and Census data: Implications for social and economic research. In: AEA Papers and Proceedings (2019)
40. Sherwood, B., Wang, L., Zhou, X.H.: Weighted quantile regression for analyzing health care cost data with missing covariates. *Stat. Med.* (2013)
41. Silverman, B.: Density estimation for statistics and data analysis. Routledge (2018)

42. Smith, A.: Privacy-preserving statistical estimation with optimal convergence rates. In: Proceedings of the 43th ACM Symposium on Theory of Computing (2011)
43. Snöke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. *J. R. Stat. Soc. Ser. A Stat. Soc.* **181**(3), 663–688 (2018)
44. Tan, K.M., Wang, L., Zhou, W.X.: High-dimensional quantile regression: convolution smoothing and concave regularization. *J. R. Stat. Soc. Ser. B Stat Methodol.* **84**(1), 205–233 (2022)
45. Whang, Y.J.: Smoothed empirical likelihood methods for quantile regression models. *Economet. Theor.* **22**(2), 173–205 (2006)
46. Wood, A., et al.: Differential privacy: a primer for a non-technical audience. *Vand. J. Ent. Tech. L.* **21**, 209 (2018)
47. You, W., Guo, Y., Zhu, H., Tang, Y.: Oil price shocks, economic policy uncertainty and industry stock returns in china: asymmetric effects with quantile regression. *Energy Econ.* **68**, 1–18 (2017)
48. Zheng, S.: Gradient descent algorithms for quantile regression with smooth approximation. *Int. J. Mach. Learn. Cybern.* (2011)



# Utility Analysis of Differentially Private Anonymized Data Based on Random Sampling

Takumi Sugiyama<sup>1(✉)</sup>, Hiroto Oosugi<sup>3,4</sup>, Io Yamanaka<sup>2</sup>,  
and Kazuhiro Minami<sup>4(✉)</sup>

<sup>1</sup> Graduate School of Economics, Chuo University, Hachioji, Japan  
[a19.dafe@g.chuo-u.ac.jp](mailto:a19.dafe@g.chuo-u.ac.jp)

<sup>2</sup> Rakus Co., Ltd., Shinagawa City, Japan  
<sup>3</sup> SECOM CO., LTD., Shibuya City, Japan

<sup>4</sup> Department of Interdisciplinary Statistical Mathematics,  
The Institute of Statistical Mathematics, Tachikawa, Japan  
[kminami@ism.ac.jp](mailto:kminami@ism.ac.jp)

**Abstract.** It is possible to produce differentially private  $k$ -anonymized data based on the method of random sampling followed by full-domain generalization for  $k$ -anonymization. We previously evaluate the performance of that method, which is implemented as the SafePub algorithm in the ARX anonymization tool. However, since the SafePub algorithm uses the maximum sampling rate that satisfies the requirements for differential privacy, we observe paradoxical results where data utility diminishes as the privacy budget for differential privacy increases.

In this paper, we, therefore, conduct preliminary experiments to explore the parameter space for privacy budget and sampling rate by setting the sampling rates explicitly through modifications to the implementation of ARX. Our initial results show the possibility of improving the utility of anonymized data by properly setting the sampling rate below its maximum value.

**Keywords:** Differential Privacy · Random Sampling ·  $k$ -anonymity

## 1 Introduction

Random sampling is a promising way to produce  $k$ -anonymized data [18] that satisfies differential privacy [3]. Originally, Li et al. [10] proposed a scheme of producing differentially private  $k$ -anonymized data by applying random sampling as preprocessing and then using a  $k$ -anonymization algorithm based on full-domain generalization in which the generalization level of each variable is determined in a differentially private manner. Subsequently, Bild et al. implemented this scheme as the SafePub algorithm in the ARX anonymization tool [2], which determines an appropriate anonymization parameter  $k$  and a sampling rate  $\beta$  to achieve  $(\epsilon, \delta)$ -differential privacy.

We previously conduct experiments with the SafePub algorithm to verify whether it can produce anonymized data of high utility [17]. However, we find that SafePub fails to produce high utility data when we set privacy budget  $\epsilon$  to be greater than 3 because the SafePub algorithms uses the maximum sampling rate  $\beta_{max}$  that satisfies  $(\epsilon, \delta)$ -differential privacy. As sampling rate  $\beta_{max}$  converges to 1 when  $\epsilon \geq 3$ , the value of anonymization parameter  $k$ , which determines the minimum allowable size of equivalence classes in  $k$ -anonymized data, increases rapidly, thereby deteriorating the utility of the resulting anonymized data. This phenomenon interestingly contradicts our expectation for differential privacy: as we allocate more privacy budget  $\epsilon$ , the utility of produced anonymized data degrades significantly.

The SafePub algorithm chooses the maximum sampling rate  $\beta_{max} = 1 - e^{-\epsilon}$ , which is the upper bound of the safety condition  $\epsilon \geq -\ln(1 - \beta)$  in Theorem 3 of [2]. The developers of the *SafePub* algorithm makes this decision based on the assumption that information loss due to unsampled records in the sampling process dominantly determines the utility of anonymized data. However, when we evaluate the utility of anonymized data, we find that it is crucial to consider the intricate balance between information loss due to unsampled records and the utility of anonymized data produced from the remaining sampled records. For example, statistical methods in sampling theory [12] could guarantee accurate confidence intervals for estimates of population statistics from sampled data.

In this paper, we, therefore, examine the possibility of producing anonymized data of high data utility by exploring lower sampling rates within the parameter space below the upper bound  $\beta_{max}$ . We conduct experiments with *US Census Adult Dataset* [1] to study how the combination of privacy budget  $\epsilon$  and sampling rate  $\beta$  affects the prediction accuracy of a classifier trained on the resulting anonymized data. Our preliminary results show the possibility of improving the utility of anonymized data by setting the sampling rate below its upper bound such that the prediction accuracy of the trained model is significantly improved. Although lowering the sampling rate decreases the number of preserved records, which seems to affect the utility of the data, we observe that reducing the value of the anonymization parameter  $k$  is more crucial for retaining the utility of anonymized data, which is determined largely based on the sizes of the equivalence classes.

The rest of the paper is organized as follows. Section 2 describes the scheme for generating differentially private  $k$ -anonymized data based on random sampling and describes the issue of setting the sampling rate appropriately. Section 3 shows our initial experimental results, where we evaluate the utility of anonymized data based on the accuracy of statistical classification. Section 4 discusses related work on the experimental results by the authors of the *SafePub* algorithm, and Sect. 5 provides concluding remarks.

## 2 Background

This section introduces the notions of  $k$ -anonymity [16, 18] and differential privacy [4], and describes the scheme of differentially private  $k$ -anonymization based

on random sampling [10], which fills the gap between two privacy notions. We finally discuss the issue of choosing the sampling rate to obtain anonymized data of high utility.

## 2.1 $k$ -Anonymity

$k$ -anonymity [16, 18] is widely recognized privacy metric for anonymized data. In  $k$ -anonymity, anonymized data consists of equivalence classes, each containing records with the same set of quasi-identifier values.  $k$ -anonymity requires that the size of every equivalence class is greater than or equal to a given integer value  $k$  such that an adversary cannot reduce the candidate records to less than  $k$  records by conducting record linkage attacks using an external dataset. Choosing a larger value for privacy parameter  $k$  enhances the safety of anonymized data, while concurrently reducing its utility.

Previous research proposes many different  $k$ -anonymization algorithms based on generalization hierarchies [8, 9], clustering [11], and so on. This paper considers  $k$ -anonymization algorithms based on full domain generalization and suppression [9] where a generalization operator  $g$  maps the domains of the quasi-identifier attributes to generalized or altered values.

## 2.2 $\epsilon$ -Differential Privacy

Differential privacy [4] is a rigorous privacy metric that ensures uncertainty concerning the existence of each record in a dataset while allowing statistical data analysis.  $\epsilon$ -Differential privacy requires a randomized function  $\mathcal{A}$ , which is called a privacy mechanism, to generate an output with similar probability distributions for any adjacent datasets  $D$  and  $D'$  as follows.

**Definition 1** ( $\epsilon$ -Differential Privacy [3]). *A randomized function  $\mathcal{A}$  gives  $\epsilon$  differential privacy if for all data sets  $D$  and  $D'$  differing on at most one element, and all  $S \subseteq \text{Range}(\mathcal{A})$ ,*

$$\Pr[\mathcal{A}(D) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{A}(D') \in S] \quad (1)$$

Differential privacy offers robust privacy security, ensuring that an adversary with knowledge of all records except one cannot determine whether the dataset includes the target record.  $\epsilon$ -Differential privacy is overly restrictive in many realistic situations. The introduction of the relaxation parameter  $\delta$  allows for a small probability of failure to meet the criteria.

**Definition 2** (( $\epsilon, \delta$ )-Differential Privacy [5]). *A randomized function  $\mathcal{A}$  gives  $(\epsilon, \delta)$ -differential privacy if for all data sets  $D$  and  $D'$  differing on at most one element, and all  $S \subseteq \text{Range}(\mathcal{A})$ ,*

$$\Pr[\mathcal{A}(D) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{A}(D') \in S] + \delta \quad (2)$$

### 2.3 Differentially Private $k$ -Anonymization Based on Random Sampling

Li et al. [10] introduce the notion of  $(\beta, \epsilon, \delta)$ -differential privacy in Definition 3, which captures the privacy amplification effect of random sampling for an algorithm  $\mathcal{A}$ .

**Definition 3** (*Differential Privacy Under Sampling* [10]). *An algorithm  $\mathcal{A}$  gives  $(\beta, \epsilon, \delta)$ -differential privacy if and only if the algorithm  $\mathcal{A}^\beta$  gives  $(\epsilon, \delta)$ -differential privacy where  $\mathcal{A}^\beta$  denotes the algorithm that first performs Bernoulli sampling with probability  $\beta$  and applies algorithm  $\mathcal{A}$  to the sampled dataset.*

Furthermore, Theorem 6 in [10] instantiates an algorithm  $\mathcal{A}$  for  $k$ -anonymization, and shows that a  $k$ -anonymization algorithm that determines the generalization levels of quasi-identifier variables for full-domain generalization [9] in a differentially private way satisfies  $(\beta, \epsilon, \delta)$ -differential privacy under the constraint on sampling rate  $\beta$  below.

$$\epsilon \geq -\ln(1 - \beta) + \epsilon_{search}. \quad (3)$$

where  $\epsilon_{search}$  is the privacy budget for determining the generalization levels of each quasi-identifier for anonymization. This inequality gives an upper bound of sampling rate  $\beta_{max} = 1 - e^{\epsilon - \epsilon_{search}} = 1 - e^{\epsilon_{anon}}$  where  $\epsilon_{anon}$  is privacy budget for anonymizing sampled data. The relaxation parameter  $\delta$  of  $(\epsilon, \delta)$ -differential privacy is computed by the function  $d$  in Theorem 5 of [10] as follows.

$$d(k, \beta, \epsilon) := \max_{n: n \geq \lceil \frac{k}{\gamma} - 1 \rceil} \sum_{j > \gamma n}^n f(j; n, b) \quad (4)$$

where  $\gamma = \frac{e^\epsilon - 1 + \beta}{e^\epsilon}$ . We denote by  $f(j; n, \beta)$  the probability mass function for the binomial distribution, which gives the probability of getting exactly  $j$  records from an equivalence class of  $n$  records where each record is chosen with probability  $\beta$ . Intuitively, the conditions of  $\epsilon$ -differential privacy are violated when we happen to sample a large number of records where  $j \geq \gamma n$ , making the difference between  $f(j; n, b)$  and  $f(j; n - 1, b)$  noticeable such that  $\frac{f(j; n, b)}{f(j; n - 1, b)} > e^\epsilon$ .

### 2.4 SafePub Algorithm in ARX

ARX anonymization tool [14] includes the *SafePub* algorithm [2], which realizes the Li's sampling scheme for differentially private  $k$ -anonymization. The SafePub algorithm takes as inputs privacy budget  $\epsilon$ , which is divided into two pieces  $\epsilon_{search}$  and  $\epsilon_{anon}$ , and relaxation parameter  $\delta$ . The algorithm uses  $\epsilon_{search}$  to choose a combination of generalization levels for quasi-identifiers with the exponential mechanism [13], and uses  $\epsilon_{anon}$  to conduct random sampling and full-domain generalization for producing differentially private  $k$ -anonymized data, respectively.

Before anonymizing a given dataset  $D$ , the algorithm first sets the sampling rate  $\beta_{max}$  to  $1 - e^{-\epsilon_{anon}}$ , which is the maximum sampling for  $\epsilon_{anon}$ -differential privacy. The algorithm also obtains anonymization parameter  $k$  by computing the following inverse function  $d'$  of  $\delta = d(k, \beta, \epsilon)$  in Theorem 5 of [10].

$$d'(\delta, \beta, \epsilon) := \min\{k \in \mathbb{N} : d(k, \beta, \epsilon) \leq \delta\} \quad (5)$$

Note that the function  $d$  in Eq. (4) computes the upper bound of the probability of failing to satisfy  $\epsilon$ -differential privacy. Since the upper bound of the failure probability computed by the function  $d$  decreases with increasing  $k$ , we use the function  $d'$  to obtain the minimum value for  $k$  that keeps the failure probability below  $\delta$ , aiming to maximize the utility of the anonymized data.

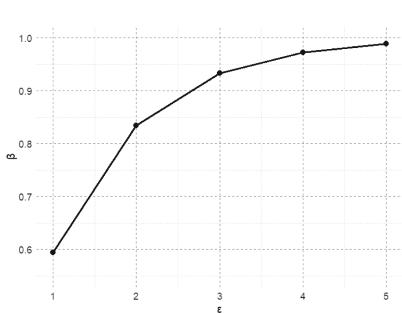
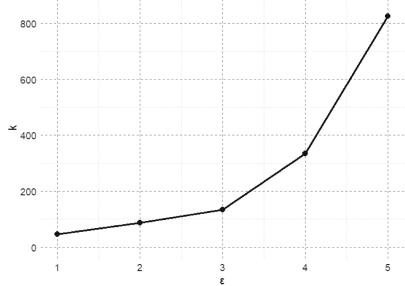
The *SafePub* algorithm first performs Bernoulli sampling to a given dataset  $D$  with the sampling rate  $\beta_{max}$ , which introduces uncertainty regarding the inclusion of each record  $t$  of  $D$  in the anonymized data. The algorithm next generalizes the sampled data with a generalization operator  $g$  for full-domain generalization and suppresses records that appear fewer than  $k$  times. The algorithm iterates every  $g$  in the set  $G$  of all generalization operators to produce a set of candidate anonymized datasets. Finally, the algorithm probabilistically selects one anonymized dataset of high utility using the exponential mechanism based on the utility of anonymized datasets.

## 2.5 Paradoxical Behavior of Deteriorating Utility of Anonymized Data

Since the *SafePub* algorithm uses the maximum sampling rate  $\beta_{max} = 1 - e^{\epsilon_{anon}}$ ,  $\beta_{max}$  converges to 1 as  $\epsilon_{anon}$  increases, as shown in Fig. 1a. Consequently, anonymization parameter  $k$  rapidly increases as  $\epsilon_{anon}$  exceeds 3 to maintain the indistinguishability regarding the existence of each record in the original dataset. Therefore, allocating more privacy budget  $\epsilon_{anon}$  paradoxically leads to a deterioration in the utility of the sampled anonymized data.

Bild et al. [2] analyzed the number of preserved records that are selected through random sampling and are also not suppressed in the following anonymization process. They examine the three cases with low, medium, and high relative generalization levels on the attributes. The results indicate that lowering the sampling rate is the main reason for reducing the number of preserved records. Thus, they make the design choice of using the maximum sampling rate  $\beta_{max}$  under the assumption that information loss of anonymized data is predominantly influenced by the number of records that are not preserved.

However, this assumption does not obviously hold when  $\epsilon_{anon}$  exceeds 3. Our previous research [17] shows that the data utility of anonymized data in terms of the discernibility metric [2] becomes lower as more privacy budget is allocated. Conversely, low-dimensional tables with low  $\beta$  still preserve the utility of anonymized data while adhering to differential privacy. In this paper, we thus explore the possibility of selecting a sampling rate less than  $\beta_{max}$  to improve the utility of anonymized data when privacy budget  $\epsilon_{anon}$  is relatively large.

(a) Sampling rate  $\beta$  on  $\epsilon_{anon}$ (b) Anonymization parameter  $k$  on  $\epsilon_{anon}$ 

**Fig. 1.** The dependencies of sampling rate  $\beta_{max}$  and anonymization parameter  $k$  on privacy budget  $\epsilon_{anon}$

## 2.6 Classification Accuracy

We measure the utility of anonymized data based on the suitability as a training set for statistical classifiers since we evaluate the accuracy of the classifiers trained with anonymized data in Sect. 3. We use the score function proposed by Iyengar [7], which gives penalties on records that do not accommodate the most common combination of characteristics and class attribute values. The score function counts the number of non-penalized records so that a higher score implies higher utility.

For a given dataset  $D$ , anonymization parameter  $k$ , and generalization operator  $g$ , which generalize each record  $r \in D$  into  $r'$ , the function  $S$  computes the anonymized data as follows.

$$S(D) := suppress(g(D), k),$$

where  $g(D) := \bigcup_{r \in D} \{g(r)\}$ . The function *suppress* takes the generalized dataset  $S(D)$  and  $k$  as inputs and produces a dataset  $D'$  in which every record that appears less than  $k$  times is removed from  $D$ .

**Definition 4.** For every anonymization parameter  $k \in \mathbb{N}$ , the score function  $class_k : (\mathcal{D}_m \times \mathcal{G}_m) \rightarrow \mathbb{R}$  is definition as follows:

$$class_k(D, g) := \sum_{r' \in S(D)} w(S(D), r')$$

where  $w(S(D), r')$  is defined as follows.

$$w(S(D), r') := \begin{cases} 1, & \text{if } fv(r') \text{ is not suppressed and} \\ & cv(r') = cv_{maj}(S(D), r') \text{ holds.} \\ 0, & \text{otherwise} \end{cases}$$

We denote by  $fv(r)$  and  $cv(r)$  the values of explanatory variables and that of class variable in record  $r$ , respectively. In statistical classification, the classifier takes the values of explanatory variables in record  $r$  as inputs and predicts the value of the class variable in  $r$ .

### 3 Evaluation

We conduct an experimental evaluation to analyze the utility of anonymized data across a wide range of sampling rates, contrasting with the approach in [2] which utilizes the maximum sampling rate  $\beta_{max}$ .

#### 3.1 ARX Anonymization Tool

We conduct our experiments with the Java implementation of the *SafePub* algorithm [2] in the ARX anonymization tool (ARX) [14]. We assign 90% and 10% of given total privacy budget  $\epsilon$  to  $\epsilon_{anon}$  and  $\epsilon_{search}$ , respectively. As we discuss in Sect. 2.5, the original *SafePub* algorithm in ARX is implemented to choose the maximum sampling rate  $\beta_{max}$  by default. We, therefore, modify the Java source code of the class ParameterCalculation.java to explicitly set the value of the sampling rate between 0.2 to  $\beta_{max}$ .

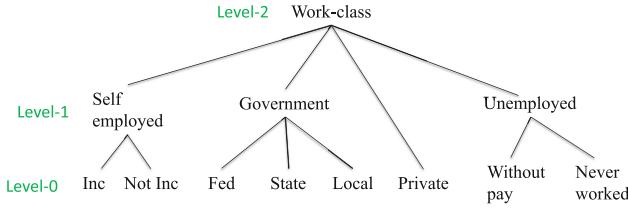
#### 3.2 Dataset and Generalization Hierarchies

We use the *UC Census* dataset from the UCI Machine Learning Repository [1]. The dataset comprises 32,560 records across 15 variables. After excluding records with missing values, we retain 30,162 records for our experiments. We designate the attributes {age, workclass, education, occupation, race, sex, native-country} as quasi-identifiers in this dataset.

We set a generalization hierarchy for each variable. For the *age* variable, we defined seven levels where each level contains age classes of increasing interval size. Specifically, levels 0, 1, and 2 correspond to exact ages, 5-year age groups, and 10-year age groups, respectively. For the variable *sex*, we define categorical values {Male, Female} at level 0 and *Any* at level 1, respectively. Figure 2 displays the generalization hierarchy for the variable *workclass*, sourced from [7]. We also adjusted the highest level of each attribute \* to *Any* for consistency across other attributes. Due to space constraints, we omit the additional generalization hierarchies, which are available on GitHub for reference in the ARX tool [6].

#### 3.3 Utility Analysis of Statistical Classifications

For utility analysis, we focus on the performance of statistical classification using anonymized data generated by the *SafePub* algorithm. ARX offers functionality to solve a statistical classification problem on anonymized data using models like Logistic Regression, Naive Bayes, and Random Forest. We choose the Logistic Regression to assess the prediction accuracy of the model trained with



**Fig. 2.** A generalization hierarchy for variable *workclass* [7].

anonymized data and to investigate the relationship between sampling rate  $\beta$  and the prediction accuracy of the classifier model.

To address multi-nominal classification problems using the Logistic Regression model, ARX employs the “one-vs-all” approach. This method [15] creates  $N$  separate binary classifiers, where each classifier distinguishes between one specific class and all other classes in the dataset. We select two attributes, *marital-status* and *income*, each with eight and two class values, respectively, as the target variables for our prediction model. All other variables in the dataset are anonymized, and we designate them as explanatory variables for the classifier.

We explore the parameter space of privacy budget with three fixed pairs:  $(\epsilon_{anon}, \epsilon_{search}) = (0.9, 0.1), (1.8, 0.2), (2.7, 0.3)$  while varying the sampling rate  $\beta$  from 0.2 to  $\beta_{max}$ . Additionally, we set  $\delta = 0.0001$  uniformly across all pairs. Each  $\epsilon_{anon}$  is associated with its respective upper bound of the sampling rate  $\beta_{max}$ , specifically set as  $(\epsilon_{anon}, \beta_{max}) = (0.9, 0.59), (1.8, 0.83), (2.7, 0.93)$ . We specify the suppression limit, which is the maximum ratio of the allowable suppressed records in the original dataset, as 5%. The hyperparameters for Logistic Regression are configured to default settings in ARX, which can be adjusted across six sections: Alpha, Decay exponent, Lambda, Learning rate, Prior function, and step offset.

We conduct experiments 10 times to measure the accuracy of the classifier and average the results. Figure 3 shows the *relative* accuracies of the classifier for the *marital-status* and *income* attributes respectively. We normalize the values for prediction accuracy such that 0% represents the accuracy of the baseline classifier, which always predicts the most frequent value of the class attribute, while 100% represents that of the classifier trained with the original data. We selected various sampling rates  $\beta$  from 0.2 to  $\beta_{max}$ , with intervals of 0.1, to study the relationship between the sampling rate and the performance of the classifier.

Figure 3a depicts the relative accuracy for the *marital-status* attribute at  $\epsilon = 1, 2, 3$ . We observe a significant improvement in relative accuracy when  $\epsilon = 3$  as we decrease the sampling rate  $\beta$ . For instance, the relative accuracy increases from 63.1% at  $\beta = 0.93$  to 76.7% at  $\beta = 0.2$ , marking a 22% improvement. While the improvement less pronounced at  $\epsilon = 1, 2$  as the case at  $\epsilon = 3$ , we still note enhanced relative accuracy when  $\epsilon = 1$ , decreasing  $\beta$  from 0.59 to 0.3. Although the improvement is not as significant as the case of the *marital-status* attribute,

as shown in Fig. 3b, we observe the tendency of improving the relative accuracy for the *income* attribute at  $\epsilon = 3$ .

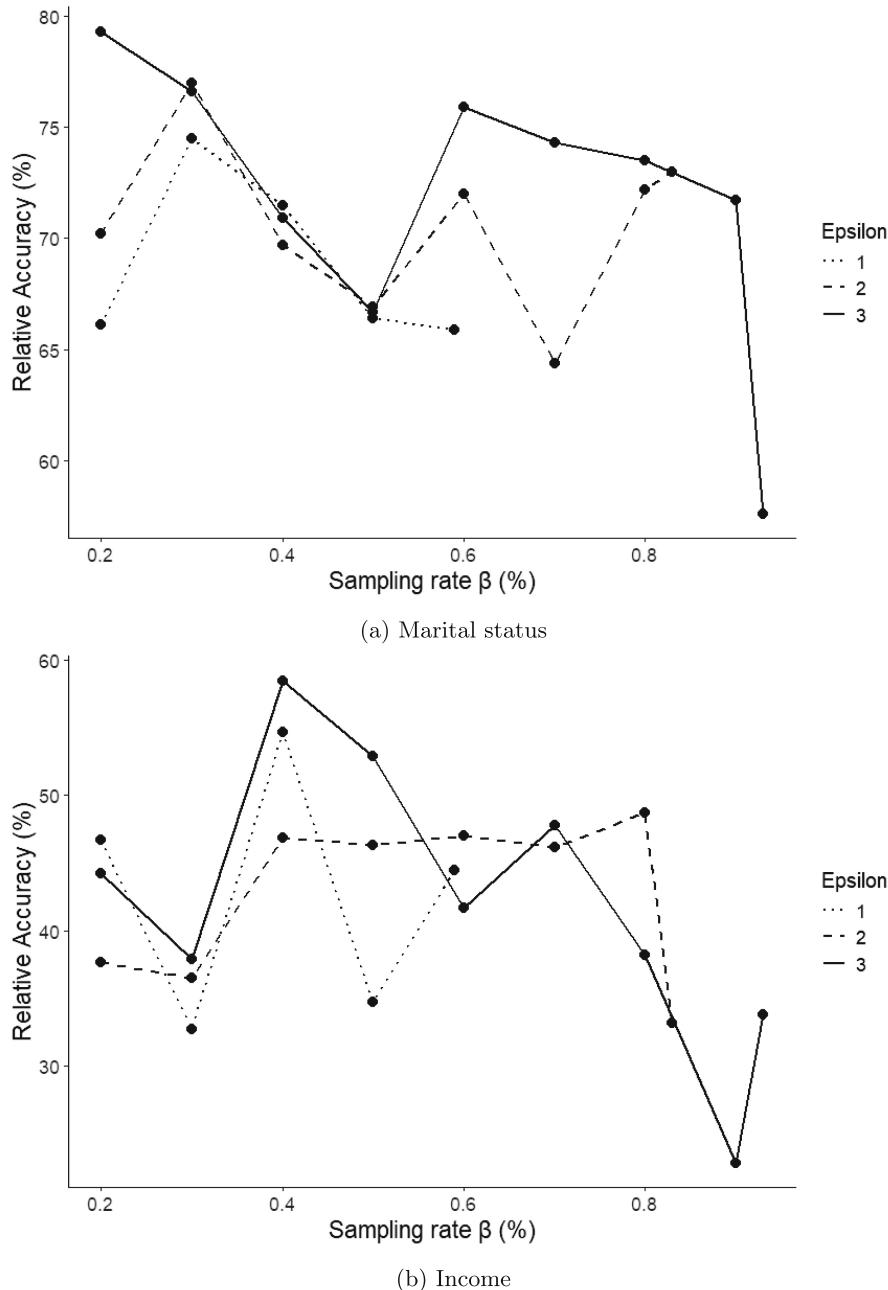
Table 1 shows the characteristics of anonymized data generated by *SafePub* algorithm at different sampling rates  $\beta$ , illustrating the trade-off between sample data size and anonymization parameter  $k$  for each  $\epsilon$ . While we observed no significant improvement in relative accuracy at  $\epsilon = 1$ , substantial improvements were noted for predicting the *marital-status* and *income* attributes at  $\epsilon = 3$ . For example, Table 1c indicates that the number of sampled records decreases from 28,149 to 6,042 as  $\beta$  decreases from 0.93 to 0.2, with the anonymization parameter  $k$  decreasing from 134 to 7 under  $\epsilon = 3$ . Despite the large size of sampled data at  $\beta = 0.93$ , the high value of  $k = 134$  results in large equivalence classes. Therefore, the number of equivalence classes shows minimal changes across the  $\beta$  range of 0.2 to 0.93. When  $\epsilon = 2$ , we observe improvements only for the prediction of *income* attribute. Given the improvement in relative accuracy for both target attributes, as depicted in Fig. 3, we conclude that the utility of anonymized data is primarily influenced by the sizes of equivalence classes rather than the number of unsampled records when the privacy budget  $\epsilon$  is relatively large (greater than 2).

## 4 Related Work

Bild et al. [2] conduct the data utility of anonymized data produced by the *SafePub* algorithm extensively. Claiming that there is not a strong correlation between information loss and the actual usefulness of data, they evaluate the prediction accuracy of a classifier based on a decision tree in the range of privacy budget  $\epsilon$  between 0.1 and 2. As we describe in Sect. 2.5, the *SafePub* algorithm chooses to use the maximum sampling rate because the lower sampling rate dominates information loss of anonymized data when  $\epsilon \geq 1.5$ .

Bild et al. [2] also mention the issue of too much generalization of data when anonymization parameter  $k$  increases as  $\epsilon$  becomes greater, exceeding the range of their experiments. We experimentally confirm this issue of deteriorating data utility when  $\epsilon = 1, 2$ , and 3 in this paper and also show the possibility of improving the utility by adjusting the sampling rate in a more flexible way. Although our experiments in this paper are supplemental to those in [2], we plan to investigate whether the approach of adjusting the sampling rate explicitly also improves the utility of anonymized data in the future.

Sugiyama et al. [17] study the utility of  $k$ -anonymized data produced by the *SafePub* algorithm. They discover that employing the maximum sampling rate  $\beta_{max}$  causes the paradoxical behavior of decreasing data utility when more privacy budget is allocated, as we discuss in Sect. 2.5. However, they do not examine various sampling rates below the maximum sampling rate  $\beta_{max}$ , which requires the modification of the ARX code described in Sect. 3.1. Also, they evaluate the utility of anonymized data based on the discernibility metric, whereas we use the prediction accuracy of the classifiers to directly compare our results with those in [2].



**Fig. 3.** Relationship between the relative accuracy of the Logistic model and sampling rate  $\beta$  at privacy budget  $\epsilon = 1, 2, 3$ .

**Table 1.** Characteristics of anonymized data for different privacy budgets in the dataset.

(a) privacy budget $\epsilon = 1$									
Sampling rate $\beta$ (%)	0.59	0.5	0.4	0.3	0.2				
#Sampled records	17875	15059	12021	9050	6040				
#Equivalence classes	32	28	44	32	34				
#Suppressed records	286.3	167.9	196.1	137.2	86.1				
Anonymization parameter $k$	45	31	22	16	11				
Ratio of suppressed records (%)	1.6	1.1	1.6	1.5	1.4				
Relative accuracy for marital status (%)	65.9	66.4	71.5	74.5	66.1				
Relative accuracy for income (%)	44.5	34.7	54.7	32.7	46.7				

(b) privacy budget $\epsilon = 2$									
Sampling rate $\beta$ (%)	0.83	0.8	0.7	0.6	0.5	0.4	0.3	0.2	
#Sampled records	25184	24135	21104	18153	15121	12104	9029	6021	
#Equivalence classes	27	24	32	43	40	46	50	43	
#Suppressed records	428.7	532.9	274.4	257.4	118.9	166.2	91	98.2	
Anonymization parameter $k$	78	63	34	24	18	14	11	7	
Ratio of suppressed records (%)	1.7	2.2	1.3	1.4	0.8	1.4	1.0	1.6	
Relative accuracy for marital status (%)	73.0	72.2	64.4	72.0	66.9	69.7	77.0	70.2	
Relative accuracy for income (%)	33.2	48.7	46.2	47.0	46.3	46.9	36.5	37.7	

(c) privacy budget $\epsilon = 3$									
Sampling rate $\beta$ (%)	0.93	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
#Sampled records	28149	27131	24134	21081	18044	15055	12062	9059	6042
#Equivalence classes	65	23	37	46	44	55	59	61	47
#Suppressed records	486.5	414.7	289.1	251.3	227.9	245.6	164	107.7	102.1
Anonymization parameter $k$	134	59	43	27	20	15	12	9	7
Ratio of suppressed records (%)	1.9	1.5	1.2	1.2	1.3	1.6	1.4	1.2	1.7
Relative accuracy for marital status (%)	57.6	71.7	73.5	74.3	75.9	66.7	70.9	76.6	79.3
Relative accuracy for income (%)	33.8	22.8	38.2	47.8	41.7	52.9	58.5	37.9	44.2

## 5 Conclusion

In this paper, we study the issue of choosing a sampling rate for the privacy mechanism of creating differentially private  $k$ -anonymized data and conduct experiments to determine how to balance the trade-off between information loss due to sampling and that caused by generalization of variables in the records.

Our preliminary results show the possibility of improving the utility of anonymized data by setting the sampling rate below its upper bound, significantly enhancing the prediction accuracy of the trained model when privacy budget  $\epsilon > 2$ . As future work, we plan to investigate whether the approach of adjusting the sampling rate in a flexible manner is effective when the privacy budget  $\epsilon$  is in the lower range, especially if we apply the proposed method to much larger datasets.

**Acknowledgments.** We would like to express our sincere gratitude to the developers of the ARX anonymization tool. This work was supported by JSPS KAKENHI Grant Numbers supported this work; JP22H00521, JP22K01427, JP21H04403, and by MHLW Program Grant Number JPMH20EA1007.

## References

1. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996). <https://doi.org/10.24432/C5XW20>
2. Bild, R., Kuhn, K.A., Prasser, F.: SafePub: a truthful data anonymization algorithm with strong privacy guarantees. *Proc. Priv. Enhancing Technol.* **2018**(1), 67–87 (2018). <https://doi.org/10.1515/popets-2018-0004>
3. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). [https://doi.org/10.1007/11787006\\_1](https://doi.org/10.1007/11787006_1)
4. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-79228-4\\_1](https://doi.org/10.1007/978-3-540-79228-4_1)
5. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: privacy via distributed noise generation. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 486–503. Springer, Heidelberg (2006). [https://doi.org/10.1007/11761679\\_29](https://doi.org/10.1007/11761679_29)
6. Fabian Prasser: arx-deidentifier/arx. <https://github.com/arx-deidentifier/arx/tree/master/data>
7. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 279–288. Association for Computing Machinery, New York (2002). <https://doi.org/10.1145/775047.775089>
8. Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., Kuhn, K.A.: Flash: efficient, stable and optimal k-anonymity. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 708–717 (2012). <https://doi.org/10.1109/SocialCom-PASSAT.2012.52>
9. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, SIGMOD 2005, pp. 49–60. Association for Computing Machinery, New York (2005). <https://doi.org/10.1145/1066157.1066164>
10. Li, N., Qardaji, W., Su, D.: On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In: Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ASIACCS 2012, pp. 32–33. Association for Computing Machinery, New York (2012). <https://doi.org/10.1145/2414456.2414474>
11. Lin, J.L., Wei, M.C.: An efficient clustering method for k-anonymization. In: Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, PAIS 2008, pp. 46–50. Association for Computing Machinery, New York (2008). <https://doi.org/10.1145/1379287.1379297>
12. Lohr, S.L.: Sampling Design and Analysis. Texts in Statistical Science, 3rd edn. CRC Press, Boca Raton (2021)
13. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), pp. 94–103 (2007). <https://doi.org/10.1109/FOCS.2007.66>
14. Prasser, F., Eicher, J., Spengler, H., Bild, R., Kuhn, K.A.: Flexible data anonymization using ARX-current status and challenges ahead. *Softw. Pract. Exp.* **50**, 1277 – 1304 (2020). <https://api.semanticscholar.org/CorpusID:213656080>
15. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. *J. Mach. Learn. Res.* **5**, 101–141 (2004)

16. Samarati, P.: Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001). <https://doi.org/10.1109/69.971193>
17. Sugiyama, T., Minami, K.: Differentially private frequency tables based on random sampling. In: 2023 IEEE International Conference on Big Data (BigData), pp. 5608–5613 (2023). <https://doi.org/10.1109/BigData59044.2023.10386984>
18. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **10**, 557–570 (2002). <https://api.semanticscholar.org/CorpusID:361794>

# **Microdata Protection**



# Asymptotic Utility of Spectral Anonymization

Katariina Perkonoja<sup>1,2</sup> and Joni Virta<sup>1</sup>

<sup>1</sup> Department of Mathematics and Statistics, University of Turku, Turku, Finland  
`{kakype, jomivi}@utu.fi`

<sup>2</sup> Department of Computing, University of Turku, Turku, Finland

**Abstract.** In the contemporary data landscape characterized by multi-source data collection and third-party sharing, ensuring individual privacy stands as a critical concern. While various anonymization methods exist, their utility preservation and privacy guarantees remain challenging to quantify. In this work, we address this gap by studying the utility and privacy of the spectral anonymization (SA) algorithm, particularly in an asymptotic framework. Unlike conventional anonymization methods that directly modify the original data, SA operates by perturbing the data in a spectral basis and subsequently reverting them to their original basis. Alongside the original version  $\mathcal{P}$ -SA, employing random permutation transformation, we introduce two novel SA variants:  $\mathcal{J}$ -spectral anonymization and  $\mathcal{O}$ -spectral anonymization, which employ sign-change and orthogonal matrix transformations, respectively. We show how well, under some practical assumptions, these SA algorithms preserve the first and second moments of the original data. Our results reveal, in particular, that the asymptotic efficiency of all three SA algorithms in covariance estimation is exactly 50% when compared to the original data. To assess the applicability of these asymptotic results in practice, we conduct a simulation study with finite data and also evaluate the privacy protection offered by these algorithms using distance-based record linkage. Our research reveals that while no method exhibits clear superiority in finite-sample utility,  $\mathcal{O}$ -SA distinguishes itself for its exceptional privacy preservation, never producing identical records, albeit with increased computational complexity. Conversely,  $\mathcal{P}$ -SA emerges as a computationally efficient alternative, demonstrating unmatched efficiency in mean estimation.

**Keywords:** Anonymization · Limiting distribution · Privacy protection · Singular value decomposition · Utility

## 1 Introduction

Various anonymization techniques and privacy protocols, such as  $k$ -anonymity [17],  $l$ -diversity [10],  $t$ -closeness [12], and differential privacy [6], have been proposed to protect data privacy. While these tools are conceptually simple, their deeper mathematical properties are often intractable, meaning that the related

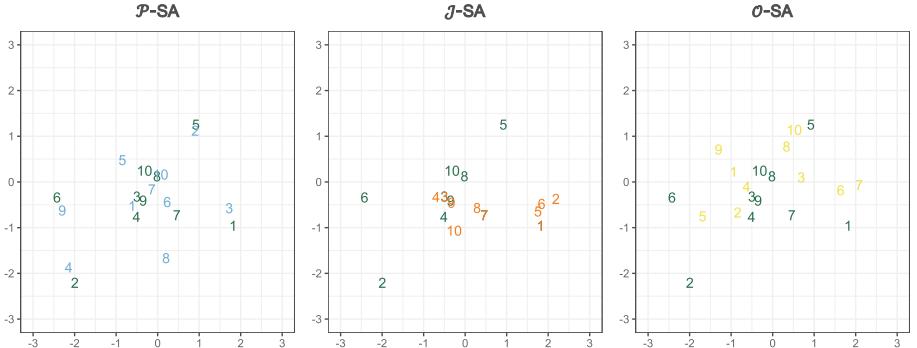
anonymization methods generally lack proven theoretical guarantees, particularly regarding utility preservation, although some attempts have been made, see, e.g., [1, 5, 14, 20]. As a result, decision-makers encounter uncertainty when determining acceptable privacy thresholds for data disclosure, compounded by the trade-off between privacy and utility. Therefore, there exists a pressing need for theoretical results addressing the utility preservation of anonymization techniques, alongside their privacy guarantees.

In this work we study the utility and privacy of the spectral anonymization (SA) algorithm of Lasko and Vinterbo [9] in the asymptotic framework where the sample size of  $n$  of the observed data grows without bounds. Given an observed data matrix  $X_n \in \mathbb{R}^{n \times p}$  with  $n$  observations of  $p$  variables, SA anonymizes it as follows: Letting  $H_n := I_n - (1/n)1_n 1_n'$  be the  $n \times n$  centering matrix, we first find the singular value decomposition (SVD) of the centered data  $H_n X_n = U_n D_n V_n'$ , where the matrix of left singular vectors is  $U_n = (u_{n1} \mid \cdots \mid u_{np}) \in \mathbb{R}^{n \times p}$ . Note that the subscript  $n$  in all quantities serves to remind that our viewpoint is asymptotic (i.e.,  $n \rightarrow \infty$ ) and that we are in fact dealing with sequences of data matrices ( $X_1, X_2, \dots$ ). Let then  $P_{n1}, \dots, P_{np} \in \mathbb{R}^{n \times n}$  be i.i.d. random permutation matrices sampled uniformly from the set  $\mathcal{P}_n$  of all  $n \times n$  permutation matrices. Denoting by  $U_{n0} := (P_{n1}u_{n1} \mid \cdots \mid P_{np}u_{np})$  the matrix of permuted singular vectors, the spectral anonymization of  $X_n$  is

$$X_{n,\mathcal{P}} = U_{n0} D_n V_n' + 1_n \bar{x}_n', \quad (1)$$

where  $\bar{x}_n := (1/n)X_n' 1_n$ . Note that each column of  $U_n$  is permuted independently, meaning that  $U_{n0}$  is not, in general, a simple row permutation of  $U_n$ . Intuitively, SA uses a very simple form of perturbation (permuting observations), which by itself would not guarantee anonymity, but since it is applied in another basis (given by the right singular vectors  $V_n$ ), anonymity in the *original basis* is reached. Technically, any basis could be used, but the *spectral basis* given by the SVD is particularly appealing as the matrix  $U_n$  has uncorrelated columns, implying that, after returning to the original basis, the anonymized data has approximately the same first and second moments as the original data. As argued also by [9], since many standard statistical methods, such as regression, discriminant analysis, support vector machines,  $k$ -means clustering, etc., are based on means and correlations, this preservation of resemblance (moments) can be seen, more or less, as equivalent to the preservation of general data utility.

While permuting is possibly the most direct form of perturbation one can apply in the spectral basis, it is not the only option, and as our first main contribution, we propose two novel forms of SA, called  $\mathcal{J}$ -spectral anonymization and  $\mathcal{O}$ -spectral anonymization. The original SA defined in (1) will in the sequel be denoted as  $\mathcal{P}$ -SA, to distinguish it from these variants. In  $\mathcal{J}$ -SA, the transformation matrices used to perturb  $U_n$  are drawn uniformly from the set  $\mathcal{J}_n$  of all  $n \times n$  sign-change matrices (diagonal matrices with diagonal elements  $\pm 1$ ) and in  $\mathcal{O}$ -SA these matrices are drawn from the Haar uniform distribution on the set  $\mathcal{O}_n$  of all  $n \times n$  orthogonal matrices [16]. The resulting spectral anonymizations of  $X_n$  are denoted by  $X_{n,\mathcal{J}}$  and  $X_{n,\mathcal{O}}$ , respectively. Figure 1 illustrates the effect of the three perturbations in a simple bivariate scenario.



**Fig. 1.** Illustration of the effects of  $\mathcal{P}$ -SA (blue),  $\mathcal{J}$ -SA (orange) and  $\mathcal{O}$ -SA (yellow) to the original data (green/darkest), with numerical labels corresponding to the row indices of original data, in a scenario with  $n = 10$  and  $p = 2$ .  $\mathcal{P}$ -SA and  $\mathcal{J}$ -SA occasionally result in unwanted overlap, wherein a row in the anonymized data matches another row in the original dataset. In the case of  $\mathcal{J}$ -SA, this occurrence is due to coincidental alignment of signs with those in the original singular vectors, causing the observation to be duplicated (overlap of same indices), whereas for  $\mathcal{P}$ -SA, random permutation of values may cause another row to align with one in the original dataset (overlap of different indices). Conversely,  $\mathcal{O}$ -SA induces arbitrary rotations in the spectral space that prevent exact matches altogether with probability 1. (Color figure online)

As our second main contribution, we investigate the earlier observation, stated also by Lasko and Vinterbo [9], that  $\mathcal{P}$ -SA approximately preserves the means, variances, covariances, and linear correlations of the original data  $X_n$ . In this work, we quantify this statement in an asymptotic framework. While SA itself makes no distributional assumptions and is applicable to any continuous data, we derive our results, for simplicity, under the assumption that the rows of  $X_n$  are sampled i.i.d. from the normal distribution  $\mathcal{N}_p(\mu, \Sigma)$  for some fixed  $\mu \in \mathbb{R}^p$  and positive definite covariance matrix  $\Sigma$ . Extensions to other distributional assumptions are investigated through simulations in Sect. 3 and discussed further in Sect. 4. As the multivariate normal distribution is entirely determined by its means and covariances, evaluating the resemblance of anonymized data with respect to these moments is equivalent to assessing general data utility in the present scenario.

As our third main contribution, we conduct a simulation study that investigates how well the three forms of SA preserve utility and privacy under a wide range of data scenarios and sample sizes. We are not aware of an equally extensive benchmarking study having been carried out for the original SA of [9], although [8] compared the prediction utility of SA to that of a private Bayesian factor model. Also, closely related to this, [3] conducted a simulation study assessing utility loss when anonymizing data using another form of spectral anonymization based on factor analysis and principal component analysis.

In Sect. 2 we present our results on the asymptotic utility of the three forms of SA. The corresponding finite-sample behavior is studied using simulations in

Sect. 3.1, whereas finite-sample privacy preservation is investigated in Sect. 3.2. In Sect. 4 we conclude with discussion. The proofs are presented in Appendix A and Appendix B contains additional simulation figures.

## 2 Asymptotic Utility

We divide our theoretical results in two parts, investigating first how well the SA-methods preserve the first moment (mean vector) and then doing the same for the second moment (covariance matrix). Recall from Sect. 1 that we assume the rows of  $X_n$  to be i.i.d. from  $\mathcal{N}_p(\mu, \Sigma)$ . Throughout this section we impose the following technical condition on the covariance matrix  $\Sigma$ , which guarantees that the singular spaces of  $H_n X_n$  are asymptotically identifiable up to sign. This assumption is mild and practically always satisfied for continuous real data.

**Assumption 1.** *The eigenvalues of  $\Sigma$  are distinct.*

The classical central limit theorem gives for  $\bar{x}_n$  (the mean of the original data) the following limiting distribution as  $n \rightarrow \infty$ :

$$\sqrt{n}(\bar{x}_n - \mu) \rightsquigarrow \mathcal{N}_p(0, \Sigma),$$

where  $\rightsquigarrow$  denotes convergence in distribution. As our first result in this section, we derive the limiting distributions of the sample means  $\bar{x}_{n,\mathcal{P}}, \bar{x}_{n,\mathcal{J}}, \bar{x}_{n,\mathcal{O}}$  of the different spectral anonymizations of  $X_n$ .

**Theorem 1.** *Under Assumption 1, we have the following, as  $n \rightarrow \infty$ .*

$$\begin{aligned}\sqrt{n}(\bar{x}_{n,\mathcal{P}} - \mu) &= \sqrt{n}(\bar{x}_n - \mu) \rightsquigarrow \mathcal{N}_p(0, \Sigma) \\ \sqrt{n}(\bar{x}_{n,\mathcal{J}} - \mu) &\rightsquigarrow \mathcal{N}_p(0, 2\Sigma) \\ \sqrt{n}(\bar{x}_{n,\mathcal{O}} - \mu) &\rightsquigarrow \mathcal{N}_p(0, 2\Sigma)\end{aligned}$$

Two estimators of the same parameter can be compared based on the “magnitudes” of their limiting covariance matrices (“smaller” covariance matrix implies more accurate estimator). Thus, Theorem 1 shows that  $\mathcal{P}$ -SA yields mean estimation efficiency equivalent to the original data (in fact, the two estimators are always exactly the same). Whereas  $\mathcal{J}$ -SA and  $\mathcal{O}$ -SA incur a cost in efficiency. Interestingly, this extra cost has a very simple form and the asymptotic efficiency of these two methods compared to mean estimation from the original data turn out to be exactly one half (due to the factor 2 in front of  $\Sigma$ ).

We next conduct an analogous study of covariance matrix estimation. We denote the sample covariance matrices produced by the three forms of SA as  $S_{n,\mathcal{P}}, S_{n,\mathcal{J}}, S_{n,\mathcal{O}}$ . As a baseline, Theorem 1 in [18], shows that the limiting distribution of  $S_n$ , the sample covariance matrix of the original data, is

$$\sqrt{n}\text{vec}(S_n - \Sigma) \rightsquigarrow \mathcal{N}_{p^2}(0, (\Sigma^{1/2} \otimes \Sigma^{1/2})(I_{p^2} + K_{p,p})(\Sigma^{1/2} \otimes \Sigma^{1/2})'),$$

where  $\text{vec}(\cdot)$  denotes column-wise vectorization,  $\otimes$  is the Kronecker product,  $K_{p,p}$  is the  $(p,p)$ -commutation matrix [7] and the square root matrix  $\Sigma^{1/2} := O\Lambda^{1/2}O'$  is computed using the eigendecomposition  $\Sigma = O\Lambda O'$ .

**Theorem 2.** Under Assumption 1, we have for  $\mathcal{A} \in \{\mathcal{P}, \mathcal{J}, \mathcal{O}\}$  the following, as  $n \rightarrow \infty$ .

$$\sqrt{n}\text{vec}(S_{n,\mathcal{A}} - \Sigma) \rightsquigarrow \mathcal{N}_{p^2}(0, (\Sigma^{1/2} \otimes \Sigma^{1/2})(2I_{p^2} + 2K_{p,p} - 2V_p)(\Sigma^{1/2} \otimes \Sigma^{1/2})'),$$

where  $V_p := \text{diag}\{\text{vec}(I_p)\}$ .

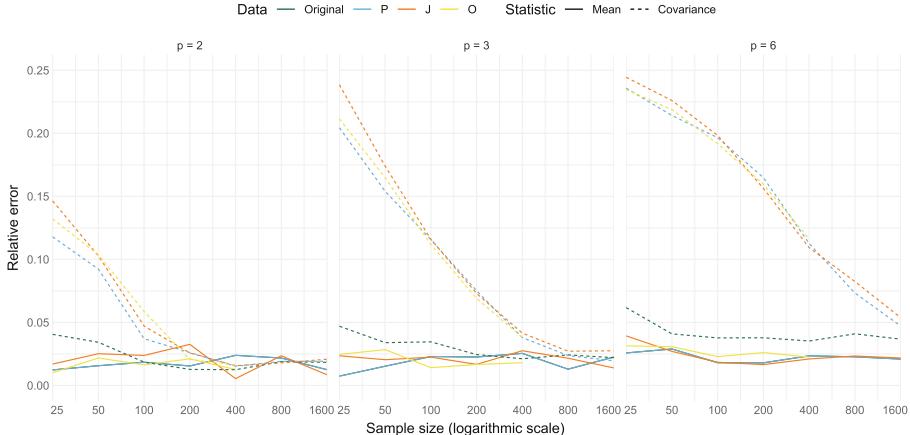
Theorem 2 offers us two insights: (a) All three forms of SA are equally efficient in covariance estimation. (b) Assume next that  $O = I_p$  in the eigendecomposition  $\Sigma = O\Lambda O'$  (since the rotation  $O$  only fixes the coordinate system in which the data is observed, this choice does not limit our interpretations). Then, for instance when  $p = 2$ , the limiting covariance matrices of  $S_n$  and  $S_{n,\mathcal{A}}$  take the forms:

$$\begin{pmatrix} 2\lambda_1^2 & 0 & 0 & 0 \\ 0 & \lambda_1\lambda_2 & \lambda_1\lambda_2 & 0 \\ 0 & \lambda_1\lambda_2 & \lambda_1\lambda_2 & 0 \\ 0 & 0 & 0 & 2\lambda_2^2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 2\lambda_1^2 & 0 & 0 & 0 \\ 0 & 2\lambda_1\lambda_2 & 2\lambda_1\lambda_2 & 0 \\ 0 & 2\lambda_1\lambda_2 & 2\lambda_1\lambda_2 & 0 \\ 0 & 0 & 0 & 2\lambda_2^2 \end{pmatrix}, \quad (2)$$

respectively, where  $\lambda_1, \lambda_2$  are the marginal variances of the data. We observe that the elements (1,1), (4,4), giving the asymptotic variances of the sample variances, are equal in the two matrices, meaning that the estimation of variances is equally efficient in original data and in SA-generated data. Whereas, the elements (2,2), (2,3), (3,2), (3,3) in the matrices (2), corresponding to the asymptotic variances and covariances of the sample covariances are two times larger for SA than for the original data. Thus, we conclude that SA makes the estimation of the cross-terms in the covariance matrix  $\Sigma$  more difficult. This is entirely natural considering that each column of  $U_n$  is in SA mixed independently of each other, compromising our ability to detect interactions (sample covariances) between the columns, but leaving the marginal distribution (sample variances) of each column intact. For arbitrary rotation  $O$ , the equivalent interpretation holds, only this time in the coordinate system specified by  $O$ .

### 3 Simulations

To complement the earlier asymptotic results, in this section we study the finite-sample utility of  $\mathcal{P}$ -SA,  $\mathcal{J}$ -SA and  $\mathcal{O}$ -SA through a simulation study. As discussed in Sect. 2, achieving the exact limiting distribution necessitates an infinite sample size. Therefore, our focus in Sect. 3.1 is to evaluate how good of an approximation the limiting distribution is for finite sample sizes. Rather than directly comparing empirical distributions to the limiting distribution, we examine the similarity between the empirical covariance matrices and their limiting counterparts. Moreover, since  $\mathcal{P}$ -SA,  $\mathcal{J}$ -SA and  $\mathcal{O}$ -SA are designed for data anonymization, we evaluate their capability in protecting privacy in Sect. 3.2 by employing distance-based record linkage. The source code used for conducting the simulation study presented in this work is available on GitLab <https://gitlab.utu.fi/kakype/asymptotic-utility-of-spectral-anonymization.git>.



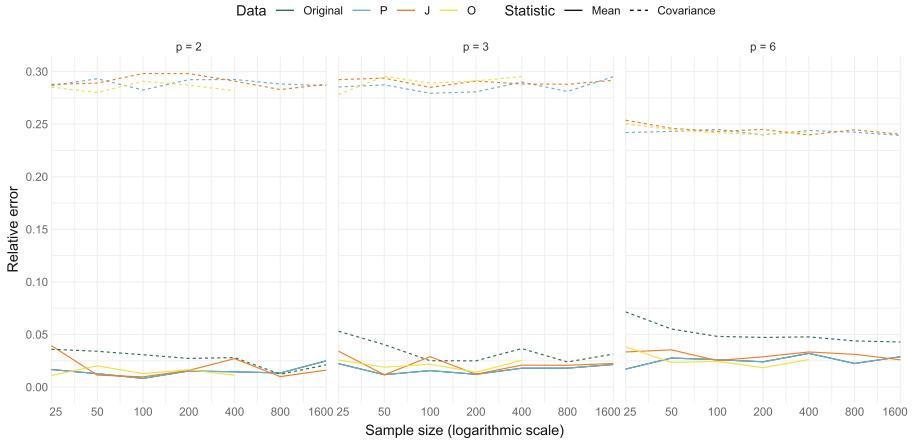
**Fig. 2.** The relative error of the empirical covariance matrices of sample mean (solid line) and sample covariance (dotted line) compared to their asymptotic covariance matrices for  $\mathcal{P}$ -SA (blue),  $\mathcal{J}$ -SA (orange), and  $\mathcal{O}$ -SA (yellow), when original data (green/darkest) is sampled from a normal distribution meeting the Assumption 1. Sample size is presented on a logarithmic scale. (Color figure online)

### 3.1 Finite-Sample Utility

We explore the impact of finite data on convergence, considering sample sizes  $n = \{25, 50, 100, 200, 400, 800, 1600\}$  and number of variables  $p = \{2, 3, 6\}$ , reflecting plausible real-world scenarios. For each combination of  $n$  and  $p$ , we sample original data  $X_{n1}$  from  $\mathcal{N}_p(\mu, \Sigma)$ , where  $\mu = (3, 3, \dots, 3) \in \mathbb{R}^p$  and  $\Sigma$  represents a diagonal matrix with elements  $(p, p-1, \dots, 1)$  to fulfill Assumption 1. Additionally, we explore the consequences of deviating from normality and Assumption 1 by generating three alternative datasets:  $X_{n2}$ ,  $X_{n3}$  and  $X_{n4}$ , whose rows are sampled i.i.d. from  $\mathcal{N}_p(\mu, I_p)$ ,  $Poisson(\lambda_1)$  and  $Poisson(\lambda_2)$ , respectively, where  $\lambda_1 = (p, p-1, \dots, 1)$ ,  $\lambda_2 = (1, 1, \dots, 1) \in \mathbb{R}^p$  and  $Poisson(c)$  denotes a distribution with independent  $Poisson(c_k)$ -distributed marginals. Note that, by the equivariance properties of the mean and covariance matrix, restricting our attention to data with uncorrelated variables is without loss of generality.

Each dataset was sampled  $M = 10000$  times and anonymized using  $\mathcal{P}$ -SA,  $\mathcal{J}$ -SA and  $\mathcal{O}$ -SA, respectively. However, due to the  $O(n^3)$  complexity of  $\mathcal{O}$ -SA, we opted to perform this form of anonymization only for sample sizes  $n \leq 400$ . Subsequently, empirical values were computed according to the left-hand sides of Theorems 1 and 2 and the relative errors (RE) of their covariance matrices across the  $M$  datasets were then calculated. By our theoretical results, these values approach zero as  $n \rightarrow \infty$ . For instance, in the case of sample mean (Theorem 1) and  $\mathcal{P}$ -SA, the calculated value is

$$RE_{\bar{x}_{n,\mathcal{P}}} = \frac{\|\widehat{Cov}_M(\sqrt{n}(\bar{x}_{n,\mathcal{P}} - \mu)) - \Sigma\|_F}{\|\Sigma\|_F},$$



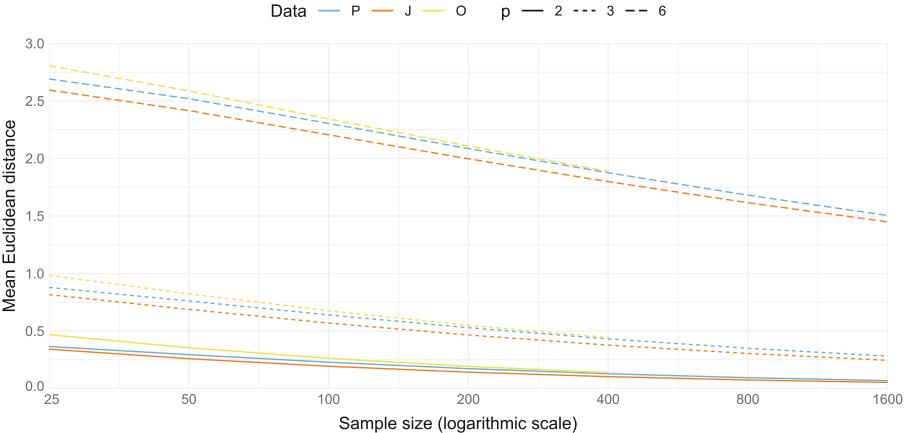
**Fig. 3.** The relative error of the empirical covariance matrices of sample mean (solid line) and sample covariance (dotted line) compared to their asymptotic covariance matrices for  $\mathcal{P}$ -SA (blue),  $\mathcal{J}$ -SA (orange), and  $\mathcal{O}$ -SA (yellow), when original data (green/darkest) is sampled from a normal distribution violating the Assumption 1. Sample size is presented on a logarithmic scale. (Color figure online)

where  $\mu$  and  $\Sigma$  are the mean and covariance parameters derived from the aforementioned data distributions, and  $\|\cdot\|_F$  represents the Frobenius norm. Figures 2 and 3 illustrate the results for normally distributed data, with corresponding representations for Poisson distribution found in Appendix B.

As predicted by Theorems 1 and 2, all curves in Fig. 2 approach zero when  $n$  grows (notably, even with small sample sizes, the convergence for the sample mean is eminent). There are no discernible differences in the convergence among  $\mathcal{P}$ -SA,  $\mathcal{J}$ -SA, and  $\mathcal{O}$ -SA towards their asymptotic covariance matrices. The relative error for the sample covariance depends on  $p$ , higher dimensions requiring a larger sample size for the RE to converge. When Assumption 1 is violated, Fig. 3 indicates that the empirical covariance matrix of the sample covariance matrix converges to a different constant than outlined in Theorem 2. This implies that Assumption 1 is necessary for our results to hold. Similar results were obtained regarding deviation from the normality assumption (Appendix B), where the sample covariance curves, even for the original data, failed to converge to zero. Hence, the results in Sect. 2 are not guaranteed to hold if the normality assumption is discarded, see Sect. 4 for further discussion of this point.

### 3.2 Privacy

In addition to assessing empirical convergence with finite data, we wanted to evaluate the privacy provided by  $\mathcal{P}$ -SA,  $\mathcal{J}$ -SA and  $\mathcal{O}$ -SA. We used distance-based record linkage, akin to the approach outlined in [4], with the distinction that all variables were utilized to compute the shortest Euclidean distance (EUC)



**Fig. 4.**  $\text{EUC}_{\mathcal{A}}$  when the original data is sampled from normal distribution meeting Assumption 1. The y-axis illustrates the mean distance  $\text{EUC}_{\mathcal{A}}$  between records in anonymized data and any record in the original data across all datasets, while the x-axis denotes the sample size on a logarithmic scale. The number of variables  $p$  is distinguished by different linetypes and each anonymization approach is represented by a unique color. (Color figure online)

between an anonymized record and any original record. The mean distance across all  $m = 1, \dots, M$  simulated datasets was then calculated as follows:

$$\text{EUC}_{\mathcal{A}} = \frac{1}{M} \sum_{m=1}^M \frac{1}{n} \sum_{i=1}^n \min_j \|x_{(m),i,\mathcal{A}} - x_{(m),j}\|_2,$$

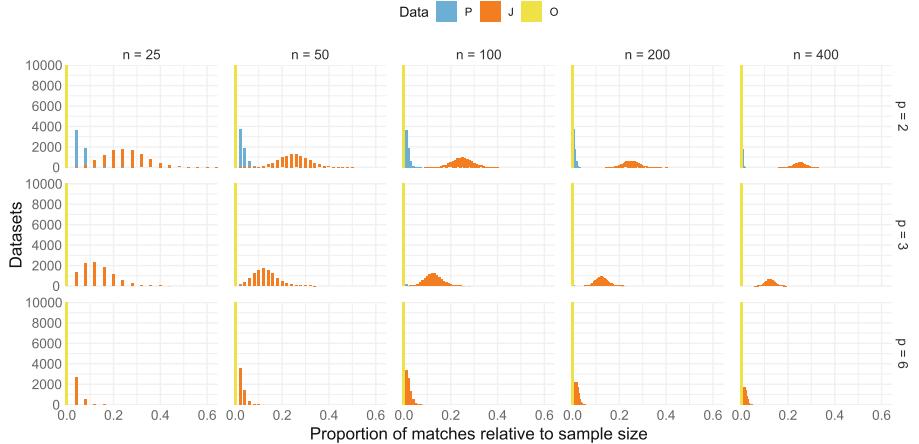
where  $x_{(m),i,\mathcal{A}}$  denotes the  $i$ th observation vector in the  $m$ th anonymized data and  $x_{(m),i}$  denotes the equivalent for the  $m$ th original data. The distances were computed to the unstandardized data, aligning with the common practice of publishing datasets in their original scale. This anticipates potential linkage attempts by adversaries with external data sources.  $\text{EUC}_{\mathcal{A}}$  embodies the privacy-utility trade-off, where values closer to zero indicate better utility to the original data but also imply higher privacy risk, with a distance of zero denoting essentially identical datasets. Consequently, higher values signify stronger privacy protection. Results of this comparison for normally distributed data meeting Assumption 1 are depicted in Fig. 4.

While all SAs demonstrated comparable performance,  $\mathcal{O}$ -SA exhibited the best privacy protection, consistently yielding higher distances compared to  $\mathcal{P}$ -SA and  $\mathcal{J}$ -SA, as illustrated in Fig. 4. These findings remained consistent across all simulation settings (results not shown). However,  $\text{EUC}_{\mathcal{A}}$  is never precisely zero, as this would imply essentially complete equivalence between the anonymized and original data, which is highly unlikely for the given SAs. Yet, these algorithms can produce some matches by chance, as demonstrated in Fig. 1, and for any individual, a direct match would constitute a privacy violation. Therefore,

we supplemented our analysis by evaluating the proportion of matches, i.e., correctly linked records, relative to the sample size across all simulated datasets  $m = 1, \dots, M$ :

$$\text{Matches}_{m,\mathcal{A}} = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1 & \text{if } \min_j \|x_{(m),i,\mathcal{A}} - x_{(m),j}\|_2 < \delta, \\ 0 & \text{otherwise,} \end{cases}$$

where the tolerance  $\delta$  was set to  $10^{-6}$ . Results for normally distributed data meeting Assumption 1 are presented in Fig. 5, with a lower proportion indicating better privacy protection (fewer matches).



**Fig. 5.** Histograms of match proportions when the original data is sampled from normal distribution meeting Assumption 1. The y-axis represents the number of simulated datasets while the x-axis indicates the proportion of matches relative to the sample size, with different SAs represented by distinct colors. Sample sizes  $n > 400$  have been omitted here, as they introduce no new information. (Color figure online)

$\mathcal{O}$ -SA did not produce any matches, as illustrated in Fig. 5, with a proportion of zero observed across all datasets. Conversely, both  $\mathcal{P}$ -SA and  $\mathcal{J}$ -SA approaches resulted in matches, with the former indicating a lower frequency compared to the latter. However, as  $p$  increased, the number of matches decreased, ultimately leading to no matches for  $\mathcal{P}$ -SA and approximately rate of 0.03 for  $\mathcal{J}$ -SA. This trend remained consistent across all simulation scenarios, with the Poisson distribution exhibiting slightly higher match proportion than the normal distribution, and deviations from Assumption 1 having only a marginal effect (results not shown).

## 4 Conclusion

Based on our findings with finite data, none of the three methods demonstrated clear superiority in utility convergence. However, concerning privacy,  $\mathcal{O}$ -SA out-

performed  $\mathcal{P}$ -SA and  $\mathcal{J}$ -SA by never generating identical records. Hence, we advocate for employing  $\mathcal{O}$ -SA for data anonymization, notwithstanding its computational intensity compared to  $\mathcal{P}$ -SA and  $\mathcal{J}$ -SA. For a dataset size of  $n = 1000$  and  $p = 6$ , the runtime for  $\mathcal{O}$ -SA is on average 2.44s, while for  $\mathcal{P}$ -SA and  $\mathcal{J}$ -SA, it is 0.002 and 0.001s, respectively (using AMD Ryzen 5 5600X). The complexity of SVD is the same as for principal component analysis, meaning that the method scales well. In practical scenarios, data anonymization is typically a one-time operation, mitigating concerns over computational overhead, and all SA algorithms can be applied to larger datasets (both in terms of  $n$  and  $p$ ) provided the data is continuous. However, ensuring asymptotic utility (Sect. 2) necessitates additional assumptions. Should computational efficiency become paramount,  $\mathcal{P}$ -SA emerges as the secondary choice (no matches in our simulations with  $p = 6$ ). In case privacy is given secondary priority, then  $\mathcal{P}$ -SA offers superior mean estimation efficiency to its two competitors (Theorem 1) and should be used. Nonetheless, further research is warranted to evaluate these SA algorithms using empirical real-world data and utility scenarios.

Our simulations showed that Assumption 1 cannot be dispensed with. SA is based on perturbing the singular vectors of  $H_n X_n$  and, intuitively, the role of Assumption 1 is to ensure that these vectors are asymptotically identifiable. Similarly, application to Poisson-distributed data revealed that our conclusions do not directly extend outside of normality. However, we expect that equivalent results could be obtained also under other distributional assumptions, with suitable techniques of proof. And while we chose to derive our theoretical results under the normality assumption due to the ubiquitousness of the Gaussian distribution, the SA-method itself remains applicable also outside of normal data.

A prospective research direction would be to conduct analogous studies for competing methods, enabling utility and privacy comparisons between the methods. E.g., data shuffling [11] and correlated noise addition [15] have been shown to preserve second-order statistics. Unlike SA, data shuffling and noise addition use distributional assumptions as part of their data generation and their impact on the methods' utility could be quantified with results such as Theorem 1 and 2. For categorical data, Post-randomization Method (PRAM) leads to unbiased estimates of univariate moments, see, e.g. [15], and also seems applicable to asymptotic analysis. However, PRAM operates on a single variable at a time and more elaborate techniques are required to preserve the dependency structures between multiple categorical variables. Finally, investigating modification constraints, such as selectively altering the first  $j$  columns of  $U_n$ , presents a compelling research area, especially for high-dimensional datasets.

**Acknowledgments.** The study was supported by the Finnish Cultural Foundation (grant 00220801) and the Research Council of Finland (grants 347501 and 353769). The authors would like to thank two anonymous reviewers for their valuable comments.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## A Proofs of Technical Results

*Proof of Theorem 1.* We begin with the  $\mathcal{P}$ -SA. Denoting  $Y_n := X_{n,\mathcal{P}}$ , we have  $\bar{y}_n = \frac{1}{n} Y'_n 1_n = \frac{1}{n} V_n D_n U'_{n0} 1_n + \bar{x}_n$ . Observing now that  $U'_{n0} 1_n = U'_n 1_n$  and that  $V_n D_n U'_{n0} 1_n = 0$ , the desired result follows from the central limit theorem.

For the  $\mathcal{J}$ -SA, we first consider the case where the data come from  $\mathcal{N}_p(0, \Lambda)$  for some diagonal matrix  $\Lambda$  with strictly positive and mutually distinct diagonal elements. We then write

$$\sqrt{n}\bar{y}_n = \frac{1}{\sqrt{n}} Y'_n 1_n = \frac{1}{\sqrt{n}} V_n D_n U'_{n0} 1_n + \sqrt{n}\bar{x}_n. \quad (3)$$

Now, the  $k$ th element of  $D_n U'_{n0} 1_n$  equals

$$d_{nk} u'_{nk} J_{nk} 1_n = e'_k D_n U'_n J_{nk} 1_n = e'_k V'_n X'_n H_n J_{nk} 1_n. \quad (4)$$

We next show that  $X'_n H_n J_{nk} 1_n / \sqrt{n}$  admits a limiting distribution and is, as such, stochastically bounded. To see this, we let the  $i$ th diagonal element of  $J_{nk}$  be  $s_{n,i} \sim \text{Uniform}\{-1, 1\}$  and write  $\frac{1}{\sqrt{n}} X'_n H_n J_{nk} 1_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{n,i} s_{n,i} - \sqrt{n}\bar{x}_n \frac{1}{n} \sum_{i=1}^n s_{n,i}$ . Now, the first term has a limiting distribution (by CLT) and the second term is  $o_p(1)$  (by CLT and Slutsky's lemma). Thus, fixing the sign of  $V_n$  such that  $V_n \rightarrow_p I_p$ , we have, by (4) that  $\frac{1}{\sqrt{n}} d_{nk} u'_{nk} J_{nk} 1_n = \frac{1}{\sqrt{n}} e'_k X'_n H_n J_{nk} 1_n + o_p(1)$ . Using the same in (3), we also obtain  $\sqrt{n}\bar{y}_n = \sqrt{n}\bar{x}_n + \frac{1}{\sqrt{n}} D_n U'_{n0} 1_n + o_p(1)$ . This implies that the  $k$ th element of  $\sqrt{n}\bar{y}_n$  has the expansion  $(\sqrt{n}\bar{y}_n)_k = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{n,ik} (1 + s_{n,i}) + o_p(1)$ . As  $x_{n,ik} (1 + s_{n,i})$ ,  $i = 1, \dots, n$ , are i.i.d. with mean zero and variance  $2\text{Var}(x_{n,ik})$ , and as  $x_{n,ik} (1 + s_{n,i})$  and  $x_{n,i\ell} (1 + s_{n,i})$  are uncorrelated for  $k \neq \ell$ , the limiting distribution of  $\sqrt{n}\bar{y}_n$  is  $\mathcal{N}_p(0, 2A)$ .

Let now  $Z_n = X_n O' + 1_n \mu'$  for some  $p \times p$  orthogonal matrix  $O$  and  $\mu \in \mathbb{R}^p$ . Then,  $Z_{n,\mathcal{P}} = (X_{n,\mathcal{P}} - 1_n \bar{x}'_n) O' + 1_n (O \bar{x}_n + \mu)'$ , thus implying that  $\frac{1}{n} Z'_{n,\mathcal{P}} 1_n = O \frac{1}{n} X'_{n,\mathcal{P}} 1_n - O \bar{x}_n + O \bar{x}_n + \mu$  and  $\sqrt{n} (\frac{1}{n} Z'_{n,\mathcal{P}} 1_n - \mu) = \sqrt{n} O \bar{y}_n \rightsquigarrow \mathcal{N}_p(0, 2O\Lambda O')$ . This completes the proof for  $\mathcal{J}$ -SA.

For  $\mathcal{O}$ -SA, the proof is similar to  $\mathcal{J}$ -SA, and we point out only the differences next. To see that  $X'_n H_n O_{nk} 1_n / \sqrt{n}$  is  $\mathcal{O}_p(1)$ , we observe that  $O_{nk} 1_n \sim \sqrt{n} O_{nk} e_1 \sim \sqrt{n} u_n$ , where  $u_n$  is uniform on the unit sphere  $\mathbb{S}^{n-1}$ . Moreover, letting  $\chi_n$  denote a random variable obeying a  $\chi$ -distribution with  $n$  degrees of freedom that is independent of  $u_n$ , we have  $b_n := \chi_n u_n \sim \mathcal{N}_n(0, I_n)$  and  $\chi_n / \sqrt{n} \rightarrow_p 1$ . Consequently, denoting a generic column of  $X_n$  by  $x_n \sim \mathcal{N}_n(0, \lambda I_n)$ , we have  $(1/\sqrt{n}) x'_n H_n O_{nk} 1_n \sim x'_n H_n u_n = x'_n u_n - \bar{x}_n 1'_n u_n = b_n^{-1} (1/\sqrt{n}) x'_n z_n - \sqrt{n} \bar{x}_n \bar{z}_n b_n^{-1}$ , showing that  $X'_n H_n O_{nk} 1_n / \sqrt{n}$  is stochastically bounded. Hence,

$$(\sqrt{n}\bar{y}_n)_k = (\sqrt{n}\bar{x}_n)_k + b_n^{-1} \frac{1}{\sqrt{n}} e'_k X'_n z_n + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{n,ik} (1 + z_{n,i}) + o_p(1).$$

The limiting distribution of this is  $\mathcal{N}(0, 2\text{Var}(x_{n,ik}))$ , and rest of the proof follows similarly as for  $\mathcal{J}$ -SA.

The following lemma is a direct consequence of Proposition 4.1 in [19].

**Lemma 1.** *Let  $P$  be uniformly random  $n \times n$  permutation matrix. Then,  $E(P) = (1/n)1_n 1_n'$ ,  $E\{\text{tr}(P^2)\} = 2$  and  $E\{\text{tr}(P)\}^2 = 2$ .*

*Proof of Theorem 2.* Starting again with the case of  $\mathcal{N}_p(0, \Lambda)$ -distributed data and using the notation of the proof of Theorem 1, we have

$$\sqrt{n}(S_{n,\mathcal{P}} - \Lambda) = \sqrt{n} \left( \frac{1}{n} Y'_n Y_n - \Lambda - \bar{y}_n \bar{y}'_n \right) = \sqrt{n} \left( \frac{1}{n} Y'_n Y_n - \Lambda \right) + o_p(1),$$

as  $\sqrt{n}\bar{y}_n = \mathcal{O}_p(1)$ . From Theorem 1 we have  $V_n D_n U'_{n0} 1_n \bar{x}_n / \sqrt{n} = o_p(1)$ , giving

$$\begin{aligned} \sqrt{n}(S_{n,\mathcal{P}} - \Lambda) &= \sqrt{n} \left( \frac{1}{n} V_n D_n U'_{n0} U_{n0} D_n V'_n - \Lambda \right) + o_p(1), \\ &= \sqrt{n} \left( \frac{1}{n} X'_n H_n X_n - \Lambda \right) + \frac{1}{\sqrt{n}} V_n D_n G_n D_n V'_n + o_p(1), \quad (5) \\ &= \sqrt{n} \left( \frac{1}{n} X'_n X_n - \Lambda \right) + \frac{1}{\sqrt{n}} V_n D_n G_n D_n V'_n + o_p(1), \end{aligned}$$

where  $G_n \in \mathbb{R}^{p \times p}$  has  $g_{n,kk} = 0$ ,  $g_{n,k\ell} = u'_{nk} T'_{n\ell} T_{n\ell} u_{n\ell}$ , where  $T_{nk}$  are either permutation, sign-change or orthogonal matrices, depending on the type of SA. We next show that, for  $k \neq \ell$ , the quantity  $f_{n,k\ell} := d_{nk} d_{n\ell} g_{n,k\ell} / \sqrt{n}$  is  $\mathcal{O}_p(1)$ .

For  $\mathcal{P}$ -SA, this is seen by writing  $\sqrt{n}f_{n,k\ell} = e'_k D_n U'_n P'_{nk} P_{n\ell} U_n D_n e_\ell = e'_k V'_n X'_n H_n P'_{nk} P_{n\ell} H_n X_n V_n e_\ell = e'_k V'_n (X'_n P'_{nk} P_{n\ell} X_n) V_n e_\ell - n e'_k V'_n (\bar{x}_n \bar{x}'_n) V_n e_\ell$ , giving  $f_{n,k\ell} = n^{-1/2} e'_k V'_n (X'_n P'_{nk} P_{n\ell} X_n) V_n e_\ell + o_p(1) = n^{-1/2} x'_{n,k} P'_{nk} P_{n\ell} x_{n,\ell} + o_p(1)$ , where  $x_{n,k}$  is the  $k$ th column of  $X_n$ . For the second equality, we have used  $\frac{1}{\sqrt{n}} X'_n P'_{nk} P_{n\ell} X_n = \mathcal{O}_p(1)$  which holds for the off-diagonal elements by the independence of  $x_{n,k}$  and  $x_{n,\ell}$ . For the diagonal elements, the boundedness is equivalent to  $x'_n P_n x_n / \sqrt{n} = \mathcal{O}_p(1)$  when  $x_n \sim \mathcal{N}_n(0, I_n)$  and  $P_n$  is a uniformly random  $n \times n$  permutation matrix. Since  $E(x'_n P_n x_n) = 0$ , by [2, Theorem 14.4-1],

$$\frac{1}{\sqrt{n}} x'_n P_n x_n = \frac{1}{\sqrt{n}} \text{SD}(x'_n P_n x_n) \cdot \mathcal{O}_p(1). \quad (6)$$

Now,  $E(x_n x'_n P_n x_n x'_n | P_n) = \text{tr}(P_n) I_n + P_n + P'_n$ , giving, using Lemma 1, that  $\text{Var}(x'_n P_n x_n) = E\{\text{tr}(P_n)\}^2 + E\{\text{tr}(P_n^2)\} + n = n + 4$ . Plugging in to (6), the stochastic boundedness of  $X'_n P'_{nk} P_{n\ell} X_n / \sqrt{n}$  and  $f_{n,k\ell}$  now follows.

For  $\mathcal{J}$ -SA, the reasoning is similar but uses the facts that  $\frac{1}{n} X'_n J_{nk} J_{n\ell} 1_n = o_p(1)$  and  $\frac{1}{n} 1'_n J_{nk} J_{n\ell} 1_n = o_p(1)$  where the former holds because  $J_{n\ell} J_{nk} X_n \sim X_n$ . Furthermore, instead of Lemma 1, we use  $\text{tr}(J_n^2) = n$  and  $E\{\text{tr}(J_n)\}^2 = n$ .

For  $\mathcal{O}$ -SA, we similarly have that  $\frac{1}{n} X'_n O_{nk} O_{n\ell} 1_n = o_p(1)$  and  $\frac{1}{n} 1'_n O_{nk} O_{n\ell} 1_n = o_p(1)$ , since  $O_{n\ell} O_{nk} X_n \sim X_n$  and since, reasoning as in the proof of Theorem 1, we have  $\frac{1}{n} 1'_n O_{nk} O_{n\ell} 1_n = \frac{1}{n} z'_{n1} z_{n2} + o_p(1)$ , where  $z_{n1}, z_{n2} \sim \mathcal{N}_n(0, I_n)$  are independent of each other. Also, instead of Lemma 1, we use for  $\mathcal{O}$ -SA the bounds  $\text{tr}(O_n^2) \leq n$  and  $E\{(P_{n,11} + \dots + P_{n,nn})^2\} \leq$

$\{[E(P_{n,11}^2)]^{1/2} + \dots + [E(P_{n,nn}^2)]^{1/2}\}^2$ , together with  $E(P_{n,ii}) = 0$  and  $E(P_{n,ii}^2) = 1/n$ .

Thus, for every form of SA, by (5), we have  $\sqrt{n}(S_{n,\mathcal{P}} - \Lambda) = \sqrt{n}(\frac{1}{n}X_n'X_n - \Lambda) + \frac{1}{\sqrt{n}}D_nG_nD_n + o_p(1)$ . Writing out the individual elements of  $B_n := \sqrt{n}(S_{n,\mathcal{P}} - \Lambda)$ , the diagonal and off-diagonal thus read

$$b_{n,kk} = \sqrt{n} \left( \frac{1}{n}x_{n,k}'x_{n,k} - \lambda_k \right) + o_p(1), \quad (7)$$

$$b_{n,k\ell} = \frac{1}{\sqrt{n}}x_{n,k}'(I_n + T_{nk}'T_{n\ell})x_{n,\ell} + o_p(1). \quad (8)$$

What remains now is to show that  $b_{n,kk}, b_{n,k\ell}$ ,  $k, \ell = 1, \dots, p$ ,  $k \neq \ell$ , have a limiting joint normal distribution with the desired covariance matrix. We begin by computing the covariance matrix. For  $\mathcal{P}$ -SA, direct computation gives  $\text{Var}(b_{n,kk}) = 2\lambda_k^2$  and  $\text{Var}(b_{n,k\ell}) = 2\lambda_k\lambda_\ell\{1 + (1/n)\text{Etr}(P_{nk}'P_{n\ell})\} = 2(1 + 1/n)\lambda_k\lambda_\ell = 2\lambda_k\lambda_\ell + o(1)$ , where we have used the fact that  $E(P_{nk}) = 1_n 1_n'/n$ . Similarly, we obtain the following non-zero covariances (rest of the covariances are zero):  $\text{Cov}(b_{n,k\ell}, b_{n,\ell k}) = \text{Var}(b_{n,k\ell}) = 2\lambda_k\lambda_\ell + o(1)$ . Analogous computation reveals that the same asymptotic variances and covariances are obtained also for  $\mathcal{J}$ -SA and  $\mathcal{O}$ -SA. Consequently, in each case, the limiting covariance matrix of  $\sqrt{n}\text{vec}(S_{n,\mathcal{P}} - \Lambda)$  is  $(\Lambda^{1/2} \otimes \Lambda^{1/2})[2I_p + 2K_{p,p} - 2\text{diag}\{\text{vec}(I_p)\}](\Lambda^{1/2} \otimes \Lambda^{1/2})$ . The form for a general normal distribution now follows with equivariance arguments as in the proof of Theorem 1. That the moments of the limiting distribution actually are the limits of the moments we computed earlier, is guaranteed by showing that  $b_{n,kk}$  and  $b_{n,k\ell}$  are uniformly integrable. E.g., for  $b_{n,k\ell}$ , this can be done by denoting  $A := (1/\sqrt{n})x_{n,k}'x_{n,\ell}$  and  $B := (1/\sqrt{n})x_{n,k}'T_{nk}'T_{n\ell}x_{n,\ell}$ , using Young's inequality to bound  $E\{(A + B)^4\} \leq aE(A^4) + bE(B^4)$  for some scalars  $a, b \in \mathbb{R}$  and observing that  $T_{nk}x_{n,k} \sim \mathcal{N}_n(0, I_n)$  for each form of SA. The boundedness of  $E(A^4)$  and  $E(B^4)$  now follows from the moments of the Gaussian distribution.

Next, we show that the limiting distribution exists. By the equivariance properties of the multivariate normal, it is sufficient to consider only the case  $\Lambda = I_p$ . We begin with  $\mathcal{O}$ -SA. Letting  $Q_{n1}, \dots, Q_{np}$  be random orthogonal matrices uniform from the Haar distribution (and independent of all our other random variables), we have  $x_{n,k} \sim Q_{nk}x_{n,k}$ . Hence, (7) and (8) can be written as  $b_{n,kk} = \sqrt{n}\{(1/n)x_{n,k}'x_{n,k} - 1\} + o_p(1)$  and  $b_{n,k\ell} = (1/\sqrt{n})x_{n,k}'(Q_{nk}'Q_{n\ell} + R_{nk}'R_{n\ell})x_{n,\ell} + o_p(1)$  where  $R_{nk} := O_{nk}Q_{nk}$  is Haar-uniformly distributed orthogonal matrix and independent of  $Q_{nk}$ . We then decompose  $x_{n,k} = \chi_{nk}u_{nk}$  where  $\chi_{nk} \sim \chi_n$  and  $u_{nk}$  is uniform on the unit sphere and let  $\chi_{nk1}, \chi_{nk2}$  be independent  $\chi_n$ -random variables. Using these we write  $b_{n,kk}$  and  $b_{n,k\ell}$  as  $b_{n,kk} = \sqrt{n}\{(1/n)(\chi_{nk}u_{nk})'(\chi_{nk}u_{nk}) - 1\} + o_p(1)$  and as

$$\begin{aligned} b_{n,k\ell} &= \frac{\chi_{nk}\chi_{n\ell}}{\chi_{nk1}\chi_{n\ell1}} \frac{1}{\sqrt{n}}(\chi_{nk1}Q_{nk}u_{nk})'(\chi_{n\ell1}Q_{n\ell}u_{n\ell}) \\ &\quad + \frac{\chi_{nk}\chi_{n\ell}}{\chi_{nk2}\chi_{n\ell2}} \frac{1}{\sqrt{n}}(\chi_{nk2}R_{nk}u_{nk})'(\chi_{n\ell2}R_{n\ell}u_{n\ell}) + o_p(1). \end{aligned}$$

Now, above  $(\chi_{nk}\chi_{n\ell})/(\chi_{nk1}\chi_{n\ell1}) \rightarrow_p 1$  and  $(\chi_{nk}\chi_{n\ell})/(\chi_{nk2}\chi_{n\ell2}) \rightarrow_p 1$ . Moreover,  $\chi_{nk}u_{nk}$ ,  $\chi_{nk1}Q_{nk}u_{nk}$  and  $\chi_{nk2}R_{nk}u_{nk}$  are independent of each other and all obey the  $\mathcal{N}_n(0, I_n)$ -distribution. Hence, by CLT and Slutsky's lemma, the desired joint limiting normal distribution for  $b_{n,kk}$  and  $b_{n,k\ell}$ ,  $k, \ell = 1, \dots, p$ ,  $k \neq \ell$ , is obtained in the case of  $\mathcal{O}$ -SA.

For  $\mathcal{J}$ -SA, joint limiting normality follows from the multivariate CLT, as  $J_{n1}, \dots, J_{np}$  do not "mix" the observations and retain their i.i.d. nature.

What is thus left is  $\mathcal{P}$ -SA. In that case, we use the Cramér-Wold device combined with the CLT for local dependency neighbourhoods [13, Theorem 3.6]. We demonstrate the proof below in the case  $p = 2$ , the general case following analogously (but with more cluttered notation). That is, the claim holds for  $p = 2$  once we show that  $S_n := a_{11}\sqrt{n}(\frac{1}{n}x'x - 1) + a_{12}\frac{1}{\sqrt{n}}x'(I_n + P)y + a_{22}\sqrt{n}(\frac{1}{n}y'y - 1)$  admits a limiting normal distribution where  $a_{11}, a_{12}, a_{22}$  are arbitrary constants,  $P$  is a uniformly random  $n \times n$  orthogonal matrix and  $x, y \sim \mathcal{N}_n(0, I_n)$  are independent of each other and  $P$ . Our objective is to use [13, Theorem 3.6], where we will take (using their notation)  $X_i = a_{11}(x_i^2 - 1)/\sqrt{n} + a_{12}x_i(y_i + y_{\delta(i)})/\sqrt{n} + a_{22}(y_i^2 - 1)/\sqrt{n}$ , where  $\delta$  is the permutation mapping corresponding to the permutation matrix  $P$ , and  $\sigma^2 = \text{Var}(S_n) = 2a_{11} + 2a_{12}(1 + 1/n) + 2a_{22}$ .

Inspection of the proof of their Theorem 3.6 reveals that the dependency neighbourhood  $N_i$  can be taken to be random (as they are for us), as long as their maximal degree is almost surely some finite  $D$  (as it is for us, since  $x_i(y_i + y_{\delta(i)})$  depends on maximally two other terms), when specific modifications to the proof/result are done: (i) The first term on the RHS of the statement of Theorem 3.6 can be kept as such, but when deriving it, the sums of the form " $\sum_{j \in N_i}$ " have to be replaced with " $\sum_{j=1}^n \mathbb{I}(j \in N_i)$ ", since  $N_i$  are random. (ii) In their formula (3.12), we decompose the term  $\text{Var}(\sum_{i=1}^n \sum_{j \in N_i} X_i X_j)$  as

$$\begin{aligned} & \mathbb{E} \left\{ \text{Var} \left( \sum_{i=1}^n \sum_{j \in N_i} X_i X_j \mid P \right) \right\} + \text{Var} \left\{ \mathbb{E} \left( \sum_{i=1}^n \sum_{j \in N_i} X_i X_j \mid P \right) \right\} \\ & \leq 14D^3 \sum_{i=1}^n \mathbb{E}(X_i^4) + \text{Var}\{\text{Var}(S_n \mid P)\}, \end{aligned}$$

where the inequality uses the variance bound derived in the proof of [13, Theorem 3.6] (for our conditioned-upon dependency neighborhoods) and the second term uses the fact that  $\text{Var}(\sum_{i=1}^n X_i \mid P) = \mathbb{E}(\sum_{i=1}^n \sum_{j \in N_i} X_i X_j \mid P)$ .

Now, as  $D$  is fixed, as  $\sigma^2$  approaches a non-zero constant when  $n \rightarrow \infty$ , and as the  $X_i$  are identically distributed, the desired claim holds once we show that

$$\mathbb{E}|X_1|^3 = o(1/n), \quad \mathbb{E}X_1^4 = o(1/n) \quad \text{and} \quad \text{Var}\{\text{Var}(S_n \mid P)\} = o(1). \quad (9)$$

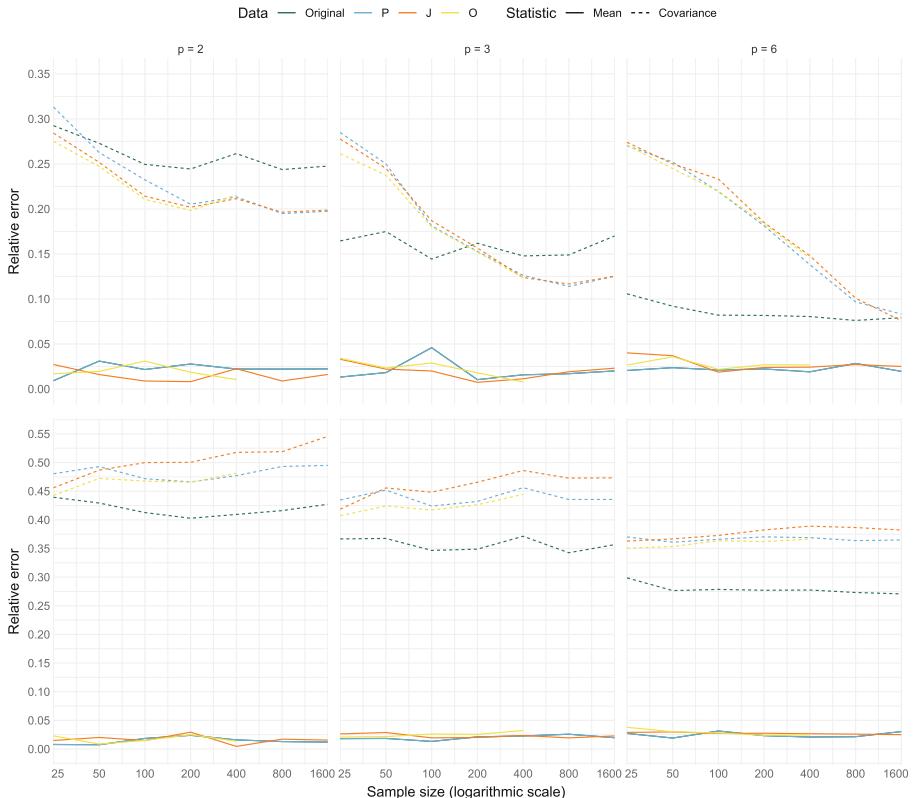
For the third claim in (9), we have  $\text{Var}(S_n \mid P) = 2a_{11} + 2a_{12}\{1 + (1/n)\text{tr}(P)\} + 2a_{22}$ . Hence,

$$\text{Var}\{\text{Var}(S_n \mid P)\} = \frac{4a_{12}^2}{n^2} \text{Var}\{\text{tr}(P)\} \leq \frac{4a_{12}^2}{n^2} \mathbb{E}[\{\text{tr}(P)\}^2] = \frac{8a_{12}^2}{n^2},$$

where we use Lemma 1, taking care of the third claim. Then, as  $X_1$  equals  $1/\sqrt{n}$  times a random variable with all moments finite,  $X_1 = (1/\sqrt{n})\{a_{11}(x_1^2 - 1) + a_{12}x_1(y_1 + y_{\sigma(i)}) + a_{22}(y_1^2 - 1)\}$ , the first two claims in (9) also trivially hold. This concludes the proof in the case  $p = 2$  and the general case is handled analogously.

## B Additional Figures

Figure 6 shows the results of the utility simulation study in Sect. 3.1 for Poisson-distributed data. The top (bottom) row of Fig. 6 is analogous to Fig. 2 (Fig. 3).



**Fig. 6.** Relative error of the empirical covariance matrices of sample mean (solid line) and sample covariance (dotted line) compared to their asymptotic covariance matrices for  $\mathcal{P}$ -SA (blue),  $\mathcal{J}$ -SA (orange), and  $\mathcal{O}$ -SA (yellow). In the top row, the original data (green/darkest) is sampled from a Poisson distribution meeting Assumption 1 while on the second row the assumption is violated. Sample size is on a logarithmic scale. (Color figure online)

## References

1. Awan, J., Kenney, A., Reimherr, M., Slavković, A.: Benefits and pitfalls of the exponential mechanism with applications to Hilbert spaces and functional PCA. In: Proceedings of the 36th International Conference on Machine Learning, vol. 97, pp. 374–384. PMLR (2019)
2. Bishop, Y.M., Fienberg, S.E., Holland, P.W.: Discrete Multivariate Analysis: Theory and Practice. Springer, New York (2007)
3. Calviño, A., Aldeguer, P., Domingo-Ferrer, J.: Factor analysis for anonymization. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 984–991 (2017)
4. Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical data protection. *J. Comput. Appl. Math.* **164–165**, 285–293 (2004)
5. Dunsche, M., Kutta, T., Dette, H.: Multivariate mean comparison under differential privacy. In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 31–45. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_3](https://doi.org/10.1007/978-3-031-13945-1_3)
6. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). [https://doi.org/10.1007/11787006\\_1](https://doi.org/10.1007/11787006_1)
7. Kollo, T., von Rosen, D.: Advanced Multivariate Statistics with Matrices. Springer, Dordrecht (2005)
8. Kundu, S., Suthaharan, S.: Privacy-preserving predictive model using factor analysis for neuroscience applications. In: IEEE International Conference on Big Data Security on Cloud (BigDataSecurity), High Performance and Smart Computing (HPSC) and Intelligent Data and Security (IDS), pp. 67–73. IEEE (2019)
9. Lasko, T.A., Vinterbo, S.A.: Spectral anonymization of data. *IEEE Trans. Knowl. Data Eng.* **22**(3), 437–446 (2009)
10. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity. *ACM Trans. Knowl. Discov. Data* **1**, 3-es (2007)
11. Muralidhar, K., Sarathy, R.: Data shuffling-a new masking approach for numerical data. *Manag. Sci.* **52**(5), 658–670 (2006)
12. Ninghui, L., Tiancheng, L., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the 23rd International Conference on Data Engineering, pp. 106–115 (2007)
13. Ross, N.: Fundamentals of Stein’s method. *Probab. Surv.* **8**, 210–293 (2011)
14. Seeman, J., Reimherr, M., Slavković, A.: Exact privacy guarantees for Markov chain implementations of the exponential mechanism with artificial atoms. In: Advances in Neural Information Processing Systems, vol. 34, pp. 13125–13136. Curran Associates, Inc. (2021)
15. Shlomo, N., De Waal, T.: Protection of micro-data subject to edit constraints against statistical disclosure. *J. Off. Stat.* **24**(2), 229–253 (2008)
16. Stewart, G.W.: The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM J. Numer. Anal.* **17**(3), 403–409 (1980)
17. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**, 557–570 (2002)
18. Tyler, D.E.: Radial estimates and the test for sphericity. *Biometrika* **69**(2), 429–436 (1982)
19. Viana, M.A.: The covariance structure of random permutation matrices. *Contemp. Math.* **287**, 303–326 (2001)
20. Xiao, H., Ye, Y., Devadas, S.: Local differential privacy in decentralized optimization. arXiv preprint (2019)



# Robin Hood: A De-identification Method to Preserve Minority Representation for Disparities Research

James Thomas Brown<sup>1</sup> , Ellen W. Clayton<sup>1</sup> , Michael Matheny<sup>1</sup> , Murat Kantarcioglu<sup>2</sup> , Yevgeniy Vorobeychik<sup>3</sup> , and Bradley A. Malin<sup>1</sup>

<sup>1</sup> Vanderbilt University, Nashville, TN, USA

jthomasbrown1@gmail.com

<sup>2</sup> University of Texas at Dallas, Dallas, TX, USA

<sup>3</sup> Washington University in Saint Louis, Saint Louis, MO, USA

**Abstract.** Data stewards often turn to de-identification to make data available for research while complying with privacy law. A primary challenge to de-identification is balancing the privacy-utility tradeoff, but optimizing the tradeoff with respect to a complete dataset has been shown to create both privacy risk and data utility disparities between subgroups of individuals represented in the dataset. Notably, the minority populations incur the greatest utility loss and privacy risks. Recent studies have shown that utility inequalities can mask disparities and bias algorithms trained on such data. Yet achieving equal privacy and utility is inherently constrained by the fact that each subgroup has a different privacy-utility tradeoff, differences that are exacerbated by the deterministic transformations that standard de-identification models typically employ. To address this problem, we introduce Robin Hood, a de-identification method that leverages non-deterministic transformations to more equally distribute risk and utility in a de-identified dataset. It does so by transforming majority groups' records in a way that gives minorities privacy. We show how Robin Hood can provide equal privacy protections to all records in a dataset at expectation while supporting more accurate and consistent disparity estimation than standard  $k$ -anonymity methods in simulated and real-world Census data.

**Keywords:** De-identification · Anonymization · Fairness

## 1 Introduction

Data stewards frequently turn to the principles of de-identification (in the United States and Canada) and anonymization (in the Europe Union and elsewhere) to make data available for research and innovation while complying with privacy law and regulation. Since data transformations that reduce privacy risk also degrade data utility, one of the primary challenges to implementing such methods is optimizing their inherent privacy-utility tradeoff. However, recent studies have shown that optimizing the privacy-utility tradeoff with respect to the full dataset frequently induces risk and utility disparities

that disadvantage underrepresented populations [1]. Notably, such inequalities occur regardless of the data transformation method, be it generalization and suppression [2], noise addition via differential privacy [3], or synthetic data generation [4].

The inequality, or unfairness, of current de-identification methods can diminish the promise of the data-driven initiatives that employ them. For example, if data utility is not equally distributed across records, disparities could be obscured and subsequently under-addressed [5]; resource allocation could systematically disadvantage the populations with greatest need [6]; and artificial intelligence (AI) models' based on such data could potentially induce or exacerbate discrimination against minorities [7, 8]. If the privacy risk is not equally distributed, more distinguishable populations remain more exposed to privacy intrusions.

An ideal de-identification solution would simultaneously equalize risk and utility among all records; however, beyond the trivial solution of sharing no data, this optimization problem appears to also possess inherent limitations. Specifically, achieving equal utility has thus far required unequal privacy risk and vice versa [3, 5]. This is likely due to minority groups' possessing a less favorable privacy-utility tradeoff than the majority [1], making it difficult - if not impossible - to find a transformation solution that equalizes risk and utility across all groups in the data. In fact, standard de-identification methods exacerbate the inequalities imposed by the differential tradeoffs. In particular, generalization and suppression methods generally apply deterministic transformations [9], such that every record with the same set of quasi-identifying features (e.g., age, gender, and race) is transformed in the same manner by the de-identification algorithm. This convention requires the records of individuals in a minority to be transformed to increase their own privacy protections. At the same time, it does not allow the majority's records, whose greater numbers decrease their initial privacy risk and increase their robustness against utility degradation relative to the minority records, to be transformed in a way that gives privacy protections to the smaller groups. In effect, deterministic transformation strategies require that every group fends for itself.

To support more equitable privacy-preserving data sharing more broadly, in this paper, we introduce a de-identification method that leverages non-deterministic transformations to relax the fairness constraints imposed by unequal privacy-utility tradeoffs. The transformations ambiguates whether a subset of records correspond to the majority or the minority groups, decreasing the distinguishability and subsequently the privacy risk of both. We call this method Robin Hood, as it aims to transform the majority group's records in a way that gives minorities privacy.

The paper is structured as follows. We first describe Robin Hood, its implementation, and formalize the privacy protections it shares between subgroups of records in a de-identified dataset. We then demonstrate Robin Hood's ability to preserve minorities' utility in the context of disparity estimation, given standard de-identification methods' tendency to distort evidence of disparities between majority and minority groups [5]. We compare Robin Hood's performance to that of two standard  $k$ -anonymization implementations [10], using both simulated and real-world Census data.

## 2 Robin Hood De-identification Method

### 2.1 Conceptualization

We begin by defining several concepts to facilitate our description of Robin Hood. We define a **quasi-identifier** as the set of features in the dataset that can enable re-identification and an **equivalence class** as the set of records that share the same quasi-identifying values. We define the **fairness attribute** as the attribute for which the data steward aims to improve the fairness of de-identification transformations (i.e., equalize privacy risk and data utility across subgroups of records) and assume it is part of the quasi-identifier such that Robin Hood applies its transformations to this attribute. Finally, we define a **masking class** as the set of records that share the same set of quasi-identifying features sans the fairness attribute.

Figure 1 illustrates Robin Hood as applied to a single masking class. For comparison, Fig. 1B shows the 4-anonymous version of the same dataset. While every record must be transformed to achieve 4-anonymity in this example, Robin Hood achieves similar privacy protections (derived in Sect. 2.3) by transforming a subset of records in a way that ambiguates to which equivalence class they correspond by suppressing their fairness attribute value. We call this transformation **masking**. Since minority populations frequently fall into smaller equivalence classes than the majority [2], to preserve the utility of minorities, Robin Hood primarily masks records belonging to the larger equivalence class (“M” in Fig. 1) within the masking class. However, we make the conservative assumption that the data recipient knows the original distribution of the dataset prior to Robin Hood’s masking. In such a case, the smaller equivalence class (“m” in Fig. 1) must also contribute at least one record to be masked to receive privacy protections from the set of masked records. Otherwise, the recipient could infer that all masked records derive from the larger equivalence class.

A) Original			B) $k$ -anonymity			C) Robin Hood		
ID	Age	Race <sup>†</sup>	ID	Age	Race <sup>†</sup>	ID	Age	Race <sup>†</sup>
1	21	M	1	21	M or m	1	21	M
2	21	M	2	21	M or m	2	21	•
3	21	M	3	21	M or m	3	21	M
4	21	M	4	21	M or m	4	21	•
5	21	M	5	21	M or m	5	21	•
6	21	M	6	21	M or m	6	21	M
7	21	m	7	21	M or m	7	21	M
8	21	m	8	21	M or m	8	21	•
9	21	m	9	21	M or m	9	21	m
10	21	m	10	21	M or m	10	21	•
11	21	m	11	21	M or m	11	21	m

**Fig. 1.** An example of  $k$ -anonymity (B) and Robin Hood masking (C) applied to an example dataset (A). All records belong to the same masking class. M: majority equivalence class; m: minority equivalence class. <sup>†</sup>Fairness attribute.  $k = 4$ .

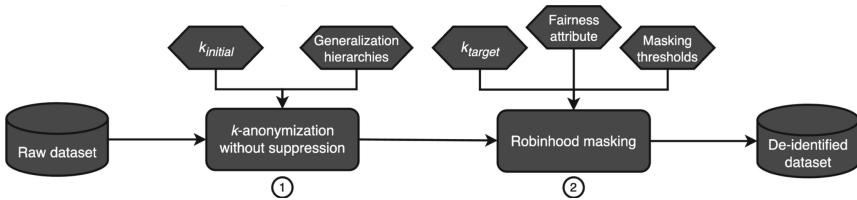
The set of masked records decrease the distinguishability of the equivalence classes that contributed records to be masked. This is because, when targeting a specific individual, the recipient must assume the target could have been masked. The more records that are masked, the larger the set of records that potentially correspond to the target.

This effectively reduces the target individual's privacy risk by increasing the effort to re-identify them [11].

The masked records provide privacy protections to the unmasked records, such that the unmasked records are distorted less (Fig. 1C) than had all records in the equivalence class been transformed together (Fig. 1B). While the best way to analyze such data may vary by application, our disparity estimation evaluation in this paper only uses the unmasked records under the hypothesis that fewer, less distorted records will better represent underlying disparities.

## 2.2 Implementation

Figure 2 describes the Robin Hood implementation pipeline. For the smaller equivalence classes to contribute at least one record to be masked without being masked completely, the first step of the pipeline is to  $k$ -anonymize the raw dataset to a large enough value of  $k$  to support masking. The user specifies this initial  $k$  value,  $k_{initial}$ , as well as the generalization hierarchies that guide  $k$ -anonymization. In this work, we  $k$ -anonymize the raw dataset via global recoding using the Optimal Lattice de-identification (OLA) algorithm without suppression [10]. We do not permit OLA to suppress records in this pipeline to prevent the algorithm from suppressing minority groups' records.



**Fig. 2.** Complete pipeline for Robin Hood de-identification.

The second step of the pipeline applies Robin Hood's masking transformations. Robin Hood increases an adversary's (i.e., a data recipient who attempts re-identification) expected effort to re-identify any record to at least that of a  $k$ -anonymous dataset, where  $k$  is equal to the user-defined  $k_{target}$ . The derivation for the number of records that must be masked to achieve such privacy protections is described in Sect. 2.3. The user also specifies the fairness attribute as well as thresholds that limit the extent to which the majority and minority equivalence classes can be masked. The algorithm defines majority equivalence classes (note, not necessarily the same as the majority subpopulation in the dataset) as those that have at least  $k_{target}$  records in the original table. Minority equivalence classes have less than  $k_{target}$  records. In our experimental evaluation, we set masking thresholds that allow all of the majority equivalence classes' records be masked but only half of the minority equivalence classes' records be masked. The dataset is then masked according to the Robin Hood algorithm, producing the final de-identified dataset. The algorithm's pseudocode is provided in the Appendix.

### 2.3 Privacy Protections of Robin Hood

De-identification methods that leverage generalization and suppression and the privacy models that underlie them typically assume an adversary makes a single re-identification attempt. Based on this assumption, a record's re-identification risk is the inverse of its equivalence class size. However, as shown by Xia et al. [11], a rational adversary (i.e., an adversary that considers both the cost and benefit of attempting re-identification and attacks in a manner that maximizes their payoff) may attack more than one record in an equivalence class to re-identify the target. The authors also showed that an adversary's motivation to attack, and ultimately a record's risk of re-identification, depends on the amount of effort the adversary must exert to re-identify the individual. Following this premise, we show how Robin Hood can increase the effort required to re-identify target individual to that of  $k$ -anonymity for a specified value of  $k$ . However, to compare against standard assumptions, we first derive Robin Hood's privacy protections against a single re-identification attempt.

#### Preliminaries

Table 1 summarizes the notation used to facilitate the derivation and description of the privacy protections.  $A_i$  represents the size of equivalence class  $i$  prior to masking.  $B_i$  and  $C_i$  represent the number of records from  $i$  that are not masked and masked, respectively. Therefore,  $A_i = B_i + C_i$ . The symbol  $D_{ij}$  represents the number of records that are masked within masking class  $j$  that do not belong to  $i$ . Finally,  $E_j$  represents the total number of records masked in  $j$ , such that  $E_j = C_i + D_{ij}$  for any equivalence class  $i$  belonging to masking class  $j$ .

To illustrate, for the majority equivalence class  $M$  in masking class  $j$  shown in Fig. 1:  $A_M = 7$ ,  $B_M = 4$ ,  $C_M = 3$ ,  $D_{Mj} = 1$ , and  $E_j = 4$ . For the minority equivalence class,  $m$ :  $A_m = 3$ ,  $B_m = 2$ ,  $C_m = 1$ , and  $D_{mj} = 3$ .

**Table 1.** Summary of notation used in this paper.

$A_i$	Size of equivalence class $i$ in the original table.
$B_i$	Number of records from $i$ that are not masked.
$C_i$	Number of records from $i$ that are masked.
$D_{ij}$	Number of records masked within masking class $j$ and outside equivalence class $i$ .
$E_j$	Total number of records masked within masking class $j$ .

#### Adversarial Assumptions

We make the following adversarial assumptions. First, we assume the adversary attempts to re-identify a target individual in the dataset. Second, we assume the adversary knows 1) that the target individual's record is in the population, 2) all values of the target individual's quasi-identifier, and 3) the distribution of equivalence classes in the population. Third, we assume that the data steward does not know which individual the adversary is targeting such that similar privacy protections should be applied to every

record in the dataset. Finally, we assume an adversary attempts re-identification following the attack strategy that maximizes their expected rate of success. We model three alternative re-identification strategies the adversary can take: 1) the adversary attacks all unmasked records before attacking the masked records, 2) the adversary attacks all masked records before attacking the unmasked, and 3) the adversary attacks masked and unmasked records with equal probability. As shown in the Appendix, strategy (1) maximizes the adversary's expected rate of successful re-identification and we thus assume the adversary only takes strategy (1).

### Re-identification Risk on the First Attempt

We first define the expected probability that an adversary re-identifies a target individual with a single attempt. Let target individual  $t$  belong to equivalence class  $i$  and masking class  $j$ . The expected re-identification risk against attack strategy (1) is defined in Eq. 1. For comparison against the more traditional notion of  $k$ -anonymity, Eq. 2 displays the expected re-identification risk for individual  $t$  in a  $k$ -anonymous dataset.

$$\begin{aligned} E(reid_t | \text{Robin Hood, Attack strategy 1}) \\ = P(t \text{ is unmasked}) * P(\text{re-id}_t | t \text{ is unmasked, attacks unmasked}) \\ = \left( \frac{B_i}{A_i} \right) * \left( \frac{1}{B_i} \right) = \frac{1}{A_i} \end{aligned} \quad (1)$$

$$E(reid_t | k\text{-anonymity}) \leq \frac{1}{k} \quad (2)$$

Notably, Robin Hood's non-deterministic masking does not reduce the expected probability of re-identification when an adversary makes a single attempt with attack strategy (1) (i.e., they attack an unmasked record). Without any masking, the expected risk would also be  $1/A_i$ , which is equal to that of a  $k$ -anonymous dataset when  $A_i = k$ . However, were the steward to increase the value of  $k$ , such that the new  $k > A_i$ ,  $k$ -anonymity provides greater privacy protections than Robin Hood against a single re-identification attempt.

### Risk Against Multiple Re-identification Attempts

The risk that an adversary re-identifies an individual  $t$  is inversely proportional to the effort required to do so. We define effort as the expected number of attempts until correct re-identification. We use a hypergeometric distribution to model the attack scenario, such that the expected number of attacks until re-identification is defined as  $(N + 1)/2$ , where  $N$  is the size of the pool of records the adversary attacks. The expected number of attempts to re-identify  $t$  in a Robin Hood-transformed dataset against attack strategy (1) is defined in Eq. 3; for a  $k$ -anonymized dataset in Eq. 4.

$$\begin{aligned} E(\text{number of attempts until reid}_t | \text{Robin Hood, Attack strategy 1}) \\ = P(t \text{ is unmasked})E(\#\text{attempts}|t \text{ is unmasked, attacks unmasked}) \\ + P(t \text{ is masked})E(\#\text{attempts}|t \text{ is masked, attacks masked}) \\ = \left( \frac{B_i}{A_i} \right) \left( \frac{B_i + 1}{2} \right) + \left( \frac{C_i}{A_i} \right) \left( B_i + \frac{E_j + 1}{2} \right) \end{aligned} \quad (3)$$

$$E(\text{number of attempts until } \text{reid}_t | k\text{-anonymity}) \geq \frac{k+1}{2} \quad (4)$$

While Robin Hood’s masking does not reduce the expected re-identification risk against a single attempt, it does reduce the risk against an adversary who takes multiple attempts as  $(A_i + 1)/2$  is less than Eq. 3 when  $C_i, D_{ij} \geq 1$ . We next derive the number of records that must be masked in order for Robin Hood to impose the same amount of effort at expectation as  $k$ -anonymity, for a specified value of  $k$ .

Specifically, Eq. 5 shows that the privacy gain of Robin Hood ( $\Delta_i$ ) – equal to the increase in the expected effort required to re-identify a target individual – is proportional to both the number of records masked within  $t$ ’s equivalence class ( $C_i$ ) and the number of records masked outside  $t$ ’s equivalence class but within  $t$ ’s masking class ( $D_{ij}$ ). The latter allows the majority equivalence classes’ records to be masked in a way that contributes privacy protections to the minority equivalence classes.

$$\begin{aligned} \frac{k+1}{2} &= \left( \frac{B_i}{A_i} \right) \left( \frac{B_i + 1}{2} \right) + \left( \frac{C_i}{A_i} \right) \left( B_i + \frac{E_j + 1}{2} \right) \\ k &= A_i + \frac{C_i D_{ij}}{A_i} \\ k - A_i &= \frac{C_i D_{ij}}{A_i} \\ \Delta_i &= \frac{C_i D_{ij}}{A_i} \end{aligned} \quad (5)$$

### 3 Utility Evaluation

#### 3.1 Evaluation Overview

Since de-identification has been shown to mask evidence of racial disparities when it disproportionately degrades racial minorities’ data utility [5], we compare Robin Hood’s ability to preserve such evidence to that of two standard  $k$ -anonymization methods. Both  $k$ -anonymization implementations apply global recoding using the OLA algorithm, where the first allows up to 1% of records to be suppressed and the second does not apply suppression. Both implementations are optimized according to entropy as defined in Eq. S5.

We evaluate performance on two data sources. The first is a controlled simulated dataset, in which we systematically vary the population distribution. The second is the Adult dataset from UC Irvine, a real-world dataset upon which many de-identification methods have been tested [12].

Each simulated dataset contains four attributes: 1) age, 2) sex, 3) race, and 4) a binary outcome. We define the quasi-identifier as  $\{\text{age}, \text{sex}, \text{race}\}$ . Figure 7 displays their corresponding generalization hierarchies. The demographic values for the dataset are selected by randomly sampling with replacement according to a feature-specific probability distribution. The age distribution is based on the Adult dataset. The sex distribution is a uniformly distribution of Female and Male. The race distribution includes three possible

values – majority, minority, super-minority (the smallest group) – and varies between experiments as described below. The binary outcome is assigned to records according to a race-specific rate: {majority: 0.1, minority: 0.2, super-minority:0.5}. Each simulated dataset contains 100,000 records and each experiment is repeated 100 times.

The Adult income dataset is a sample of U.S. Census data containing several demographic and socioeconomic features. We define the quasi-identifier as *{age, gender}* (equivalent to biological sex), *race*, *native-country*, *educational-num*, *workclass*, *marital-status*, *occupation*. We apply the age and gender/sex generalization hierarchies shown in Figure 7. Figure 8 displays hierarchies for the other quasi-identifying attributes, including a different race hierarchy. We define whether an individual has an annual salary above \$50,000 dollars as the binary outcome for disparity estimation. Pre-processing includes dropping duplicate records and records with null values. The final dataset includes 45,175 records: 86% of records correspond to the White race, 9% to Black, 3% to Asian-Pac-Islander, 1% to Amer-Indian-Eskimo, and 0.8% to Other.

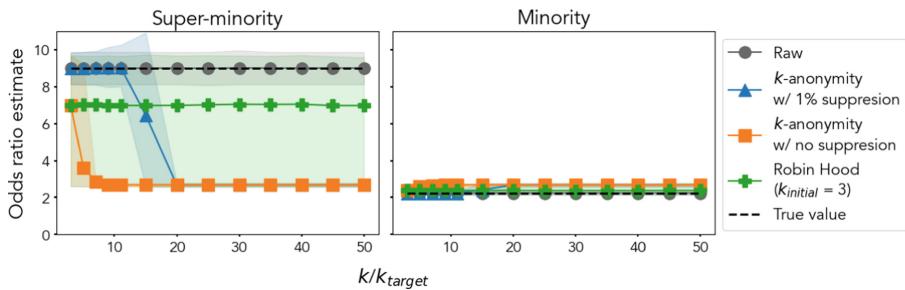
We estimate disparities in the binary outcomes using logistic regression. For the Robin Hood-transformed data, the models are built from the unmasked records. When  $k$ -anonymity is applied with suppression, the models are built from the unsuppressed records. Otherwise, all records are used to build the models. The more accurate the estimated odds ratio, the better we consider the de-identified method to preserve evidence of the differences between groups. For the simulated dataset, the dependent variables include the three quasi-identifying features and the baseline values for categorical variables are white race and male gender. *Age* is treated as a continuous variable, regardless of generalization. For the Adult dataset, the dependent variables include the seven quasi-identifying attributes and the *hours-per-week* variable. The baseline values for categorical variables are white race, male gender, married marital status, native country of United States, private work class, and white-collar occupation. *Age*, *educational-num*, and *hours-per-week* are treated as continuous variables. In scenarios in which race values are generalized into coarser representations (e.g., minority and super-minority combined into “Other”), we assign the same estimated odds ratio to each racial subgroup corresponding to the generalized value. When an odds ratio cannot be estimated (e.g., when the records for a particular racial subgroup are not present in the anonymized dataset) we assign an odds ratio of 1.

### 3.2 Simulated Data

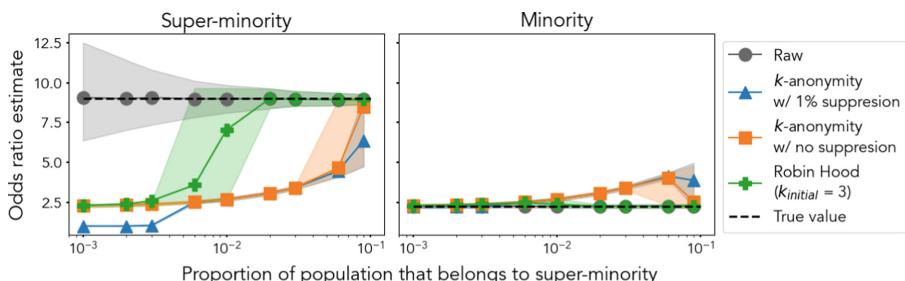
We first evaluate disparity estimation performance when varying the level of  $k/k_{target}$  on a fixed racial distribution: {majority = 0.9, minority = 0.09, super-minority = 0.01}. Figure 3 displays the odds ratio estimates for the minority and super-minority populations.  $k$ -anonymity with suppression supports the most accurate odds ratio estimates up to  $k = 11$ . Afterward, the minority and super-minority race values are generalized to “Other”, and the same odds ratio is estimated for both groups.  $k$ -anonymity without suppression runs into the same problem at  $k = 5$ . The expected odds ratio estimates from the datasets transformed via Robin Hood are less than the true value (derived from the simulation parameters) for the super-minority race and greater than the true value for the minority race. This is because initializing the dataset to  $k_{initial} = 3$  involves generalizing the minority and super-minority race value to “Other” in a fraction of the simulations.

Nonetheless, increasing the number of records masked (i.e., as  $k_{target}$  increases while  $k_{initial}$  remains constant) does not change the odds ratios estimates at expectation. Therefore, on average, Robin Hood supports the most accurate odds ratio estimation at higher  $k$  values.

We repeated the experiment when varying the relative size of the super-minority population and fixing  $k/k_{target}$  to 30. The race probability distribution is defined as:  $\{\text{majority} = 0.9 - x, \text{minority} = 0.09, \text{super-minority} = x\}$ . As shown in Fig. 4, Robin Hood supports the most accurate odds ratio estimates at all proportion values. In fact, Robin Hood's performance equals that of the raw data when the proportion values are  $\geq 0.02$ .  $k$ -anonymity without suppression's performance never equals that of the raw data but improves as the super-minority's proportion increases.  $k$ -anonymization with suppression supports the least accurate disparity estimation performance. When the proportion of the population corresponding to the super-minority group is  $< 0.006$ , this  $k$ -anonymity implementation wholly suppresses the super-minority population's representation in the anonymized dataset such that no disparity can be estimated (denoted by an estimated odds ratio of 1).



**Fig. 3.** Odds ratio estimates for racial disparities in simulated data that has been transformed by different de-identification methods, when varying the level of  $k$ . Race probability distribution is defined as  $\{\text{majority} = 0.9, \text{minority} = 0.09, \text{super-minority} = 0.01\}$ . Expected values are denoted as lines and 95% quantile ranges as shaded areas.

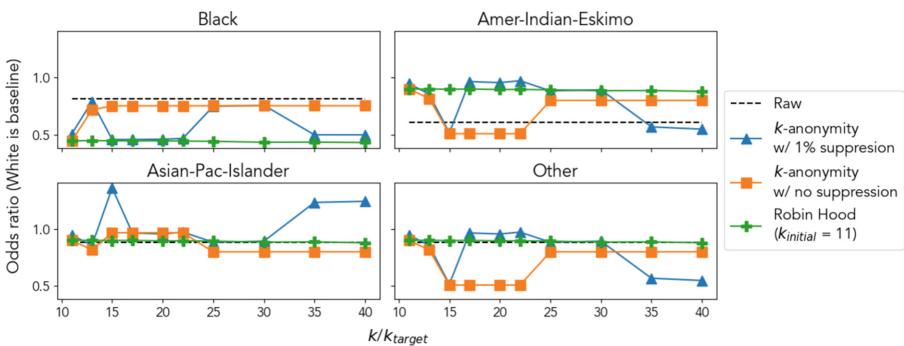


**Fig. 4.** Odds ratio estimates for racial disparities in simulated data transformed by different de-identification methods, when varying proportion of the dataset corresponding to the super-minority race. Expected values are denoted as lines and 95% quantile ranges as shaded areas.

### 3.3 Adult Dataset

The race-specific odds ratio estimates for the Adult dataset, when varying the value of  $k$ , are displayed in Fig. 5. In this case, the true odds ratios for racial disparities’ having a salary above \$50,000 dollars are not known. As such, the results highlight how the odds ratio estimates change when using the raw data vs. the de-identified data.

The  $k$ -anonymization implementations have an inconsistent effect on the odds ratio estimates, particularly at smaller  $k$  values.  $k$ -anonymization with suppression is the least stable, where, at times, it flips the direction of the disparity (e.g., the odds ratio for Asian-Pac-Islander changes from  $<1$  to  $>1$ ).  $k$ -anonymization without suppression does not flip the disparity’s direction, but it does change its estimated magnitude. By contrast, Robin Hood provides consistent odds ratio estimates across  $k$  values, where the accuracy of the estimates relative to the raw data depends more on the generalization applied when initializing to  $k_{initial}$  than by the increasing number of records masked. The initial generalization combines Amer-Indian-Eskimo, Asian-Pac-Islander, and Other to “Not Black or White”. As such, the same odds ratio is estimated for the three original racial subpopulations. Whereas the Amer-Pac-Islander and Other estimates from Robin Hood are very similar to those from the raw data, the Amer-Indian-Eskimo odds ratio from Robin Hood underestimates the disparity compared to the raw data.



**Fig. 5.** Odds ratio estimates for racial disparities, with respect to the binary outcome of having an annual salary greater than \$50,000, in the Adult dataset that has been transformed by different de-identification methods.

## 4 Discussion

Robin Hood is a de-identification method that leverages non-deterministic data transformations to more equally distribute privacy risk and data utility across subgroups of records in a dataset compared to current deterministic methods. Robin Hood’s transformations ambiguate which equivalence class a subset of records originates from, thus increasing the set of records that correspond to a target individual and the expected effort to re-identify them. Robin Hood strategically masks records to raise the expected effort required to re-identify any record to that of a  $k$ -anonymous dataset per a user-defined

value of  $k$ . By this risk measure, Robin Hood sets a fair or equal floor to the privacy protections given to all records in the de-identified dataset. At the same time, the masking transformations allow privacy protections to be shared between equivalence classes, such that larger equivalence classes' records can be transformed in a way that decreases smaller equivalence classes' risk. The cumulative effect of such transformations is a more equal distribution of data utility between subgroups in a dataset. Robin Hood, however, cannot guarantee utility equality. The experiments show the utility distribution is constrained by the initial  $k$ -anonymization, where  $k = k_{initial}$ . Nevertheless, Robin Hood improves utility equality to support more accurate and consistent disparity estimation than deterministic de-identification methods.

Still, Robin Hood's more equal distribution of data utility comes at a price. Unlike  $k$ -anonymity and other group-based de-identification methods, Robin Hood cannot guarantee that every record will be in an equivalence class of a certain size. Rather, it provides privacy protections in expectation, and the risk of records that are ultimately left unmasked is likely to be more than those masked. And though it increases the expected effort of re-identification to deter a rational adversary from attacking in the first place, Robin Hood does not decrease the expected re-identification risk on the adversary's first attempt. These nuances highlight yet another constraint to achieving fairness in de-identified data, as more distinguishable populations are the disadvantaged.

Despite the progress Robin Hood makes toward more equitable de-identification, we wish to highlight several limitations to guide future investigations and improvement. First, Robin Hood's privacy protections depend on deterring a rational adversary from attempting re-identification. While the effectiveness of such deterrence has been shown theoretically [11, 13], and arguably implied by the scant evidence of real-world re-identifications [14, 15], it is still possible an adversary will repeatedly attack the dataset to re-identify a target individual. In such case, more distinguishable records that are not masked could remain susceptible to re-identification. There are two potential solutions to this problem. First, data transformations could be supplemented with sociotechnical deterrents, such as a data use agreement [13]. However, this comes at the cost of data accessibility. Second, a data steward could reapply the Robin Hood algorithm such that a different version of the dataset is shared with each data recipient. This would ensure that the same record does not remain unmasked against every potential adversary, making it less likely the adversary's target individual is unmasked. However, such a strategy would create potential for collusion. Were data recipients to combine several versions of the dataset, they may be able to reverse the masked values. Nevertheless, data sharing frameworks that prevent collusion could mitigate this risk. Future work should evaluate the privacy protections of Robin Hood, with and without additional deterrents and recipient-specific masking methods, against real-world attacks.

Second, we developed Robin Hood to protect against a re-identification attack in which a single individual is targeted. Robin Hood could provide better or worse privacy protections against other re-identification attacks, particularly those in which the adversary attempts to re-identify more than one individual [16]. The fairness of the re-identification protections may also vary against diverse attacks. Furthermore, we did not consider how to incorporate Robin Hood into privacy models that protect against

other types of privacy disclosures, such as  $l$ -diversity [17]. Future work should develop masking methods that consider diverse attack methods and disclosures.

Third, as imputation methods continuously improve, it may be possible that the masked values could be imputed from residual information [18]. It may also be possible for the adversary to reverse-engineer the original dataset with knowledge of the Robin Hood algorithm, the masked dataset, and certain background knowledge. Nevertheless, the principle of transforming records in a way that ambiguates their correspondence to a particular equivalence class could still be achieved with more sophisticated transformation strategies. Future work should investigate how to adapt Robin Hood against imputation and reverse-engineering.

Fourth, the utility evaluation revealed there are more nuances to fairness in de-identification. The simulated data experiments clearly show that, when the disparities correlate with a single attribute, Robin Hood preserves each subgroup's utility in a manner that supports more accurate disparity estimation than  $k$ -anonymity. The Adult dataset experiments are more complex; the true disparities are not known (requiring estimates from the raw data to serve as proxy) and may correlate with multiple attributes. In this real-world dataset,  $k$ -anonymization without suppression supported the most similar disparity estimates to that of the raw data for the Black and Amer-Indian-Eskimo groups. And while Robin Hood supported the most consistent disparity estimates across  $k$  values, suggesting Robin Hood's non-deterministic masking increases certain privacy protections with less utility degradation than  $k$ -anonymity, future work should evaluate Robin Hood's ability to support estimation of more complex disparities.

Fifth, the utility evaluation was limited to disparity estimation via logistic regression. The data recipient may desire to have more representative data for different applications, such as developing fair machine learning models. Where the use case is not defined, the generalizable utility of the data could be estimated using intrinsic utility measures. Future work should evaluate Robin Hood's ability to preserve overall utility and support fair utility in the context of diverse applications and utility measures.

Finally, we only considered fairness with respect to a single attribute. Ideally, de-identified data could fairly distribute privacy risk and utility with respect to several attributes. Future work should investigate how to mask data in a way that supports fairness with respect to multiple attributes, and whether optimizing fairness with respect to one variable sacrifices the fairness with respect to others.

**Acknowledgments.** This research was funded, in part, by grants from the National Institutes of Health (RM1HG009034, T15LM007450, U54HG012510).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## Appendix

### Alternative Re-identification Strategies

The expected probability that an adversary re-identifies target individual  $t$  with a single re-identification attempt against attack strategies (2) and (3) (See in Sect. 2.3) are defined in Eqs. S1 and S2, respectively.

$$\begin{aligned} & E(reid_t | \text{Robin Hood, Attack strategy 2}) \\ &= P(t \text{ is masked}) * P(reid_t | t \text{ is masked, attacks masked}) \\ &= \left( \frac{C_i}{A_i} \right) * \left( \frac{1}{E_j} \right) = \frac{C_i}{A_i E_j} \end{aligned} \quad (6)$$

$$\begin{aligned} & E(reid_t | \text{Robin Hood, Alternative Strategy 3}) \\ &= P(re-id_t | \text{attacks masked or unmasked}) \\ &= \frac{1}{B_i + E_j} \end{aligned} \quad (7)$$

Assuming that the adversary knows the distribution of equivalence classes in the original table, at least one record must be masked both within and outside of the target individual's equivalence class. Therefore,  $C_i \geq 1$  and  $D_{ij} \geq 1$  such that  $C_i + D_{ij} = E_j > C_i$ . It follows that the adversary maximizes their probability of re-identifying  $t$  by prioritizing attacking unmasked records, as defined in Eq. 1.

The expected number of attempts required to re-identify  $t$  in a Robin Hood-transformed dataset against the alternative attack strategies are defined in Eqs. S3 and S4.

$$\begin{aligned} & E(\#\text{attempts until } reid_t | \text{Robin Hood, Attack strategy 2}) \\ &= P(t \text{ is masked})E(\#\text{attempts}|t \text{ is masked, attacks masked}) \\ &\quad + P(t \text{ is unmasked})E(\#\text{attempts}|t \text{ is unmasked, attacks unmasked}) \\ &= \left( \frac{C_i}{A_i} \right) \left( \frac{E_j + 1}{2} \right) + \left( \frac{B_i}{A_i} \right) \left( E_j + \frac{B_i + 1}{2} \right) \end{aligned} \quad (8)$$

$$\begin{aligned} & E(\#\text{attempts until } reid_t | \text{Robin Hood, Attack strategy 3}) \\ &= E(\#\text{attempts} | \text{attacks masked and unmasked}) \\ &= \frac{B_i + E_j + 1}{2} \end{aligned} \quad (9)$$

Now, let  $C_i \geq 1$  and  $E_j \geq 1$ . It follows that Eq. 1 < Eq. S4 < Eq. S3. Finally, it is evident that an adversary minimizes their effort to re-identify  $t$  by attacking all unmasked records before attacking masked records (Table 2).

## Robin Hood Algorithm

**Table 2.** Description of pseudocode functions in Fig. 6.

Function	Description
getMaskingClasses ( $Data'$ , $Attr_{fair}$ )	Partitions $Data'$ into unique masking classes with respect to $Attr_{fair}$
getMajorityEquivalenceClasses ( $Data'$ , $j$ , $k_{target}$ )	Returns the indices of the records corresponding to equivalence classes within masking class $j$ that have at least $k_{target}$ records
getMinorityEquivalenceClasses ( $Data'$ , $j$ , $k_{target}$ )	Returns the indices of the records corresponding to equivalence classes within masking class $j$ that have less than $k_{target}$ records. The indices are partitioned into individual equivalence classes
sizeEquivalenceClass( $i$ )	Returns the size of equivalence class $i$ before masking
alreadyMasked ( $Data^*$ , $j$ , $i$ )	Returns the number of records within masking class $j$ and outside equivalence class $i$ that have already been masked in $Data^*$
numToMask( $A_i$ , $C_i$ , $k_{target}$ )	Using Eq. 5, returns the number of records that must be masked to achieve the privacy protections at $k_{target}$
mask( <i>number to mask, indices of records that can be masked, Data*, Attr<sub>fair</sub></i> )	Randomly chooses, without replacement, which records are masked within the specified indices in $Data^*$ . Records that were previously masked are ignored. Masking involves changing $Attr_{fair}$ values to ‘?’ symbol, or equivalent

Python implementation available at: <https://github.com/j-t-brown/RobinHood>.

## Normalized Entropy

We measure utility loss according to a normalized version of entropy as defined by Gionis and Tassa [10, 19]. Let dataset  $D$  contain  $n$  records  $\{d_1, d_2, \dots, d_n\}$ . Let  $F(d_i)$  denote the size of  $d_i$ 's equivalence class prior to de-identification and  $F(\phi(d_i))$  denote the size of

**Algorithm 1:** Robinhood

---

**Input :**  $Data'$ , Dataset  $k$ -anonymized to  $k_{initial}$ ;  
 $k_{target}$ ,  $k$  value to which masking should protect;  
 $Attr_{fair}$ , fairness attribute.  
 $Threshold_{Majority}$ , maximum proportion of majority equivalence class records that can be masked;  
 $Threshold_{Minority}$ , maximum proportion of minority equivalence class records that can be masked.

**Output:**  $Data^*$ , masked Dataset;  
 $feasible$ , boolean indicating whether masking to  $k_{target}$  was feasible.

---

```

1  $Data^* \leftarrow Data'$ 
2  $J \leftarrow \text{getMaskingClasses}(Data', Attr_{fair})$ 
3 for  $j$  in  $J$  do
4    $I_{majority} \leftarrow \text{getMajorityEquivalenceClasses}(Data', j, k_{target})$ 
5    $I_{minority} \leftarrow \text{getMinorityEquivalenceClasses}(Data', j, k_{target})$ 
6    $N_{majority} \leftarrow \text{numberMajorityRecords}(Data', I_{majority})$ 
7    $feasible \leftarrow True$ 
8   for  $i$  in  $I_{minority}$  do
9      $A_i \leftarrow \text{sizeEquivalenceClass}(i)$ 
10     $C_i \leftarrow 1$ 
11     $D_{masked} \leftarrow \text{alreadyMasked}(Data^*, j, i)$ 
12    while  $C_i \leq (A_i * Threshold_{minority})$  do
13       $D_{ij} \leftarrow \text{numToMask}(A_i, C_i, k_{target}) - D_{masked}$ 
14      if  $D_{ij} \leq (N_{majority} * Threshold_{majority})$  then
15         $\text{mask}(C_i, i, Data^*, Attr_{fair})$ 
16         $\text{mask}(D_{ij}, I_{majority}, Data^*, Attr_{fair})$ 
17        break
18      else
19         $| C_i \leftarrow C_i + 1$ 
20      end if
21    end while
22    if  $C_i > (A_i * Threshold_{minority})$  then
23       $| feasible \leftarrow False$ 
24      return  $Data^*, feasible$ 
25    end if
26  end for
27 end for
28 return  $Data^*, feasible$ 
```

---

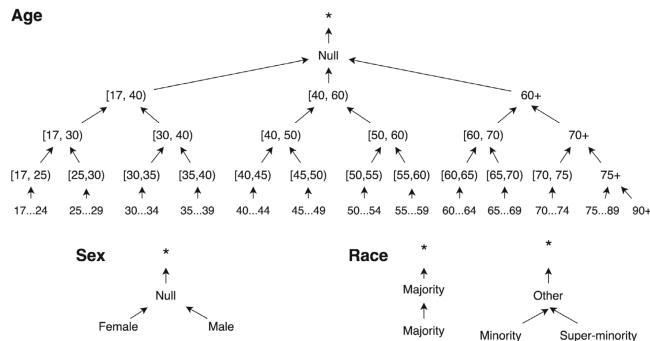
**Fig. 6.** Robin Hood algorithm.

$d_i$ 's equivalence class after de-identification by a deterministic de-identification function  $\phi$ . Then, the utility loss is defined as:

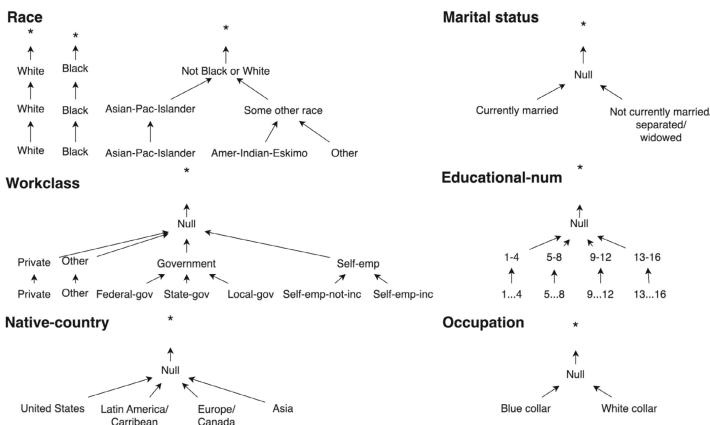
$$\frac{\sum_{d_i \in D} -\log_2 \left( \frac{F(d_i)}{F(\phi(d_i))} \right)}{n} \quad (10)$$

We normalize this measure by the number of records ( $n$ ) to enable a comparison of values between subgroups of different sizes.

## Generalization Hierarchies



**Fig. 7.** Generalization hierarchies for the simulated datasets. The '\*' symbol denotes suppression of the complete record.



**Fig. 8.** The generalization hierarchies guiding de-identification of the Adult dataset include those presented here and the age and sex hierarchies shown in Fig. 7. The '\*' symbol denotes suppression of the complete record.

## References

1. Bowen, C., Snode, J.: Do No Harm Guide: Applying Equity Awareness in Data Privacy Methods. Urban Institute, Washington, DC (2023)
2. Xu, H., Zhang, N.: Privacy in health disparity research. *Med. Care* **57**, S172 (2019). <https://doi.org/10.1097/MLR.0000000000001034>
3. Fioretto, F., Tran, C., Van Hentenryck, P., Zhu, K.: Differential privacy and fairness in decisions and learning tasks: a survey. In: Proceedings of the Thirty-First IEEE International Joint Conference on Artificial Intelligence, July 2022, pp. 5470–5477 (2022). <https://doi.org/10.24963/ijcai.2022/766>

4. Bhanot, K., Qi, M., Erickson, J.S., Guyon, I., Bennett, K.P.: The problem of fairness in synthetic healthcare data. *Entropy* **23**(9), 1165 (2021). <https://doi.org/10.3390/e23091165>
5. Xu, H., Zhang, N.: Implications of data anonymization on the statistical evidence of disparity. *Manag. Sci.* **68**(4), 2600–2618 (2021). <https://doi.org/10.2139/ssrn.3662612>
6. Steed, R., Liu, T., Wu, Z.S., Acquisti, A.: Policy impacts of statistical uncertainty and privacy. *Science* **377**(6609), 928–931 (2022)
7. Cheng, V., Suriyakumar, V.M., Dullerud, N., Joshi, S., Ghassemi, M.: Can you fake it until you make it? Impacts of differentially private synthetic data on downstream classification fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, pp. 149–160. <https://doi.org/10.1145/3442188.3445879>
8. Kenny, C.T., McCartan, C., Kuriwaki, S., Simko, T., Imai, K.: Evaluating bias and noise induced by the US Census Bureau’s privacy protection methods. *Sci. Adv.* **10**(18), eadl2524 (2024)
9. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J. Biomed. Inform.* **50**, 4–19 (2014). <https://doi.org/10.1016/j.jbi.2014.06.002>
10. El Emam, K., et al.: A globally optimal k-anonymity method for the de-identification of health data. *J. Am. Med. Inform. Assoc.* **16**(5), 670–682 (2009). <https://doi.org/10.1197/jamia.M3144>
11. Xia, W., Kantarcioğlu, M., Wan, Z., Heatherly, R., Vorobeychik, Y., Malin, B.: Process-driven data privacy. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, October 2015, pp. 1021–1030 (2015)
12. Becker, B., Kohavi, R.: Adult. UC Irvine (1996) <https://doi.org/10.24432/C5XW20>
13. Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E.W., Kantarcioğlu, M., Malin, B.: Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach. *Am. J. Hum. Genet.* **100**(2), 316–322 (2017). <https://doi.org/10.1016/j.ajhg.2016.12.002>
14. Seastedt, K.P., et al.: Global healthcare fairness: we should be sharing more, not less, data. *PLOS Digital Health* **1**(10), e0000102 (2022). <https://doi.org/10.1371/journal.pdig.0000102>
15. Emam, K.E., Jonker, E., Arbuckle, L., Malin, B.: A Systematic review of re-identification attacks on health data. *PLoS ONE* **6**(12), e28071 (2011). <https://doi.org/10.1371/journal.pone.0028071>
16. Dankar, F.K., El Emam, K.: A method for evaluating marketer re-identification risk. In: Proceedings of the 2010 EDBT/ICDT Workshops, pp. 1–10 (2010)
17. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: 22nd IEEE International Conference on Data Engineering, p. 24 (2006). <https://doi.org/10.1109/ICDE.2006.1>
18. Zou, J., Gichoya, J.W., Ho, D.E., Obermeyer, Z.: Implications of predicting race variables from medical images. *Science* **381**(6654), 149–150 (2023). <https://doi.org/10.1126/science.adh4260>
19. Gionis, A., Tassa, T.: K-Anonymization with minimal loss of information. *IEEE Trans. Knowl. Data Eng.* **21**(2), 206–219 (2009)

# **Statistical Table Protection**



# Secondary Cell Suppression by Gaussian Elimination: An Algorithm Suitable for Handling Issues with Zeros and Singletons

Øyvind Langsrud<sup>(✉)</sup>

Statistics Norway, Postboks 2633 St. Hanshaugen, 0131 Oslo, Norway  
Oyvind.Langsrud@ssb.no

**Abstract.** To protect tabular data through cell suppression, efficient algorithms are essential. Gaussian elimination can be used for secondary cell suppression to prevent exact disclosure. A beneficial feature of this method is that all tables created from the same microdata can be handled simultaneously. This paper presents a solution to the issue where suppressed zeros in frequency tables cannot protect each other. In magnitude tables, it outlines how the algorithm can be tailored to provide protection against singleton contributors using their own data for disclosure.

**Keywords:** Statistical disclosure control · Confidentiality · Tabular data · Cell suppression · Official statistics

## 1 Introduction

Cell suppression is a widely-used statistical disclosure control method for tabular data [3, 4], with several algorithms for secondary suppression implemented in well-known tools [11, 12]. One particular algorithm, Gaussian elimination, is implemented as the function `GaussSuppression` in the R package `SSBtools` [9]. A practical user interface for this function is provided through another R package, also named `GaussSuppression` [8]. The method has previously been described briefly in [10], and an application involving more than 50,000 primary suppressed cells can be found in [7]. The present paper describes secondary suppression through Gaussian elimination more thoroughly, detailing how this approach can resolve specific disclosure issues.

This paper utilizes the examples provided in Sect. 2 to describe the underlying theory throughout, while the main algorithm itself is detailed in Sect. 3.

In frequency tables, attribute disclosure can be countered by suppressing all non-structural zeros. In this scenario, however, suppressed zeros cannot protect each other. Section 4 explores solutions to this issue.

Section 5 addresses the singleton problem in magnitude tables, aiming to prevent all scenarios where individual contributors can use their own data for disclosure. Finally, Sect. 6 presents some concluding remarks.

**Table 1.** The example frequency table with row and column totals. Symbols indicate suppression: • for primary suppression, \* for simple secondary suppression and × for secondary suppression considering zeros.

Age	Iceland	Portugal	Spain	Total
Young	0•	0•	5×	5
Old	6*×	3*×	4×	13
Total	6	3	9	18

**Table 2.** The example frequency table with special totals. Total<sub>1</sub> is the overall total excluding Young~Iceland, and Total<sub>2</sub> is the overall total excluding Young~Portugal. Symbols indicate suppression: • for primary suppression, \* for ordinary secondary suppression and △ for secondary suppression using the conservative subspace method.

Age	Iceland	Portugal	Spain
Young	0•	0•	5
Old	6	3*△	4
Total <sub>1</sub> = 18		Total <sub>2</sub> = 18△	

## 2 Illustrative Examples

The frequency table to be used as an example in this paper is given in Table 1. To avoid disclosing the age of the individuals from Iceland and Portugal, the zeros have been primary suppressed. The standard solution from Gaussian elimination and other secondary suppression algorithms is that two cells need to be secondary suppressed. However, this solution does not provide sufficient protection, as it reveals that all five young individuals are from Spain. Since negative frequencies cannot occur, zeros cannot protect each other. Therefore, Table 1 needs four secondary suppressed cells to ensure adequate protection.

As we will explore further, various algorithms can be employed to achieve this result. The small example in Table 1 is not sophisticated enough to highlight the differences between alternative methods of addressing this issue. Therefore, as shown in Table 2, we have made a special modification to the example.

In Table 3, we present a magnitude table example. The values are intended to represent aggregated sales data from different companies across four specific sectors in three different countries. In this example, five cells require primary suppression. These are singleton cells, i.e., each with a single contributor. Among these five primary suppressed cells, the contributions come from just two distinct companies, identified as A and B.

The two zeros in this table result from there being no contributing company. These are considered known zeros, and therefore, these cells cannot be subjected to secondary suppression. Without taking singletons into account, only two cells need to be secondary suppressed, as marked with \* in Table 3. However, this solution is insufficient. For instance, by examining only the figures from Spain,

**Table 3.** The example magnitude table. Symbols indicate suppression: • denotes primary suppression, used with A or B to signify a singleton contributor ID. \* denotes simple secondary suppression,  $\times$  denotes secondary suppression with singletons handled by virtual primary suppressed cells,  $\circ$  denotes secondary suppression with singletons handled within Gaussian elimination, # denotes secondary suppression with singletons handled within two parallel Gaussian eliminations, and  $\triangle$  denotes secondary suppression with singletons handled within Gaussian elimination considering combinations.

Sector	Iceland	Portugal	Spain	Total
Agriculture	0.0	$3.9^{\circ}\#^{\triangle}$	$122.3^{\circ}\#^{\triangle}$	126.2
Entertainment	$2.7^{\bullet}\text{A}$	$12.2^{\bullet}\text{A}$	$109.1^{\times\triangle}$	124.0
Governmental	0.0	$64.8^{*\times\circ\triangle}$	$15.5^{\bullet}\text{A}$	$80.3^{\#^{\triangle}}$
Industry	$45.6^{*\times\circ\#^{\triangle}}$	$83.9^{\bullet}\text{B}$	$2.3^{\bullet}\text{B}$	$131.8^{\#^{\triangle}}$
Total	48.3	164.8	249.2	462.3

it is clear that A and B can reveal each other's values. This table is designed to illustrate various methods and challenges in preventing the disclosure of sensitive information by singleton contributors. As will be discussed below, the method offering the most effective protection requires a total of seven cells to be secondary suppressed.

### 3 The Main Gaussian Elimination Algorithm

The vector  $z$ , representing all cells in the table(s) to be published, can be expressed as

$$z = X^T y \quad (1)$$

where  $X$  is a matrix consisting of only 0s and 1s, and  $y$  is the vector of all inner cells obtained by crossing all the dimensional variables. Efficient R functionality for creating such  $X$ -matrices in connection with hierarchical tables was recently described in [5].

To obtain protection against exact disclosure, we will ensure that the value of any primary suppressed cell cannot be computed as a linear combination of un suppressed cells. This means that none of the columns of  $X$  representing primary suppressed cells can depend linearly on the columns of  $X$  representing un suppressed cells.

There are many methods for calculations dealing with linear dependence. In this context, limiting memory usage is important by ensuring the matrices involved remain sparse, allowing them to be stored compactly. This is one of the reasons to use Gaussian elimination, which, when expressed as matrix factorization, is known as LU decomposition.

$$\begin{array}{ccccccccc}
& \text{Tot} & \text{Tot} & \text{Tot} & \text{S} & \text{I} & \text{Yng} & \text{Tot} \\
& \text{Tot} & \text{Old} & \text{Old} & \text{I} & \text{Yng} & \text{S} & \text{Old} & \text{P} \\
& \text{Old} & \text{Old} & \text{Tot} & \text{I} & \text{Old} & \text{S} & \text{Old} & \text{P} \\
\text{Yng} & \text{S} & 1 & \cdot & 1 & \cdot & 1 & 1 & \cdot \cdot \cdot \\
\text{Old} & \text{S} & 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 \\
\text{Yng} & \text{I} & 1 & \cdot & \cdot & 1 & \cdot & 1 & \cdot \cdot \cdot \\
\text{Yng} & \text{P} & 1 & \cdot & \cdot & \cdot & 1 & \cdot & 1 \\
\text{Old} & \text{I} & 1 & 1 & \cdot & 1 & \cdot & \cdot & \cdot \\
\text{Old} & \text{P} & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1
\end{array}
\rightarrow
\begin{array}{ccccccccc}
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot & -1 & -1 & 1 & \cdot \\
\cdot & \cdot & -1 & 1 & \cdot & \cdot & -1 & \cdot & \cdot \\
\cdot & \cdot & -1 & \cdot & \cdot & \cdot & -1 & \cdot & 1 \\
\cdot & 1 & -1 & 1 & 1 & -1 & -1 & \cdot & \cdot \\
\cdot & 1 & -1 & \cdot & \cdot & -1 & -1 & 1 & 1
\end{array}
\right]$$
  

$$\stackrel{2}{\rightarrow}
\begin{array}{ccccccccc}
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot & -1 & 1 & \cdot & -1 & \cdot & \cdot & 1 \\
\cdot & \cdot & -1 & \cdot & \cdot & -1 & \cdot & 1 & \cdot \\
\cdot & \cdot & -1 & 1 & 1 & \cdot & -1 & \cdot & \cdot \\
\cdot & \cdot & -1 & \cdot & \cdot & \cdot & -1 & 1 & 1
\end{array}
\stackrel{3}{\rightarrow}
\begin{array}{ccccccccc}
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & 1 & \cdot \\
\cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & -1 \\
\cdot & \cdot & \cdot & \cdot & 1 & 1 & -1 & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & -1 & \cdot & 1 & -1 & 1 \\
\cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & 1 & -1
\end{array}$$
  

$$\stackrel{4}{\rightarrow}
\begin{array}{ccccccccc}
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & \cdot & 1 & -1 & \cdot & -1 \\
\cdot & \cdot & \cdot & \cdot & 1 & -1 & \cdot & 1 & \cdot
\end{array}
\stackrel{5}{\rightarrow}
\begin{array}{ccccccccc}
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & 1 & 1 \\
\cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & -1
\end{array}$$

**Fig. 1.** Gaussian elimination steps leading to simple secondary suppression marked as \* in Table 1. The zeros are shown as dots. The columns for primary suppressed cells are to the right of the vertical line. Compared to Table 1, the categories for age and country have been shortened to three and one letter respectively. The steps are: Col 1 subtracted from cols 3, 6, 7; Col 2 added to cols 6, 7, and subtracted from col 8; Col 3 added to cols 4, 11, and subtracted from col 7; Col 4 added to cols 9, 12, and subtracted from col 11; Col 5 added to cols 8, 11, and subtracted from col 7.

As described in the documentation for the Matrix package in R [1], the LU decomposition of  $X$  is generally expressed as

$$P_1 X P_2 = LU \quad (2)$$

where  $P_1$  and  $P_2$  are permutation matrices for row and column interchanges, respectively. Here,  $L$  and  $U$  are lower and upper trapezoidal matrices, respectively. To be precise, we have used the term trapezoidal since triangular is limited to square matrices. To preserve the triangular structure in cases with linearly dependent columns,  $P_2$  should be defined such that the first  $\text{rank}(X)$  columns are linearly independent.

Here, we are not primarily interested in the complete decomposition of  $X$ , but rather in using the elimination process to achieve secondary suppression. The choice of a specific order for the columns ( $P_2$ ) is essential, and this must

be determined based on the utility of publishing each cell. The last columns will represent the primary suppressed cells. This sequential elimination algorithm is much more straightforward than seeking an optimal solution aimed at minimizing information loss from the suppressed cells. The latter is equivalent to maximizing the total utility for all the unsuppressed cells. Although the algorithm is not optimal, the prioritization order ensures that information loss is limited. Therefore, in addition to the primary suppressed cells, the priority order is a crucial input to the algorithm. In Fig. 1, we have illustrated this process in the case of simple secondary suppression of the frequencies in Table 1. The first matrix in this figure represents a version of  $X$  where the columns and rows are sorted in a specific order. The columns are arranged by decreasing frequency values. In cases of equal frequency values, marginal cells are prioritized over inner cells.

From left to right, each column is sequentially used as the starting point for elimination. The row selected for elimination is chosen as high up as possible. Each column with an element in this row is changed by subtracting a multiple of the starting column, where the multiple itself can be negative. Besides the column that initiated the process, it is also possible for other columns to be eliminated due to linear dependence. As shown in the figure, column 6 was eliminated during the second round of elimination, a result of the fact that  $\text{Young} \sim \text{Total}$  can be computed from  $\text{Total} \sim \text{Total}$  and  $\text{Old} \sim \text{Total}$ . In this process, we will follow this rule: *Columns representing primary suppressed cells cannot be completely eliminated*. Therefore, we have to skip columns that cannot be used as starting points for elimination. The cells corresponding to these skipped columns become secondary suppressed. In the fifth elimination, we had to skip  $\text{Old} \sim \text{Iceland}$  to prevent the elimination of  $\text{Young} \sim \text{Iceland}$ . The final matrix in Fig. 1 shows two secondary suppressed cells as the result. The deviation from the first matrix can be expressed in LU form as

$$\begin{bmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & 1 & \cdot & \cdot & \cdot \\ 1 & \cdot & 1 & 1 & \cdot & \cdot \\ 1 & 1 & 1 & \cdot & 1 & \cdot \\ 1 & 1 & 1 & 1 & 1 & \end{bmatrix} \times \begin{bmatrix} 1 & \cdot & 1 & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & -1 & -1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & 1 & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & 1 & \cdot & -1 & 1 \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & -1 & \cdot & \cdot & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & -1 & \cdot & \cdot \end{bmatrix} \quad (3)$$

In this case, the row order in Fig. 1 was chosen to achieve a triangular structure.

In this example, one row in the  $X$ -matrix (Fig. 1) was required for each of the six inner cells (Table 1). When it comes to structural zeros, which are never involved in suppression, their corresponding rows can be omitted. Below, in Fig. 4, it therefore holds with ten rows in the  $X$ -matrix even though there are 12 inner cells in Table 3. In general, we have a nice way to distinguish between non-structural and structural zeros. Rows for the former are included in the  $X$ -matrix, but not for the latter.

In the small examples in this paper, the matrices within the elimination process only consist of the numbers -1, 0, and 1. Generally, it will be other numbers. In order to keep integers so that exact calculation is ensured, the columns can

be scaled along the way. This is implemented in the `GaussSuppression` function in `SSBtools`. In the text below, we describe the columns as being made up only of the rows with nonzero elements. This description corresponds to the implementation with sparse column vectors.

## 4 Handling Zeros in Frequency Tables

### 4.1 Virtual Cells of Zeros

One way to ensure that the zeros in Table 1 cannot be revealed is by creating a virtual primary suppressed cell that sums the two zero-frequency cells. In Fig. 2, we have introduced such a cell into the elimination process. In this instance, we also changed the row order. This change does not affect the result here, but we will demonstrate in the subsections below how it can be useful. The final matrix in Fig. 2 shows four secondary suppressed cells, as shown in Table 1. In more advanced tables, this method can be challenging because, to be sure, all possible combinations of zeros in  $y$  must be taken into account.

$$\begin{array}{c}
 \begin{array}{ccccccccc}
 & & \text{Tot} & \text{Tot} & & & & & \\
 & & \text{Old} & \text{Tot} & \text{S} & \text{Tot} & \text{I} & \text{Yng} & \text{Tot} \\
 & & \text{Old} & \text{Old} & \text{Tot} & \text{Old} & \text{I} & \text{Yng} & \text{S} \\
 & & & & \text{Old} & & & \text{Old} & \text{S} \\
 & & & & & \text{Old} & & \text{Tot} & \text{P} \\
 & & & & & & \text{Old} & \text{Old} & \text{P} \\
 & & & & & & & \text{Yng} & \text{I} \\
 & & & & & & & \text{Yng} & \text{P} \\
 & & & & & & & & \text{virtual}
 \end{array} \\
 \begin{array}{c|c}
 \text{Old I} & \left[ \begin{array}{cccccc|ccc}
 1 & 1 & \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\
 1 & \cdot & 1 & \cdot & \cdot & 1 & 1 & \cdot & \cdot \\
 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\
 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 \\
 1 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & 1 & \cdot \\
 1 & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & 1
 \end{array} \right] \\
 \text{Yng S} \\
 \text{Old S} \\
 \text{Old P} \\
 \text{Yng P} \\
 \text{Yng I}
 \end{array}
 \end{array}
 \\
 \xrightarrow{1} \begin{array}{c|c}
 \left[ \begin{array}{cccccc|ccc}
 \cdot & \cdot \\
 \cdot & -1 & 1 & -1 & -1 & 1 & 1 & \cdot & \cdot \\
 \cdot & \cdot & 1 & -1 & -1 & \cdot & \cdot & 1 & \cdot \\
 \cdot & \cdot & \cdot & -1 & -1 & \cdot & \cdot & 1 & 1 \\
 \cdot & -1 & \cdot & -1 & -1 & 1 & \cdot & \cdot & 1 \\
 \cdot & -1 & \cdot & \cdot & -1 & 1 & \cdot & \cdot & 1
 \end{array} \right] & \xrightarrow{2} \left[ \begin{array}{cccccc|ccc}
 \cdot & \cdot \\
 \cdot & \cdot \\
 \cdot & \cdot & 1 & -1 & -1 & \cdot & \cdot & 1 & \cdot \\
 \cdot & \cdot & \cdot & -1 & -1 & \cdot & \cdot & 1 & 1 \\
 \cdot & -1 & \cdot & \cdot & -1 & -1 & \cdot & 1 & \cdot \\
 \cdot & -1 & 1 & \cdot & -1 & \cdot & \cdot & 1 & 1
 \end{array} \right]
 \end{array} \\
 \xrightarrow{3} \begin{array}{c|c}
 \left[ \begin{array}{cccccc|ccc}
 \cdot & \cdot \\
 \cdot & \cdot \\
 \cdot & \cdot & \cdot & -1 & -1 & \cdot & \cdot & 1 & 1 \\
 \cdot & \cdot & \cdot & -1 & -1 & -1 & 1 & 1 & \cdot \\
 \cdot & \cdot & \cdot & -1 & -1 & -1 & 1 & \cdot & 1
 \end{array} \right] & \xrightarrow{4} \left[ \begin{array}{cccccc|ccc}
 \cdot & \cdot \\
 \cdot & \cdot \\
 \cdot & \cdot \\
 \cdot & -1 & 1 \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1 & -1 & -1 \\
 \cdot & \cdot & \cdot & \cdot & \cdot & -1 & -1 & 1 & \cdot
 \end{array} \right]
 \end{array}$$

**Fig. 2.** Gaussian elimination steps with virtual cells leading to secondary suppression marked as  $\times$  in Table 1.

## 4.2 The Subspace Method

In Fig. 2, the rows corresponding to zero frequencies in  $y$  (*zero rows*) are arranged at the bottom. Then, no zero rows are eliminated until a column consisting exclusively of zero rows serves as the starting point for the elimination. Assuming that the zero rows are sorted at the bottom, an alternative algorithm that avoids the use of virtual cells is as follows: *Elimination of any of the zero rows is prohibited.* We can see from Fig. 2 that this approach achieves the same result.

However, this is a conservative method that may produce unnecessary secondary cells. Table 2 presents a small toy example that is suitable to illustrate this phenomenon. As shown in Fig. 3, standard Gaussian elimination results in a single secondary suppressed cell. However, if we follow the rule that neither of the two zero rows can be eliminated, then Step 2 in the elimination process shown in Fig. 3 cannot be carried out. Consequently, Total<sub>2</sub> must also be suppressed.

$$\begin{array}{ccccccccc}
 & \text{Total}_1 & & \text{Total}_2 & & & & & \\
 & \text{Old} & \text{I} & \text{Old} & \text{S} & \text{Old} & \text{P} & \text{Yng} & \text{I} & \text{Yng} & \text{P} \\
 & \text{Old} & \text{I} & \text{Yng} & \text{S} & \text{Old} & \text{P} & \text{Yng} & \text{I} & \text{Yng} & \text{P} \\
 \text{Old I} & \left[ \begin{array}{cccc|cc} 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \cdot & 1 & \cdot & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & 1 & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \end{array} \right] & \xrightarrow{1} & \left[ \begin{array}{cccc|cc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & \cdot & 1 & \cdot \\ \cdot & -1 & -1 & \cdot & \cdot & \cdot & 1 \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \end{array} \right] & \xrightarrow{2} & \left[ \begin{array}{cccc|cc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{array} \right] \\
 \text{Yng S} & & & & & & & & \\
 \text{Old S} & & & & & & & & \\
 \text{Old P} & & & & & & & & \\
 \text{Yng P} & & & & & & & & \\
 \text{Yng I} & & & & & & & & 
 \end{array}$$
  

$$\xrightarrow{3} \left[ \begin{array}{cccc|cc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & \cdot & 1 & 1 \end{array} \right] \xrightarrow{4} \left[ \begin{array}{cccc|cc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -1 & \cdot & 1 & 1 \end{array} \right] \xrightarrow{5} \left[ \begin{array}{cccc|cc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & 1 & 1 \end{array} \right]$$

**Fig. 3.** Gaussian elimination steps leading to ordinary secondary suppression marked as \* in Table 2.

## 4.3 The Any-Sum Method

The column for Total<sub>2</sub> in the matrix after step 1 in Fig. 3 represents a difference. A difference known to be zero does not mean that the underlying numbers can be revealed. Therefore, we will modify the rule in Sect. 4.2 as follows: *A column consisting solely of elements in zero rows and with the same sign cannot be the starting point for elimination.* One issue is that allowing such a column with differing signs to be used for elimination impacts other columns. Specifically, this affects imaginary columns representing sums of zeros in  $y$ . By an imaginary column, we mean a column that is hypothetical to discuss what would have happened if it had been included in the elimination process. After such an

	Agr I	Gov I	Tot	Tot	Tot S	Tot P	Tot	Tot	Agr Tot	Ent	S	Gov Tot	Gov P	Tot	I	Ind	I	Agr P	Ent	I	Ent P	Gov S	Ind	P	Ind	S
Ind I	· · 1 · ·	1 · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1 1 ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Agr P	· · 1 · ·	1 · · 1 ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Gov P	· · 1 · ·	1 · · 1 ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1 1 ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Agr S	· · 1 1 · ·	1 1 · · 1	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Ent S	· · 1 1 · ·	1 1 · · 1	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Ind P	· · 1 · ·	1 · · 1 1	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1 ·	· · · · ·	B	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Ind S	· · 1 1 · ·	1 1 · · 1	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1	B	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Ent I	· · 1 · ·	1 · · · ·	1 · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1 · · ·	· · · · ·	1 · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1 · · ·	· · · · ·	A	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Ent P	· · 1 · ·	1 · · 1 ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1 · · ·	· · · · ·	A	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·
Gov S	· · 1 1 · ·	1 1 · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1 · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·	1 · · ·	· · · · ·	A	· · · · ·	· · · · ·	· · · · ·	· · · · ·	· · · · ·

**Fig. 4.** The initial Gaussian elimination matrix associated with Table 3. In the matrix, the columns for structural zeros are placed furthest to the left. The singleton contributors for rows are labeled on the right. Figures 5 and 6 show some steps of two alternative eliminations.

elimination, such an imaginary column may have different signs, thus the sign rule can no longer be relied upon. To ensure safety, an exception must be introduced: *Elimination based on such a column with differing signs will also not be allowed if it involves rows that have been affected by an elimination step where a zero row is eliminated.* With this rule, we have developed a robust and practical method that effectively addresses the problem of zeros. Specifically, there is no need to construct virtual cells with sums of zero-frequency cells. Regardless of which zero-frequency cells we combine into a virtual cell, the rule described here will handle these without them needing to be included in the process. It is thus a requirement that the zero rows are sorted at the bottom in this algorithm.

## 5 Handling Singletons in Magnitude Tables

Singleton cells are primary suppressed cells with a single contributor. These cells therefore correspond to specific columns in the  $X$ -matrix. Unit contributions also correspond to specific rows in the  $X$ -matrix. Thus, some of the columns and some of the rows of  $X$  can be labeled with the IDs of the contributors, also known as holding indicators [12]. In general, the same contributor ID may appear on multiple columns and rows. In Fig. 4, the rows are labeled on the right.

### 5.1 Virtual Cells with Singleton

One approach to handling singletons is to create virtual cells, as done in the Modular approach in Tau-Argus [12]. This rule can be formulated as follows: *If there are exactly two primary suppressed cells in a row or column of a subtable, with at least one being a singleton, then a new primary suppressed virtual cell is created, which represents the sum of these two cells.*

Thus, the example in Table 3 involves four virtual cells, resulting in three secondary suppressed cells, as shown in the same table. However, there are some aspects that this approach does not address. Notably, disclosures made via secondary suppressed cells are not taken into account. For example, contributor A can utilize its own values to reveal Entertainment~Spain, which is secondary suppressed, and subsequently Industry~Spain, which is primary suppressed.

## 5.2 The Within Gaussian Elimination Method

In Sect. 3, the rule for Gaussian elimination is that none of the primary suppressed cells can be completely eliminated. Problems can be seen by continuing the algorithm using the primary suppressed columns from a single contributor as starting columns. If another primary suppressed column is then completely eliminated, then we have uncovered that this single contributor can reveal this particular primary cell. To prevent such disclosure, we now introduce an additional requirement: *The remaining non-zero rows for columns corresponding to primary cells must not originate solely from one contributor.* An essential exception to this rule is needed: *A contributor must be able to reveal cells belonging to themselves.*

Figure 5 illustrates some of the steps in this process. In the seventh elimination step, we must skip, i.e., secondary suppress, the Agriculture~Spain column because using this column could allow A to reveal B's primary cell, Industry~Spain, which is the primary column on the rightmost side. After elimination based on the Agriculture~Spain column seen in the matrix after step 6, the Industry~Spain column would contain only elements from the A rows. Finally, Agriculture~Portugal must also be secondary suppressed for the same reason, as observed after the last step.

With this method, not all disclosures of primary cells are prevented. If the elimination continues after step 7 in the figure, using B's primary cells (the two farthest to the right) as starting points, A's two Entertainment cells will be completely eliminated. Consequently, these cells could be revealed by B. This oversight occurred because one of the B rows was eliminated earlier. Specifically, after the row for Industry~Portugal was eliminated in step 4, the column for Industry~Portugal consisted of rows from both A and B. The method can thus sometimes overlook disclosures due to rows that have been eliminated. To prevent this problem from becoming unnecessarily large, it is important to place the rows for the singleton contributors at the bottom of the  $X$ -matrix.

## 5.3 The Within Parallel Gaussian Eliminations Method

In the general Gaussian algorithm in Sect. 3, different initial sorting of the rows, or alternatively, different choices of elimination rows along the way, will yield the same result. To reduce the problem of overlooked disclosures, we can run two parallel eliminations with differently sorted rows. At each step, we can check for singleton disclosures in both of the parallel eliminations.

As mentioned previously, after the elimination of the Industry~Portugal row, the corresponding column consisted of rows from both A and B. This row was eliminated in step 4, when Industry~Total, as seen in the matrix after step 3 (Fig. 5), was the starting column. Now, imagine that the elimination occurs with the five singleton rows at the bottom sorted in the opposite order. In this scenario, it would be the Entertainment~Iceland row that is eliminated in step 4. However, this would mean that the Entertainment~Iceland column consists solely of two rows belonging to B, enabling B to reveal this primary cell. Therefore, we must skip this elimination and instead, secondary suppress Industry~Total. As shown in Fig. 6, all eliminations starting with the fourth step are different. The result is five secondary suppressed cells, and now all disclosures of primary cells are avoided.

In general, when two parallel eliminations are conducted, overlooked disclosures are rare, and in practice, the method works very well.

#### 5.4 Considering Combinations

In Sect. 5.3, we addressed the protection of primary cells, but we observe that it is still possible to disclose sums from the same contributor, in this case, the sum of A's two Entertainment cells. One way to prevent this is to introduce virtual cells similar to the descriptions in Sect. 4.1 and 5.1. However, managing all possible sums in this manner can be challenging. Moreover, we may also prevent the disclosure of differences and other linear combinations calculated from cells from the same contributor.

A simple way to prevent many such disclosures is to introduce a new rule: *The column that serves as the starting point for elimination must not consist solely of rows from one or two contributors.* Upon reviewing the matrix after step 5 in Fig. 6, this rule alone suffices to determine that the elimination cannot proceed, as the remaining columns consist only of rows from A and B. Consequently, there are seven secondary suppressed cells.

In the matrix after step 5, the Entertainment~Spain column, comprising two rows from A, shows that elimination using this column would completely eliminate a imaginary cell consisting of the sum of the two A cells. Therefore, Entertainment~Spain must be secondary suppressed to prevent the disclosure of the sum of A's two Entertainment cells.

From the Governmental~Portugal column, we can see that more complex disclosures are prevented. If this column is used to eliminate a imaginary cell consisting of the sum of the two B cells, it becomes evident that A can disclose the sum since, after elimination, the cell would consist solely of rows from A. Additionally, since A's cells have opposite signs, B could reveal the difference between Entertainment~Iceland and Governmental~Spain. Thus, disclosure is mutual. Each contributor can use their own numbers to reveal a linear combination of the other's numbers.

Since this method does not protect primary cells other than singleton cells, it serves as a supplement to the methods described in Sect. 5.2 and 5.3. For this method, disclosures can also be overlooked due to eliminated rows. Thus,

parallel eliminations become significantly useful again. In the example in Table 3, no rows are overlooked by single elimination, and parallel eliminations do not produce any different results.

## 6 Concluding Remarks

As described in Sect. 3, Gaussian elimination is based on a general  $X$ -matrix; there are no requirements for a hierarchical structure or any specific layout. Several tables can be included in the same  $X$ -matrix, and therefore, the handling of linked tables is automatically built into the Gaussian elimination algorithm.

Section 4 addressed the issue that arises when zeros are primary suppressed. Sometimes, methods are used where zeros are never suppressed, but other small frequencies are primary suppressed. In this scenario, the ones cannot protect each other because frequencies can only be integers. This problem can be resolved in the same way as the problem with zeros. Therefore, the rules in Sect. 4 can be directly translated by replacing “zero rows” with “one rows”. In situations where zeros can be secondary suppressed, a more complex method than the one described above is necessary. The method required to handle this is beyond the scope of this paper, but it is implemented in the `GaussSuppression` function in SSBtools. As discussed in Sect. 3, structural zeros can be managed by excluding rows from the  $X$ -matrix.

Section 5 focused on the singleton problem, and in particular, Sect. 5.3 proposed two parallel eliminations to reduce the issue of eliminated rows. However, there is no guarantee that this will resolve all issues. A possible improvement could be to use more than two parallel eliminations. Another option is to combine the methods described in Sect. 5.3 and 5.4 with the use of virtual cells, as discussed in Sect. 5.1. To prevent unnecessary secondary cells, virtual cells should be treated differently from other primary cells. A virtual cell does not require protection against disclosure by a singleton contributor. By integrating the methods from Sect. 5.1, 5.3 and 5.4, we have developed a robust and efficient routine that prevents almost all scenarios in which a singleton contributor can reveal sensitive numbers exactly. This option is available in the `GaussSuppression` function in SSBtools.

Examples illustrating the methods in Sect. 4 and 5 can be found in the appendix.

To achieve interval protection, Gaussian elimination can be combined with other methods. Specifically, additional suppression can be applied using another method to meet the interval width requirements for primary cells that did not achieve sufficient interval width after Gaussian elimination. However, this issue is beyond the scope of this paper.

**Acknowledgements.** I would like to thank my colleague Vidar Norstein Klungre at Statistics Norway, as well as two anonymous reviewers, for their valuable comments that led to improvements.

## Appendix

Figures 5 and 6 illustrate steps from two alternative eliminations starting from the matrix in Fig. 4. These figures are referred to in Sect. 5.

To illustrate the methods with realistic examples, we consider so-called hypercubes, according to the European Census 2021 [2]. We utilize data of 1,000,000 synthetic individuals based on Norwegian data, with a synthetic numeric variable and approximately 650,000 unique IDs added for use in magnitude tables [6]. Tables 4 and 5 provide examples of how the methods described in Sect. 4 and 5 affect the number of secondary suppressed cells. Each of the first three table

$$\begin{array}{c}
 \dots \xrightarrow{3} \\
 \dots \xrightarrow{6} \\
 \xrightarrow{7}
 \end{array}
 \left[ \begin{array}{cccc|cc}
 \cdot & \cdot \\
 \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & -1 & \cdot & \cdot \\
 \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & -1 & 1 & -1 & 1 & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & 1 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & 1 & -1 & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\
 \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & 1 & \cdot & \cdot & \cdot & -1 & \cdot & 1 & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & -1 & 1 & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & 1 & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & -1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
 \end{array} \right]$$

**Fig. 5.** Some Gaussian Elimination steps, including the last one, conducted with the matrix in Fig. 4 as a starting point. The treatment of singletons here leads to secondary suppression marked as  $\circ$  in Table 3.

**Fig. 6.** Some Gaussian Elimination steps, including the last one, conducted with the matrix in Fig. 4 as a starting point. Singletons are handled by two parallel eliminations, but the figure only shows the elimination sequence from one of these. The result is the secondary suppressed cells marked as # in Table 3. This figure also illustrates the singleton method which considers combinations. Then the process stops after step 5 and the secondary cells are those marked with  $\Delta$  in Table 3.

groups, consisting of four, three, and three linked hierarchical tables, respectively, is analyzed using the *GaussSuppression* package [8]. These tables also indicate the total execution time, measured on a Linux server, from input microdata to final results.

In Table 4, illustrating the methods in Sect. 4, we chose to primary suppress ones, twos, and threes, while zeros remained unprotected and were considered structural. Thus, these methods address the suppression of ones, which,

**Table 4.** Synthetic data examples illustrating methods for handling ones, equivalent to methods for handling zeros. Each of the first three hypercube groups, according to the European Census 2021, is considered.

	Group 1	Group 2	Group 3
	<u>Secondary suppressed cells</u>		
No method	552	10443	2167
Virtual cells of ones	560	10671	2285
The subspace method	625	14305	3063
The any-sum method	604	12718	2464
Cells to be published, $\text{ncol}(X)$	11356	60971	40141
Primary suppressed cells	114	2563	2093
Primary suppressed ones	54	1053	726
Inner cells, $\text{nrow}(X)$	2186	20396	43219
Time in seconds (any-sum method)	8	318	311

**Table 5.** Synthetic data examples illustrating methods for handling singletons. Each of the first three hypercube groups, according to the European Census 2021, is considered.

	Group 1	Group 2	Group 3
	<u>Secondary suppressed cells</u>		
Virtual Within Parallel Combinations	516	10150	1985
- - - -	576	11394	2238
X - - -	588	12378	2410
- X - -	588	12401	2416
X X - -	588	12501	2451
- X X -	588	12514	2451
X X X -	588	12378	2410
- X - X	588	12418	2416
- X X X	588	12513	2451
X X X X	588	12514	2451
Cells to be published, $\text{ncol}(X)$	11356	60971	40141
Primary suppressed cells	96	1877	1503
Primary suppressed cells that are singletons	54	1053	726
Inner cells, $\text{nrow}(X)$	2186	20396	43219
Unique singleton contributors to inner cells	31	1693	6838
Time in seconds (final method)	138	1073	1544

as described in Sect. 6, is analogous to handling zeros. This approach results in similar examples for frequency and magnitude tables.

The primary suppression of the magnitude tables in Table 5 is based on the  $p\%$ -rule [12], with  $p = 5$ . Note that in the `GaussSuppression` package, virtual cells are created by examining the  $X$ -matrix. This may differ somewhat from the Tau-Argus method described in Sect. 5.1. The beginning of Sect. 5 mentioned that both rows and columns of the  $X$ -matrix can be labeled with singleton contributor IDs. Generally, there may be IDs on rows without a matching column. Extra primary cells representing these *missing* IDs may be added so that all singleton contributions are considered sensitive. This is included in the singleton methods used in Table 5.

## References

1. Bates, D., Maechler, M., Jagan, M.: Matrix: Sparse and Dense Matrix Classes and Methods (2024). <https://CRAN.R-project.org/package=Matrix>, r package version 1.7-0
2. European Commission: Commission regulation (EU) 2017/712 of 20 April 2017 on statistical data and metadata for population and housing censuses. Official Journal of the European Union (2017). <https://eur-lex.europa.eu/eli/reg/2017/712/oj>
3. Fischetti, M., Salazar, J.J.: Solving the cell suppression problem on tabular data with linear constraints. *Manag. Sci.* **47**(7), 1008–1027 (2001). <http://www.jstor.org/stable/822485>
4. Hundepool, A., et al.: Statistical Disclosure Control. Wiley, Hoboken (2012). <https://doi.org/10.1002/9781118348239.ch1>
5. Langsrud, Ø.: Sparse model matrices for multidimensional hierarchical aggregation. *R J.* **15**, 150–166 (2023). <https://doi.org/10.32614/RJ-2023-088>
6. Langsrud, Ø.: About the Norwegian Hypercubes for the 2021 Census (2024). <https://github.com/statisticsnorway/sdc-census-2021-hypercubes>
7. Langsrud, Ø., Bøvelstad, H.M.: Synthetic decimal numbers as a flexible tool for suppression of post-published tabular data. In: Domingo-Ferrer, J., Laurent, M. (eds.) Privacy in Statistical Databases, pp. 105–115. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_8](https://doi.org/10.1007/978-3-031-13945-1_8)
8. Langsrud, Ø., Lupp, D.: GaussSuppression: Tabular Data Suppression using Gaussian Elimination (2024), <https://CRAN.R-project.org/package=GaussSuppression>, r package version 0.8.5
9. Langsrud, Ø., Lupp, D.: SSBtools: Statistics Norway’s Miscellaneous Tools (2024). <https://CRAN.R-project.org/package=SSBtools>, r package version 1.5.2
10. Lupp, D.P., Langsrud, Ø.: Suppression of directly-disclosive cells in frequency tables. In: Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality, Poznań, Poland, 1–3 December 2021 (2021)
11. Meindl, B.: sdcTable: Methods for Statistical Disclosure Control in Tabular Data (2023). <https://CRAN.R-project.org/package=sdcTable>, r package version 0.32.6
12. de Wolf, P.P., Hundepool, A., Giessing, S., Salazar, J.J., Castro, J.: tau-ARGUS user’s manual, version 4.1. Technical report, Statistics Netherlands (2014). <https://github.com/sdcTools/tauargus>



# Obtaining $(\epsilon, \delta)$ -Differential Privacy Guarantees When Using a Poisson Mechanism to Synthesize Contingency Tables

James Jackson<sup>1</sup> , Robin Mitra<sup>2</sup> , Brian Francis<sup>1</sup> , and Iain Dove<sup>3</sup>

<sup>1</sup> Lancaster University, Lancaster, UK

[jejackson96@gmail.com](mailto:jejackson96@gmail.com)

<sup>2</sup> Department of Statistical Science, UCL, London, UK

[robin.mitra@ucl.ac.uk](mailto:robin.mitra@ucl.ac.uk)

<sup>3</sup> Office for National Statistics, Titchfield, UK

**Abstract.** We show that differential privacy type guarantees can be obtained when using a Poisson synthesis mechanism to protect counts in contingency tables. Specifically, we show how to obtain  $(\epsilon, \delta)$ -probabilistic differential privacy guarantees via the Poisson distribution's cumulative distribution function. We demonstrate this empirically with the synthesis of an administrative-type confidential database.

**Keywords:** contingency tables · differential privacy · Poisson mechanism · synthetic data

## 1 Introduction

Differential privacy (DP) [7] is a property of a perturbation mechanism that formally quantifies how accurately any individual's true values can be established, given all other individuals' true values are known. Originally developed as a way to protect the privacy of summary statistics (queries), it soon expanded as a way to protect entire data sets. Differentially private data synthesis (DIPS) has since become a popular area of research; see, for example, [1, 4–6, 12–14].

In [10, 11], we proposed a synthesis approach for contingency tables that uses saturated count models. This approach effectively uses a count distribution to apply noise to the counts in the original data's contingency table, and therefore shares traits with DP mechanisms which apply noise in a similar way. Note that as microdata composed entirely of categorical variables can be expressed in contingency table format, this approach is suitable in the case of categorical data more generally.

In this paper, we consider the ability to obtain DP-guarantees when using the Poisson distribution to synthesize counts in contingency tables. We show that although  $\epsilon$ -DP cannot be satisfied,  $(\epsilon, \delta)$ -DP guarantees can be obtained through the use of the Poisson's cumulative distribution function (CDF).

The motivation behind this work is that, with the exception of [14], the use of count distributions has largely been overlooked as a way to satisfy DP. An obvious benefit of using count distributions is that negative counts cannot be obtained. As the Poisson has only one parameter and hence is likely to be sub-optimal, the intention is that in the future the Poisson could be replaced with more complex count distributions, such as the (discretised) gamma family distribution, where additional parameters provide scope for fine-tuning.

The paper is structured as follows. Section 2 introduces some terminology and definitions. Section 3 looks at existing DP mechanisms for contingency tables, such as the (discretised) Laplace and Gaussian mechanisms. Section 4 gives our novel contribution, the ability to obtain  $(\epsilon, \delta)$ -DP guarantees when using a Poisson synthesis mechanism. Section 5 gives an empirical example using an administrative database. Section 6 gives some concluding remarks.

## 2 Terminology and Definitions

Rinott et al. [15] set out how DP extends into a contingency table setting. Following their notation, let  $\mathbf{a} = (a_k, \dots, a_K) \in \mathcal{A}$  and  $\mathbf{b} = (b_k, \dots, b_K) \in \mathcal{B}$  denote vectors of counts in the original and synthetic data's contingency tables, respectively, where  $K$  denotes the number of cells and  $\mathcal{A}$  and  $\mathcal{B}$  denote the range of obtainable original and synthetic counts (respectively). For contingency tables, we suppose that  $\mathcal{A} = \mathcal{B} = \mathbb{Z}_{\geq 0}^K$ , where  $\mathbb{Z}_{\geq 0}$  is the set of non-negative integers.

Moreover, we describe  $\mathbf{a}$  and  $\mathbf{a}'$  as neighbours, denoted by  $\mathbf{a} \sim \mathbf{a}'$ , whenever all but one of the counts in  $\mathbf{a}$  and  $\mathbf{a}'$  are identical and the differing count differs by exactly one. Henceforth, without loss of generality, we suppose  $\mathbf{a}$  and  $\mathbf{a}'$  differ in their  $k$ th element only, i.e.  $a'_k = a_k - 1$  and  $a_i = a'_i$  for  $i = 1, \dots, K$ ,  $i \neq k$ . Thus  $\mathbf{a}'$  represents the data held by the intruder (who knows all but one of the individuals' true values) and  $\mathbf{a}$  represents the completed data where the "unknown individual" has been added to the cell in which they truly belong.

The  $\epsilon$ -DP definition revolves around the likelihood ratio, or, more accurately, around a series of likelihood ratios.

**Definition 1 ( $\epsilon$ -DP).** A perturbation mechanism  $\mathcal{M}$  satisfies  $\epsilon$ -DP ( $\epsilon > 0$ ) if:

$$\exp(-\epsilon) \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})} \leq \exp(\epsilon), \quad (1)$$

$$\forall \mathbf{a} \sim \mathbf{a}' \in \mathcal{A}, \forall \mathbf{b} \in \mathcal{B}.$$

Definition 1 is the special case of the standard DP definition, given in [7], for when the range of  $\mathcal{A}$  and  $\mathcal{B}$  are discrete. Although we appreciate that in some instances the denominator in (1) could be equal to zero, for the mechanisms we consider here this probability is always non-zero.

For any  $\mathbf{a}$ ,  $\mathbf{a}'$  and  $\mathbf{b}$ , whenever the ratio  $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})/\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})$  is either small or large, relatively too much is gleaned about the unknown individual's true values. It is worth noting, too, that the above definition considers all possible

synthetic data sets in  $\mathcal{B}$ , illustrating that DP is not a risk metric for a particular synthetic data set but rather a property of a synthesis mechanism.

Somewhat confusingly, there are two similar but different relaxations of  $\epsilon$ -DP. The first is  $(\epsilon, \delta)$ -differential privacy [8]. The second is known as  $(\epsilon, \delta)$ -probabilistic differential privacy [12]. These are given below in Definitions 2 and 3. In the remainder of this paper, we focus on  $(\epsilon, \delta)$ -probabilistic DP. Yet whenever  $(\epsilon, \delta)$ -probabilistic DP is satisfied,  $(\epsilon, \delta)$ -DP is also satisfied [9].

**Definition 2**  $((\epsilon, \delta)\text{-DP})$ . A perturbation mechanism  $\mathcal{M}$  satisfies  $(\epsilon, \delta)$ -DP ( $\epsilon > 0; 0 \leq \delta \leq 1$ ) if:

$$\frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b) - \delta}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq \exp(\epsilon) \quad \text{and} \quad \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b) - \delta}{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)} \leq \exp(\epsilon) \quad (2)$$

$\forall \mathbf{a} \sim \mathbf{a}' \in \mathcal{A}, \mathbf{b} \in \mathcal{B}.$

**Definition 3**  $((\epsilon, \delta)\text{-probabilistic DP})$ . A perturbation mechanism  $\mathcal{M}$  satisfies  $(\epsilon, \delta)$ -probabilistic DP ( $\epsilon > 0; 0 \leq \delta \leq 1$ ) if:

$$\mathbb{P}\left[\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq \exp(\epsilon)\right] > 1 - \delta \quad \forall \mathbf{a} \sim \mathbf{a}' \in \mathcal{A}, \mathbf{b} \in \mathcal{B}. \quad (3)$$

**Theorem 1**  $((\epsilon, \delta)\text{-probabilistic DP implies } (\epsilon, \delta)\text{-DP})$ . If a perturbation mechanism  $\mathcal{M}$  satisfies  $(\epsilon, \delta)$ -probabilistic DP, then it also satisfies  $(\epsilon, \delta)$ -DP. (Proof: see [9])

### 3 Examples of Existing DP Mechanisms

We now give examples of existing DP mechanisms suitable for synthesizing counts in contingency tables. Note that for the Laplace and Gaussian mechanisms, discretised noise needs to be added (unless one is willing to accept non-integer ‘‘counts’’). This can simply involve adding continuous noise before rounding the adjusted values to the nearest integer. Similarly, negative values can be rounded to zero.

*Example 1 (The Laplace mechanism).* A random variable  $X \sim \text{Laplace}(\mu, d)$  has probability density function  $f_L$ :

$$f_L(x; \mu, d) = \frac{1}{2d} \exp\left(-\frac{|x - \mu|}{d}\right).$$

The Laplace mechanism  $\mathcal{M}_L$  satisfies  $\epsilon$ -DP by using the Laplace distribution to add random noise to the original counts  $\mathbf{a}$ . Specifically, for every original count  $a_i$ , the Laplace mechanism generates a  $\text{Laplace}(a_i, 1/\epsilon)$  random variate. To show that this mechanism does indeed satisfy DP, we suppose that  $a_i = a'_i$  for  $i = 1, \dots, k-1, k+1, \dots, K$  and that  $a'_k = a_k - 1$  (i.e. the assumptions made in Sect. 2). Firstly, when  $b_k > a_k$ :

$$\begin{aligned}
\frac{\mathbb{P}(\mathcal{M}_L(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_L(\mathbf{a}') = \mathbf{b})} &= \frac{\exp(-\epsilon|b_k - a_k|)}{\exp(-\epsilon|b_k - a'_k|)} \\
&= \frac{\exp(-\epsilon|b_k - a_k|)}{\exp(-\epsilon|b_k - (a_k - 1)|)} \\
&= \exp(\epsilon).
\end{aligned} \tag{4}$$

Similarly, when  $a_k > b_k$ , (4) is equal to  $\exp(-\epsilon)$ , and when  $a_k = b_k$  it is equal to  $\exp(0)$ . Hence the DP definition in (1) holds.

*Example 2 (The Gaussian mechanism).* A random variable  $X \sim \text{Normal}(\mu, \sigma^2)$  has probability density function  $f_G$ :

$$f_G(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

In a similar way to the Laplace mechanism, the Gaussian mechanism, say  $\mathcal{M}_G$ , applies  $\text{Normal}(0, \sigma^2)$  random noise to the original counts, resulting in a mechanism that satisfies  $(\epsilon, \delta)$ -differential privacy. Using the same assumptions and notation as previous, it follows that:

$$\begin{aligned}
\frac{\mathbb{P}(\mathcal{M}_G(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}_G(\mathbf{a}') = b)} &= \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{b_k-a_k}{\sigma}\right)^2\right]}{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{b_k-a_k+1}{\sigma}\right)^2\right]} \\
&= \exp\left[-\frac{1}{2\sigma^2}(2a_k - 2b_k - 1)\right].
\end{aligned}$$

Recall that  $(\epsilon, \delta)$ -probabilistic DP is satisfied whenever

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq \exp(\epsilon) \quad \text{with probability } 1 - \delta,$$

which, in this instance, occurs whenever

$$-\epsilon \leq -\frac{1}{2\sigma^2}(2a_k - 2b_k - 1) \leq \epsilon \quad \text{with probability } 1 - \delta.$$

The probability  $1 - \delta$  can be obtained from  $\Phi$ , the normal distribution's CDF [2], as  $b_k \sim \text{Normal}(a_k, \sigma^2)$ .

$$\begin{aligned}
1 - \delta &= \mathbb{P}\left(-\epsilon \leq -\frac{1}{2\sigma^2}(2a_k - 2b_k - 1) \leq \epsilon\right) \\
&= \mathbb{P}(a_k - \sigma^2\epsilon - 1/2 \leq b_k \leq a_k + \sigma^2\epsilon - 1/2) \\
&= \Phi\left(\frac{a_k + \sigma^2\epsilon - 1/2 - a_k}{\sigma}\right) - \Phi\left(\frac{a_k - \sigma^2\epsilon - 1/2 - a_k}{\sigma}\right) \\
&= \Phi(\sigma\epsilon - 1/(2\sigma)) - \Phi(-\sigma\epsilon - 1/(2\sigma))
\end{aligned}$$

*Example 3 (Multinomial-Dirichlet synthesizer).* A multinomial-Dirichlet synthesis mechanism [1], say  $\mathcal{M}_{MD}$ , can also yield DP guarantees. The original counts  $\mathbf{a}$  can be converted to cell probabilities  $\boldsymbol{\pi}$  simply by dividing by  $n$  (the number of individuals in the data). A Dirichlet prior with concentration parameters  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K)$  is placed on  $\boldsymbol{\pi}$  (see [1] for more on this approach). Using the same “without loss of generality” assumptions as previous, it follows that

$$\begin{aligned}\frac{\mathbb{P}(\mathcal{M}_{MD}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_{MD}(\mathbf{a}') = \mathbf{b})} &= \frac{\Gamma(b_k + a_k + \alpha_k)}{\Gamma(a_k + \alpha_k)} \cdot \frac{\Gamma(a'_k + \alpha_k)}{\Gamma(b_k + a'_k + \alpha_k)} \\ &= \frac{\Gamma(b_k + a_k + \alpha_k)}{\Gamma(a_k + \alpha_k)} \cdot \frac{\Gamma(a_k - 1 + \alpha_k)}{\Gamma(b_k + a_k - 1 + \alpha_k)} \\ &= \frac{b_k + a_k - 1 + \alpha_k}{a_k - 1 + \alpha_k}.\end{aligned}\tag{5}$$

Recall again that DP is satisfied whenever

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})} \leq \exp(\epsilon).$$

As the expression in (5) is always greater than or equal to one, and hence always greater than  $1/\exp(\epsilon)$ , DP is satisfied whenever

$$\frac{b_k + a_k - 1 + \alpha_k}{a_k - 1 + \alpha_k} \leq \exp(\epsilon).$$

As  $a_k \geq 1$  and  $b_k \leq n$ , this simplifies to

$$\frac{n + \alpha_k}{\alpha_k} \leq \exp(\epsilon) \quad \Rightarrow \quad \alpha_k \geq \frac{n}{\exp(\epsilon) - 1}.$$

Considering all counts  $a_1, \dots, a_K$  gives that DP is satisfied whenever

$$\max_i \alpha_i \geq \frac{n}{\exp(\epsilon) - 1}, \quad \text{a result from [12].}$$

## 4 Satisfying $(\epsilon, \delta)$ -Probabilistic DP with a Poisson Synthesis Mechanism

When using saturated count models to synthesize contingency tables, as set out in [11], a count distribution, e.g. the Poisson, applies noise to original counts. We assume that a constant pseudocount  $\alpha > 0$  is added to every element of  $\mathbf{a}$  (i.e. to *all* original counts, not just to zero counts as in [11]), which opens up the possibility that original counts of zero can be synthesized to non-zeros. When using the Poisson we apply the following mechanism, which we denote by  $\mathcal{M}_P$ , to obtain a set of synthetic counts:

$$b_i | a_i, \alpha \sim \text{Poisson}(a_i + \alpha), \quad i = 1, \dots, K,$$

$$\text{i.e. } \mathbb{P}(\mathcal{M}_P(a_i) = b_i) = \frac{\exp(-a_i - \alpha)(a_i + \alpha)^{b_i}}{b_i!}, \quad i = 1, \dots, K.$$

Supposing once again that  $\mathbf{a}$  and  $\mathbf{a}'$  differ in their  $k$ th element only, we have:

$$\frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} = \exp(-1) \left( \frac{a_k + \alpha}{a_k - 1 + \alpha} \right)^{b_k}. \quad (6)$$

This quantity is bounded below by  $\exp(-1)$ , with this minimum occurring when  $b_k = 0$ . It is unbounded above, however, as  $b_k$  can take any integer up to infinity; i.e. the expression in (6) tends to infinity as  $b_k$  tends to infinity. Thus  $\epsilon$ -DP cannot be satisfied.

Instead, we now consider the  $(\epsilon, \delta)$ -probabilistic DP relaxation, first considering the left-hand inequality of the DP definition (Definition 1):

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} \Rightarrow b_k \geq \frac{1 - \epsilon}{\log \left( \frac{a_k + \alpha}{a_k - 1 + \alpha} \right)}.$$

When  $\epsilon \geq 1$ , this inequality holds with probability 1. When  $0 < \epsilon < 1$ , the probability that this inequality holds can be determined through the Poisson's CDF, since  $b_k$  is a realization from a Poisson random variable. This probability is given as:

$$1 - F_{a_k + \alpha}^P \left[ \frac{1 - \epsilon}{\log \left( \frac{a_k + \alpha}{a_k - 1 + \alpha} \right)} \right], \quad (7)$$

where  $F_{a_k + \alpha}^P$  is the CDF of the Poisson distribution with mean  $a_k + \alpha$ .

We next consider the right-hand inequality of Definition 1:

$$\frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} \leq \exp(\epsilon) \Rightarrow b_k \leq \frac{1 + \epsilon}{\log \left( \frac{a_k + \alpha}{a_k - 1 + \alpha} \right)}.$$

For all  $\epsilon$ , this inequality holds with probability

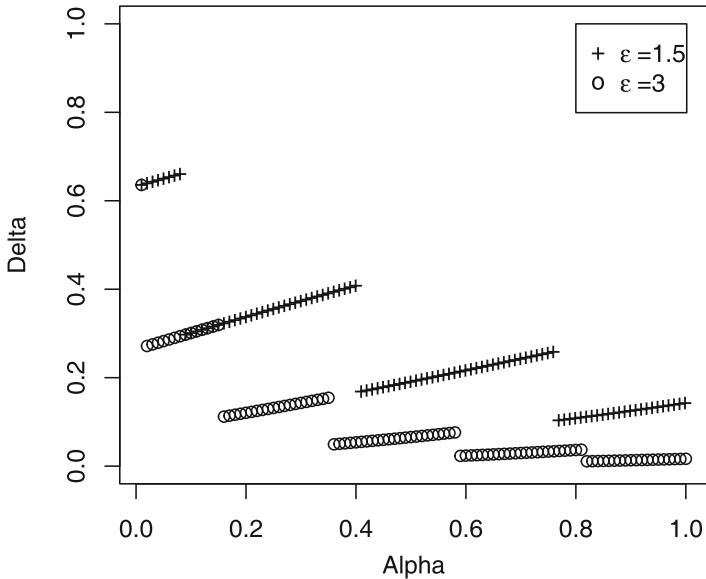
$$F_{a_k + \alpha}^P \left[ \frac{1 + \epsilon}{\log \left( \frac{a_k + \alpha}{a_k - 1 + \alpha} \right)} \right]. \quad (8)$$

Recall that in  $(\epsilon, \delta)$ -probabilistic DP,  $1 - \delta$  is the probability that DP is satisfied, i.e. the probability that both inequalities hold. A non-trivial question when  $0 < \epsilon < 1$  is how to combine the probabilities given in (7) and (8) and hence compute  $\delta$ ? This is an area of future research.

When  $\epsilon > 1$ , however, the left-hand inequality of Definition 1 always holds, thus we need only focus on (8). Although non-trivial for any  $\epsilon \geq 1$  and  $\alpha > 0$ , (8) is minimised when  $a_k = 1$  (when  $a'_k = 0$ ). Note, a formal proof has been omitted here but extensive empirical simulation results have been undertaken. Thus,

$$1 - \delta = F_{1+\alpha}^P \left[ \frac{1 + \epsilon}{\log \left( \frac{1+\alpha}{\alpha} \right)} \right]. \quad (9)$$

This also demonstrates the role of  $\alpha$  as a tuning parameter for risk. In general, a larger  $\alpha$  value corresponds to a lower  $\delta$  value. Yet  $\delta$  is not a decreasing function of  $\alpha$ . For a very brief explanation, this is because increasing  $\alpha$  increases the value of the expression inside the squared bracket in (9), but it also increases the mean of the Poisson random variable from which a synthetic count is drawn. Figure 1 illustrates the nature of the relationship between  $\alpha$  and  $\delta$  for different values of  $\epsilon$ . For example, setting  $\alpha = 0.1$  satisfies approximately  $(3,0.3)$ -probabilistic DP and  $(1.5,0.6)$ -probabilistic DP.



**Fig. 1.** The relationship between  $\alpha$  and  $\delta$  in the Poisson synthesis mechanism for  $\epsilon = 1.5$  and  $\epsilon = 3$ .

In contingency tables where there are no zero counts, a  $(\epsilon, \delta)$ -DP guarantee can be obtained when  $\alpha = 0$ . In this instance,  $\delta$  is determined by the smallest original count, i.e.:

$$1 - \delta = F_{a_i + \alpha}^P \left[ \frac{1 + \epsilon}{\log \left( \frac{\min_i a_i + 1}{\min_i a_i} \right)} \right]. \quad (10)$$

In a sense, in this example we have violated the traditional  $(\epsilon, \delta)$ -probabilistic DP definition given in (3) because  $\delta$  is dependent on a particular set of original counts  $\mathbf{a}$  – not all original counts.

We can easily replace the Poisson with any other count distribution (e.g. the negative binomial, Poisson inverse-Gaussian, Delaporte, Sichel, etc.), which of course would lead to a different expression for the ratio in (6).

## 5 An Empirical Example

### 5.1 The English School Census Administrative Database

The English School Census (ESC) is a large administrative database belonging to the UK's Department for Education (DfE), which holds information about pupils attending state-funded schools in the UK. Owing to the presence of sensitive data, strict privacy guarantees would be required for data from the ESC to be made available to researchers. There is therefore great appeal to DP-type approaches, where more formal guarantees of privacy can be obtained.

Access to the real ESC data is currently restricted, even for the sake of demonstrating the effectiveness of privacy methods. For this reason, staff at the Office for National Statistics (ONS) created a substitute data set using publicly-available data sources, such as published ESC data and 2011 UK census data. A key feature of this data set,  $\text{ESC}_{\text{rep}}$ , is that it replicates some of the statistical properties present in the actual ESC. We take a subset of this data which has approximately  $8 \times 10^6$  individuals (rows) and 5 categorical variables (columns). As all variables are categorical, the data set can be expressed as a contingency table with around  $3.5 \times 10^6$  cells. More information about the data set – as well as the data set itself – is available at [3].

### 5.2 Applying the Poisson Synthesis Mechanism

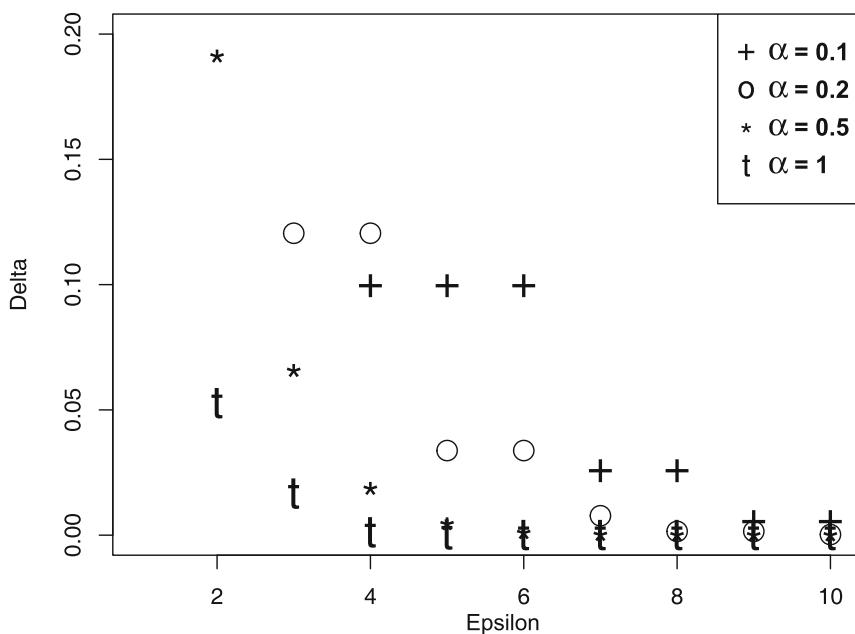
We now apply the Poisson synthesis mechanism to the  $\text{ESC}_{\text{rep}}$  data, considering different values of  $\alpha$ , and considering  $\epsilon > 1$  values.

Figure 2 gives combinations of  $(\epsilon, \delta)$  values that can be achieved for the  $\text{ESC}_{\text{rep}}$  data when using  $\alpha$  values of 0.1, 0.2, 0.5 and 1. For example, when  $\epsilon = 2$ , an  $\alpha$  value of 1 is required to obtain a  $\delta$  value of 0.05; when  $\alpha = 0.1$ , a  $\delta$  value of 0.05, is obtained only for  $\epsilon$  values greater than 6.

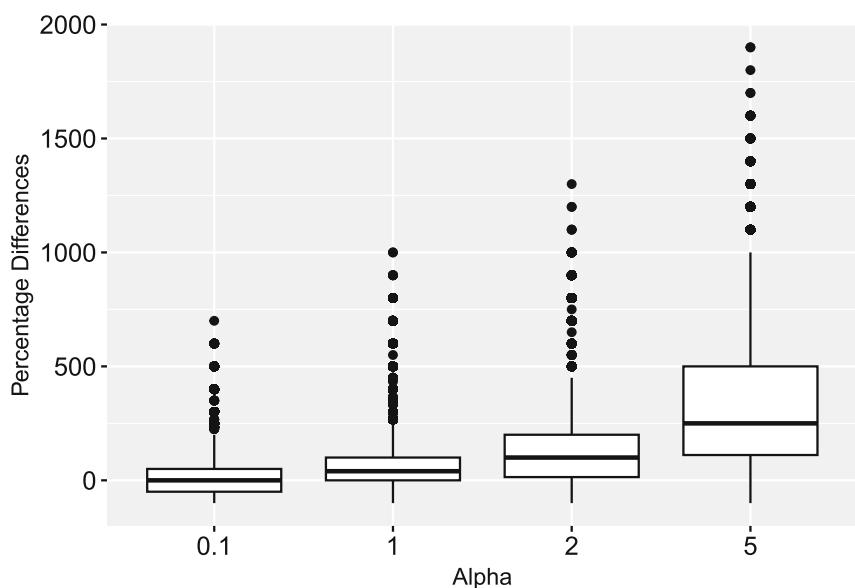
DP methods, in general are known to have a detrimental effect on utility. To gain a simple insight into general utility [16], the boxplots in Fig. 3 compare the percentage differences between original and synthetic counts for various values of  $\alpha$ , and for original counts between 1 and 10. Unsurprisingly, increasing  $\alpha$  increases the percentage differences, i.e. has an adverse effect on utility. This loss of utility is more magnified in specific analyses, especially when the analyst wishes to quantify uncertainty.

## 6 Discussion

To summarise, in this paper we have shown how to obtain  $(\epsilon, \delta)$ -DP guarantees when using a Poisson synthesis mechanism to protect the privacy of counts in contingency tables. For a given  $\epsilon > 1$ , the corresponding value of  $\delta$  that is achievable with the Poisson is relatively high; much higher than that which is achievable with other DP mechanisms. Going forward, we believe other count distributions, such as the negative binomial, are likely to be more favourable



**Fig. 2.** Combinations of  $\delta$  such that  $(\epsilon, \delta)$ -probabilistic DP is achieved when the Poisson is used, for various  $\max_i a_i$  and  $\epsilon$  equal to 1.5, 2, 2.5 and 3.



**Fig. 3.** For different values of  $\alpha$ , boxplots showing percentage differences between original and synthetic counts (utility) for original counts in the range 1–10.

(i.e. will give better utility results), while also providing the same DP-type risk guarantees, because such distributions would introduce further tuning parameters in addition to  $\alpha$ . Previous work suggests that such tuning parameter apply noise in a more efficient fashion [10]. These tuning parameters could be set to obtain certain  $\epsilon$  or  $\delta$  values.

We end with an interesting note in relation to DP. Somewhat counter-intuitively, the reason why multinomial-based synthesis mechanisms (e.g. the multinomial-Dirichlet synthesizer) can satisfy  $\epsilon$ -DP – but the Poisson cannot – is that multinomial mechanisms have a maximum synthetic count that any original count can take, namely  $n$ . With count distributions, any original count can be synthesized to any non-negative integer. To help explain why this causes the DP definition to fail, recall that with contingency tables DP definitions effectively assume that the intruder is trying to locate the cell to which just one individual belongs; i.e. in the intruder’s data set one, and only one, cell count is one less than it actually is. Now suppose that a particular count in the intruder’s data set is equal to 1, but that the corresponding synthetic count – generated by simulating from the Poisson with  $\alpha = 0$  – has a count of 5. It is 11.7 times more likely that this synthetic count originated from a cell with a count of 2 than from a count of 1, therefore the intruder can infer that that particular cell is a likely origin of the target. It is interesting therefore that, with DP, disclosure risk is deemed to be at its greatest when the scope for potential movement between original and synthetic counts is at its greatest. This largely goes against the objectives of traditional SDC methods, which typically reduce risk by increasing the divergence from the original counts.

## References

1. Abowd, J.M., Vilhuber, L.: How protective are synthetic data? In: Domingo-Ferrer, J., Saygin, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 239–246. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-87471-3\\_20](https://doi.org/10.1007/978-3-540-87471-3_20)
2. Balle, B., Wang, Y.X.: Improving the Gaussian mechanism for differential privacy: analytical calibration and optimal denoising. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 394–403. PMLR (2018). <https://proceedings.mlr.press/v80/balle18a.html>
3. Blanchard, S., Jackson, J.E., Mitra, R., Francis, B.J., Dove, I.: A constructed English School Census substitute (2022). <https://doi.org/10.17635/lancaster/researchdata/533>
4. Bowen, C.M., Liu, F.: Comparative study of differentially private data synthesis methods. Stat. Sci. **35**(2), 280–307 (2020). <https://doi.org/10.1214/19-STS742>
5. Charest, A.S.: How can we analyze differentially-private synthetic datasets? J. Priv. Conf. **2**(2) (2011). <https://doi.org/10.29012/jpc.v2i2.589>. <https://journalprivacyconfidentiality.org/index.php/jpc/article/view/589>
6. Drechsler, J.: Differential privacy for government agencies—are we there yet? J. Am. Stat. Assoc. **118**(541), 761–773 (2023). <https://doi.org/10.1080/01621459.2022.2161385>

7. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). [https://doi.org/10.1007/11681878\\_14](https://doi.org/10.1007/11681878_14)
8. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends®in Theor. Comput. Sci. **9**(3-4), 211–407 (2014). <https://doi.org/10.1561/0400000042>
9. Goetz, M., Machanavajjhala, A., Wang, G., Xiao, X., Gehrke, J.: Publishing search logs - a comparative study of privacy guarantees. IEEE Trans. Knowl. Data Eng. **24**, 520–532 (2012). <https://doi.org/10.1109/TKDE.2011.26>
10. Jackson, J., Mitra, R., Francis, B., Dove, I.: On integrating the number of synthetic data sets  $m$  into the a priori synthesis approach. In: Domingo-Ferrer, J., Laurent, M. (eds.) Privacy in Statistical Databases 2022, pp. 205–219. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_15](https://doi.org/10.1007/978-3-031-13945-1_15)
11. Jackson, J., Mitra, R., Francis, B., Dove, I.: Using saturated count models for user-friendly synthesis of large confidential administrative databases. J. R. Stat. Soc. Ser. A: Stat. Soc. **185**(4), 1613–1643 (2022). <https://doi.org/10.1111/rssa.12876>
12. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: 2008 IEEE 24th International Conference on Data Engineering, pp. 277–286. IEEE (2008)
13. McClure, D., Reiter, J.P.: Differential privacy and statistical disclosure risk measures: an investigation with binary synthetic data. Trans. Data Priv. **5**(3), 535–552 (2012)
14. Quick, H.: Generating Poisson-distributed differentially private synthetic data. J. R. Stat. Soc. A: Stat. Soc. **184**(3), 1093–1108 (2021). <https://doi.org/10.1111/rssa.12711>. <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12711>
15. Rinott, Y., O'Keefe, C.M., Shlomo, N., Skinner, C., et al.: Confidentiality and differential privacy in the dissemination of frequency tables. Stat. Sci. **33**(3), 358–385 (2018)
16. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. J. R. Stat. Soc. A: Stat. Soc. **181**(3), 663–688 (2018)

# **Synthetic Data Generation Methods**



# Generating Synthetic Data is Complicated: Know Your Data and Know Your Generator

Jonathan Latner<sup>1</sup> , Marcel Neunhoeffer<sup>1,2</sup> (✉), and Jörg Drechsler<sup>1,2,3</sup>

<sup>1</sup> Institute for Employment Research, Nuremberg, Germany

{jonathan.latner,marcel.neunhoeffer,joerg.drechsler}@iab.de

<sup>2</sup> Ludwig-Maximilians-Universität, Munich, Germany

<sup>3</sup> University of Maryland, College Park, USA

**Abstract.** In recent years, more and more synthetic data generators (SDGs) based on various modeling strategies have been implemented as Python libraries or R packages. With this proliferation of ready-made SDGs comes a widely held perception that generating synthetic data is easy. We show that generating synthetic data is a complicated process that requires one to understand both the original dataset as well as the synthetic data generator. We make two contributions to the literature in this topic area. First, we show that it is just as important to pre-process or clean the data as it is to tune the SDG in order to create synthetic data with high levels of utility. Second, we illustrate that it is critical to understand the methodological details of the SDG to be aware of potential pitfalls and to understand for which types of analysis tasks one can expect high levels of analytical validity.

**Keywords:** Synthetic · Utility · CTGAN · DataSynthesizer · synthpop

## 1 Introduction

The idea of synthetic microdata<sup>1</sup> for statistical disclosure limitation was introduced more than 30 years ago [6, 8, 14] and has become increasingly popular in recent years [2, 5]. The general appeal of synthetic data is obvious: synthetic data promises to mimic the statistical properties of the original data while maintaining the confidentiality of individual records. In practice, releasing synthetic data means fitting a model to the original data and then generating new data based on this model.<sup>2</sup> The new interest in synthetic data was spurred by the develop-

<sup>1</sup> Throughout the paper, when we refer to data, we are referring to classical microdata (i.e., one observation per individual unit), as opposed to summary tables or images.

<sup>2</sup> Note that there are different philosophies about the definition of original data and how much pre-processing (e.g., dealing with missing values or outliers) one should do to the original data before data synthesis depending on the synthesis goals (replacement of the original data vs. tool for preparing to work with the original data in a safe environment). In Sect. 3 we describe the data and any pre-processing steps in detail.

ment of new methods and algorithms to generate synthetic data, many of which are available as open-source software packages, see, e.g., [9, 12, 17].

This leaves the general impression that generating synthetic data has become easy and that software libraries offer a one size fits all solution. While it is indeed easy (even for novices) to generate synthetic data, it is hard to generate high-quality data. This paper addresses the question: What makes generating high-quality synthetic data hard? In particular, we focus on the data one wants to synthesize and the synthetic data generators (SDG) one wishes to use. We show that one must possess both good knowledge of the data and the generator when deciding which SDG might be appropriate for what data.

## 2 Study Design

In this study, we illustrate the complications in generating high quality synthetic data with limited knowledge of the data and/or the generator. To do so, we use one dataset (SD2011) and three different SDGs (CTGAN, DataSynthesizer, and synthpop). We emphasize that the goal of this paper is explicitly not to compare the performance of the three synthesizers. The SDGs we use have already been extensively compared and contrasted in previous papers [1, 7]. We only use these synthesizers to illustrate our point that knowledge of the data and of the synthesizer are key to obtain high utility synthetic datasets. We use three utility measures: pMSE and computational run-time (as defined in Appendix A) as well as one-way frequency tables to compare the graphical distribution of variables for the synthetic and original data. Replication code is available on GitHub.<sup>3</sup>

### 2.1 Data

The data we use are called Social Diagnosis 2011 - Objective and Subjective Quality of Life in Poland (SD2011) and are included as part of the synthpop package, but are also publicly available.<sup>4</sup> The data comprise 35 variables and 5000 observations, 21 variables are categorical, and 14 variables are continuous.

The reason why we use SD2011 data is that it contains many properties of real data, including missing values, outliers, ‘messy’ values, and generated variables. Previous evaluations used clean data from Census [7] or machine learning benchmarks (Kaggle, UCI, OpenML) [1]. Clean or even simulated data can be valuable, especially for evaluation purposes, but real data present challenges for SDGs that are not otherwise understood. For example, some SDGs cannot be used on data with missing values. Therefore, the ability to synthesize missing values is a stage of development that SDGs must go through before they can be applied to real data.<sup>5</sup>

---

<sup>3</sup> [https://github.com/jonlatner/KEM\\_GAN/tree/main/latner/projects/comparison](https://github.com/jonlatner/KEM_GAN/tree/main/latner/projects/comparison).

<sup>4</sup> <http://www.diagnoza.com/index-en.html>.

<sup>5</sup> Early versions CTGAN could not be used on data with missing values (<https://github.com/sdv-dev/CTGAN/issues/39>).

## 2.2 Synthetic Data Generators (SDGs)

**synthpop** (Version 1.8.0) [9] is an R package that implements parametric and ML based models (classification and regression trees (CART) and random forests) to generate synthetic data. In our application we use the default settings of the package.<sup>6</sup> synthpop follows a sequential process, where the first variable to be synthesized is generated by drawing new values from the marginal distribution of this variable (either by drawing from a parametric distribution or by sampling from the empirical distribution), and the subsequent variables are synthesized one at a time, always conditioning on those variables that have been synthesized in earlier steps.

**DataSynthesizer** (Version 0.1.13) [12] is a Python package that implements the PrivBayes algorithm [19]. PrivBayes is designed to address the challenges associated with a differentially private method for releasing synthetic data outputs from high-dimensional real data inputs. To do this, the package implements a Bayesian network model to estimate the joint distribution of the data.

To generate synthetic data from a Bayesian network, the first step is to specify a graphical model (a directed acyclical graph (DAG)) that represents how and in what way the different variables are related to each other. DataSynthesizer doesn't require this model structure as input, instead it tries to estimate the optimal structure given the data. As a hyperparameter users can set the maximum number of parents ( $k$ ) that should be considered for the model. The more parents, the more complex relationships between the variables. After specifying the model and estimating its parameters from the original data, synthetic data are generated by sampling new values based on the probabilities from the model's conditional probability table.

For example, imagine data with four columns (variables) with categorical values: age (young, middle, old), education (less than secondary, secondary, and more than secondary), gender (M and F), and income (low, middle, and high). The number of observations (or rows) are not relevant because the data are transformed into a frequency table with one cell for each unique combination of groups. In this example, there are 54 cells ( $3 \times 3 \times 2 \times 3$ ). If we assume that age, education, and gender are the parents of income, then each value in the conditional probability table represents the conditional probability of each income category given the states of the other three variables. The algorithm calculates the probabilities based on the frequencies from all possible combinations of the variables. To make this model tractable for high dimensional data, the graph structure enforces conditional independence between some of the variables, reducing the number of parameters that need to be estimated. Once the model is defined and parameters are estimated, the Bayesian network generates synthetic data by sampling new values using the estimated probabilities. In our applications, We use default settings except for the number of parents, which

---

<sup>6</sup> Default means that CART models are used for synthesis with complexity parameter = 0.001 (smaller values will grow larger trees), and minbucket = 5 (the minimum number of observations in any terminal node).

we set to  $k = 2$  (default is “greedy”, which means that DataSynthesizers tries to find the optimal value for  $k$ ).

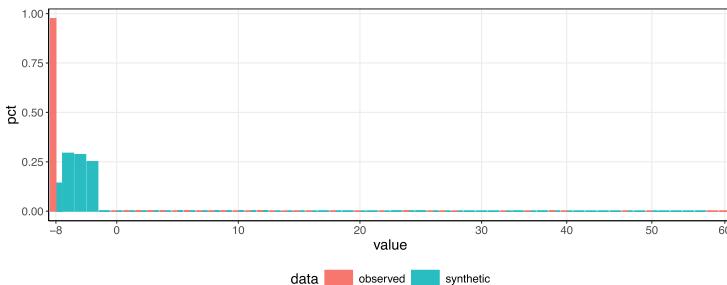
**CTGAN** (Version 1.9.0) [17] is a Python package that is part of the Synthetic Data Vault (SDV) package [11]. In their original application, generative adversarial networks (GANs) were designed to create synthetic images [4], but the approach was later adapted to also create synthetic microdata [10]. GANs simultaneously train two neural networks: a generator and a discriminator. The goal of the generator is to create synthetic data that becomes increasingly indistinguishable from original data. The goal of the discriminator is to get better at distinguishing between original and synthetic data. This adversarial process goes back and forth until the discriminator cannot distinguish between the original data and the generated data.

To illustrate how GANs generate synthetic data, imagine we have one variable: income (LN) from Census data with a mean of 10 and a standard deviation of 1. First, the generator network receives random noise vectors as input, typically sampled from a standard normal distribution with mean of 0 and a standard deviation of 1, and sends it to the discriminator to be evaluated. Second, the discriminator evaluates the synthetic data alongside real data to determine the probability that the generated data is real (1) or fake (0). If the discriminator determines that the generated data are fake, it sends feedback to the generator in the form of a loss function. Higher values indicate that it is easier for the discriminator to differentiate between real and fake data. Third, the generator then updates its parameters based on the loss function and the learning rate, which determines the magnitude of this update. The higher the learning rate, the larger the adjustments the generator will make to its parameters in response to the feedback. Fourth, updated data generated from the updated parameters are sent to the discriminator. Ultimately, this back and forth process results in a generator that produces synthetic data that ideally has the same statistical properties as the original data. In our application we mostly rely on default settings except for the number of epochs which we set to 600 (default is 300), but we also vary a number of other hyperparameters, as we explain in detail below.

### 3 Know Your Data

SD2011 contain a variety of characteristics found in real data that can present a challenge to synthetic data generators (SDGs). These challenges should be addressed prior to applying a SDG. However, cleaning the data requires knowledge of the data that is not always available to those with knowledge of a given SDG and may not be easy to detect or follow simple rules. We use DataSynthesizer ( $k = 2$ ) to demonstrate the importance of preprocessing the data, but the points raised in this section are applicable to all SDGs.

**Missing Values.** In real data, missing values are sometimes coded as either negative values or large positive values (999999). For example, in the variable `wkabdur` (Months working abroad in 2007–2011) from SD2011, 97.5% of all values are –8. The interpretation is values of –8 represent missing values and only 2.5%



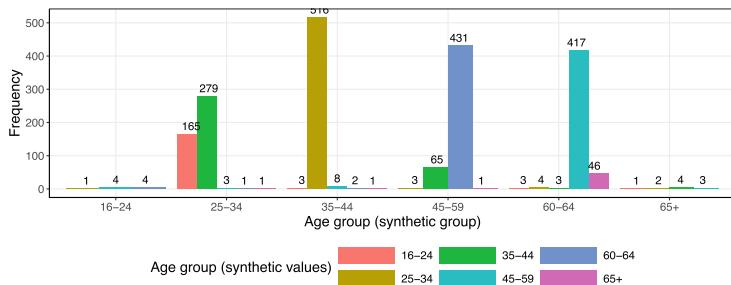
**Fig. 1.** The encoding of missing values is important. In this example, missing values of a numeric variable (on the x-axis are the values of the variable “Work abroad duration (in weeks)”) were encoded as  $-8$ . If these negative values are treated as numeric when generating the synthetic data other (meaningless) negative values can be created. In red we display the distribution of values in the original data, in blue the distribution of values in the synthetic data. (Color figure online)

of the units contained in the sample worked abroad between 2007 and 2011. If these values are not cleaned and coded as missing before applying the SDG or the SDG is informed that these values represent missing values (which is possible for example in synthpop), then the SDG will treat these values as regular values to be included in the synthetic data, which will reduce statistical utility. For example, if we did not code values of  $-8$  as missing in the original data, then values between  $-8$  and  $0$  would be created in the synthetic data that do not exist in the original data, as shown in Fig. 1.

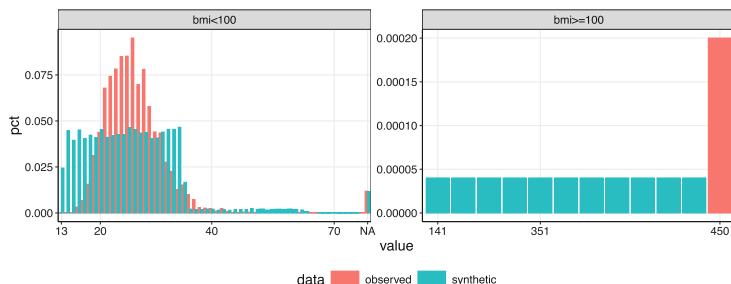
**Generated Variables.** Generated variables are variables that are deterministic functions of other variables included in the dataset. In SD2011, there are two generated variables. The variable `agegr` is generated from `age` by classifying the age variable into 6 age groups and the variable `bmi` is generated from the variables `height(cm)` and `weight(kg)` ( $bmi = weight / (height/100)^2$ ).

It is not necessary to synthesize generated variables and doing so can be problematic as inconsistencies might arise in the synthetic data. For example, in the case of `agegr`, if we apply the SDG to the original data that still includes `agegr`, then some synthetic values of `agegr` would be inconsistent with the synthetic values of `age`. In Fig. 2 the frequency of the age groups derived from the synthetic `age` variable are shown for each category of `agegr` (frequencies are based on  $m = 5$  synthetic datasets). We drop values that match so the graph only shows the mismatches. For example, for synthetic observations classified in age group 25–34 according to `agegr`, 165 (0.66%) observations have a generated synthetic age between 16–24 and 279 (1.12%) have a generated synthetic age between 35–44. In total, 7.88% of all observations would be misclassified.

To avoid these inconsistencies, generated variables should be dropped prior to applying the SDG, and then recreated based on the synthetic values of the underlying variables. However, one needs to be aware of the problem to avoid it. In practical situations, logical constraints between the variables can be much



**Fig. 2.** Synthesizing variables generated variables, i.e., variables that are deterministic functions of other variables, can lead to inconsistencies in the synthetic data. Here, the generated variable synthetic age group is not always consistent with the synthetic age values, i.e., DataSynthesizer misclassified some observations.

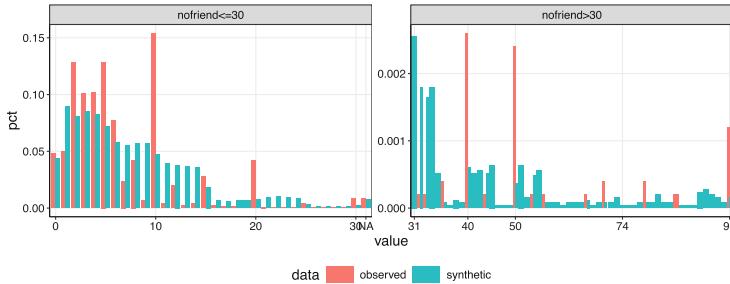


**Fig. 3.** Not cleaning the original data can be problematic. For some SDGs the presence of one (unrealistic) outlier leads to synthetic data with many outliers. Here, one false value of the variable body mass index (BMI) leads to a synthetic distribution with many unrealistic values using DataSynthesizer.

more difficult to identify and it requires good knowledge of the data to avoid implausibilities that subject matter experts would easily detect.

**Outliers.** Outliers can cause problems if they are not modeled carefully. For example, in the case of `bmi`, there is an outlier value of 450 in the original data. While this value is an obvious error,<sup>7</sup> what would be the implications if it were correct (and thus could not simply be removed before the synthesis)? If we included the value of 450 in the original data, then some SDGs would create values between 450 and 76 in the synthetic data that do not exist in the original data, as shown in Fig. 3. Fortunately, in our case, `bmi` is a generated variable, which we can drop. However, the example serves to illustrate that the existence of outlier values can affect the ability of SDGs to create synthetic data with high levels of utility and might also be problematic from a privacy perspective.

<sup>7</sup> The record with a BMI of 450 has `height` (cm) = 149 and `weight` (kg) = NA. If we calculate weight from `bmi` and `height`, then weight equals 999 or one metric ton.



**Fig. 4.** ‘Spikey’, discontinuous, or semi-continuous distributions can be problematic for SDGs. Here, DataSynthesizer seems to smooth spikes in the variable “Number of friends”.

**Messy Values.** In real data, variables include messy values, such as spikes in variables that otherwise can be treated as being continuous. An example in SD2011 is the variable `nofriend` (number of friends). In the original data, the variable `nofriend` appears to be normally distributed below 10, but then clusters at values of 10, 15, 20 and groups of 10 up to the maximum 99. This phenomenon is sometimes referred to as ‘spikey’, discontinuous, or semi-continuous distributions. As shown in Fig. 4 and discussed in more detail below, errors and spikes can pose a problem for statistical utility because most SDGs tend to smooth these spikes unless they are explicitly modeled.

To illustrate the importance of pre-processing the data, we use three versions of the original data to create synthetic data and evaluate the utility of each generated dataset using the pMSE. SD2011(a) is the raw data. SD2011(b) codes all negative values that indicate missing values in numeric variables and all empty values in character variables as missing. SD2011(c) drops the two generated variables (`agegr` and `bmi`) and then recreates them from the synthetic values. Using DataSynthesizer as our SDG and setting the DAG structure to allow a maximum of two parents, the pMSE changes from 0.2 for SD2011(a) to 0.13 for SD2011(b) to 0.07 for SD2011(c). As measured by the pMSE, the improvements in utility are substantial. In fact, when experimenting with the different tuning parameters of the different synthesizers, we found that none of the tuning parameters had such a strong impact on utility as pre-processing the data (results not reported).

## 4 Know Your Generator

In this section we discuss the importance of knowing the details of the underlying methodology of the SDG to avoid pitfalls or to at least be aware for which analysis tasks the generated data might offer reasonable analytical validity and for which not. We illustrate this point by providing an example of a methodological aspect for each synthesizer that has important impacts on the synthesis, but that might not be immediately obvious when only considering the general methodology of the SDG.

## 4.1 Synthpop

Categorical variables with large numbers of categories can substantially increase the run-time of CART based synthesizers. To understand why, it is important to understand how CART models are built. CART models operate by finding recursive binary splits to maximize the homogeneity of the values of the dependent variable in the two leaves generated by the split. To find the best possible splits, CART models search over all variables in the dataset. For each variable all possible splits are evaluated and the split that maximizes the homogeneity across all splits and all variables is selected. For continuous and ordered categorical variables the number of splits that needs to be evaluated is  $k - 1$ , where  $k$  is the number of unique values in the variable. This is because each value is considered as a possible splitting criteria with all values less than the value ending up in the left leaf and all other values ending up in the right leaf.

For unordered categorical variables, the number of splits that need to be considered is  $2^{L-1} - 1$ , where  $L$  is the number of categories, i.e., the number of splits grows exponentially with the number of categories. To illustrate, imagine one categorical variable with three values (a, b, and c). There are  $2^2 - 1 = 3$  possible options to split this variable: (1 = a; 0 = b,c), (1 = b; 0 = a,c), (1 = c; 0 = a,b). With six categories, we already need to consider 31 splits, which still doesn't pose a problem computationally. However, if there are 20 categories, then 524,288 splits need to be considered. The computational burden can be substantial with even a few categorical variables with a large number of categories.

In the sequential modeling approach that is used with synthpop, each variable that has been synthesized previously is used as a predictor in all subsequent synthesis models. If a variable with many categories is synthesized early during the synthesis process, it will always be used as a predictor for all other synthesis models imposing a high computational burden on the synthesizer. On the other hand, if the variable is the last variable to be synthesized, it will never be used as a predictor, considerably speeding up the synthesis process.

For this reason, it is recommended to synthesize categorical variables with many categories last when using CART models [13]. However, problems arise if there are a sufficient number of variables with a large number of unique values. For example, if a Census data set contained variables for 3-digit ISO country code, 3-digit ISCO codes (occupation), and 3-digit ISIC codes (industry), then it would be difficult to avoid computational problems through ordering.

Other solutions exist, but limitations remain: One option is to aggregate categorical variables with a large number of unique values. However, avoiding the information loss from aggregation is typically one of the reasons to rely on synthetic data to begin with. Alternatively, the categorical variable can be used to stratify the data and to run separate synthesis models within each stratum. Obviously, this will only be an option if the sample sizes in each stratum are still large enough to allow sufficiently rich synthesis models within each stratum.

To illustrate the problem with categorical variables, we examine the duration in time required to create synthetic data using synthpop, i.e. computational efficiency, as shown in Table 1. If we load the raw SD2011 into R as a .csv file as

**Table 1.** The effect of data pre-processing on the execution time of the SDGs measured in seconds in wall-clock time. All SDGs were executed on the same machine.

version	description	ctgan	datasynthesizer	synthpop (csv)	synthpop (package)
v00	Raw (SD2011)	331.01	245.37	2132.12	5474.39
v01	Without eduspec or wkabdur	290.30	264.43	10.99	8.45
v02	Without wkabdur	337.07	351.76	13.96	11.02
v03	Without eduspec	306.46	351.24	11.39	8.92
v04	Last variables: eduspec-wkabdur	374.57	344.02	14.23	287.85
v05	Last variables: wkabdur-eduspec	419.60	339.92	14.60	3657.55
v06	as.numeric(wkabdur) and last variable: eduspec	356.02	347.36	14.12	11.05
v07_1_20	+ 1 factor variable (20 values)	339.05	264.96	42.23	
v07_1_25	+ 1 factor variable (25 values)	400.28	326.84	137.47	
v07_1_30	+ 1 factor variable (30 values)	339.73	269.72	363.18	
v07_2_20	+ 2 factor variable (20 values)	369.74	339.45	74.96	
v07_2_25	+ 2 factor variable (25 values)	364.56	361.81	631.43	
v07_2_30	+ 2 factor variable (30 values)	373.25	346.15	1222.54	
v07_3_20	+ 3 factor variable (20 values)	393.99	369.58	122.77	
v07_3_25	+ 3 factor variable (25 values)	401.03	383.40	881.53	
v07_3_30	+ 3 factor variable (30 values)	394.44	424.64	3654.59	

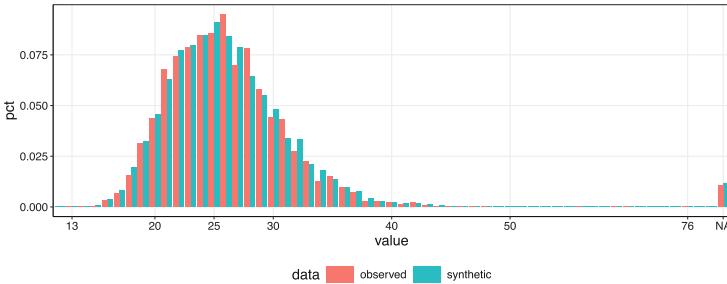
Notes: synthpop (csv) indicates that SD2011 was loaded into R from a csv file. synthpop (package) indicates that SD2011 was loaded into R from the synthpop package. CTGAN is estimated with 600 epochs. DataSynthesizer is estimated with two parents.

we normally do with a SDG, then synthpop requires 2,132 s (35 min and 30 s), but if we load SD2011 into R from the synthpop package, then synthpop requires 5,474 s (91 min). The difference in duration within synthpop is explained by two variables: **eduspec** and **wkabdur**. Both variables have a large number of unique values (28 and 33, respectively). Within the synthpop package **wkabdur** is coded as a character variable, which implies that CART automatically treats it as an unordered categorical variable. However, if SD2011 is loaded into R from a .csv file, then **wkabdur** is treated as a numeric variable. When we treat **wkabdur** as a numeric variable and place **eduspec** at the end of the synthesis chain, synthpop requires less than 15 s, regardless of how SD2011 is loaded.

However, the issue reveals a more general problem that synthpop is sensitive to the number (and order) of categorical variables with large values. To illustrate this, we added one, two, or three categorical variables with 20, 25, or 30 unique random categorical values to the end of the original SD2011 data. Results are presented in Table 1. Duration times for synthpop increase with additional variables and number of unique values. Table 1 also reveals that the other synthesizers are less sensitive to the number of categorical variables comprising many categories.

## 4.2 DataSynthesizer

The algorithm used by DataSynthesizer assumes that all variables are categorical, like most Bayesian network models [18]. Making this assumption can simplify the modeling task required to represent complex relationships. Relatedly, this increases computational efficiency and allows the model to be applied to



**Fig. 5.** Original and synthetic values of body mass index (BMI) from using DataSynthesizer and the pre-cleaned SD2011(c) dataset.

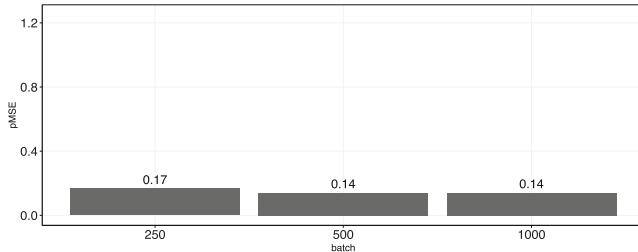
high dimensional data (see Table 1) because the information in the data can be reduced without information loss to the full conditional probability table.

When dealing with continuous variables, Bayesian Networks categorize or discretize these variables prior to estimating the model. This is done by binning the continuous variable and using the binned variable for the model. Once synthetic data have been generated based on this model, values for continuous variables are re-encoded within each bin. Given that the Bayesian Network implicitly assumes conditional independence of the values within a bin and all the other variables contained in the model, it is critical for the analytical validity of the generated data how this encoding step is implemented. DataSynthesizer implements an encoding strategy that samples values from a uniform distribution within each bin.<sup>8</sup>

To demonstrate the consequences of this strategy, we refer to Figs. 3 and 4 again, which compare frequency counts for the two variables `bmi` (body mass index) and `nofriend` (number of friends) for SD2011(a), i.e. the raw data. For `nofriend`, DataSynthesizer underestimates the frequencies for small counts and does not preserve the spikes in the data for the larger counts, as it distributes the frequencies equally within the bins. For `bmi`, synthetic values are not normal as they are in the original data because DataSynthesizer samples from a uniform distribution within the bins.

However, if we drop `bmi` before applying DataSynthesizer, and then regenerate `bmi` from `height` and `weight` as we do in SD2011(c), then synthetic values of `bmi` are distributed as in the original data, as shown in Fig. 5. While this repeats the importance of cleaning the data prior to applying the synthesizer, this is not

<sup>8</sup> We note that the DataSynthesizer paper states [12], “when invoked in correlated attribute mode, DataDescriber samples attribute values in appropriate order from the Bayesian network.” However, in the code, it seems that data are created by uniform sampling within a bin ([https://github.com/DataResponsibly/DataSynthesizer/datatypes/AbstractAttribute.py#L125](https://github.com/DataResponsibly/DataSynthesizer/blob/90722857e7f6ed736aaa25068ecf9e77f34f896a/DataSynthesizer/datatypes/AbstractAttribute.py#L125)). This illustrates the challenge in understanding the methodological details of a given SDG.



**Fig. 6.** The relationship between batch size and utility of the synthetic data generated by CTGAN (measured by the pMSE) with the number of training steps held constant (3,000).

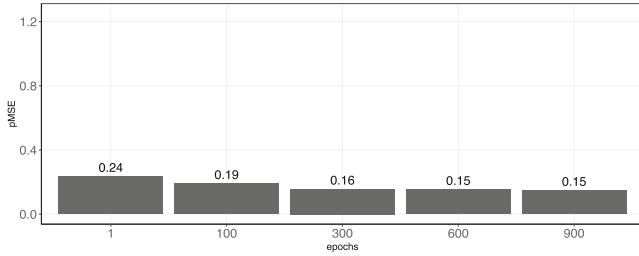
a universal solution as the spikiness of `nofriend` is a feature of the variable that cannot be cleaned.

Of course, the utility of the synthetic data could be improved by increasing the number of bins. However, increasing the number of bins will increase the computational complexity. Besides, setting the number of bins too high would lead to unstable model estimates, as a very large number of parameters would need to be estimated from the data. Furthermore, if the formal privacy guarantees of DataSynthesizer are turned off as in our application, increasing the number of bins can also lead to increased risks of disclosure. If, on the other hand, the formal guarantees should be maintained when increasing the number of bins, more noise needs to be added to each of the parameters and utility would suffer.

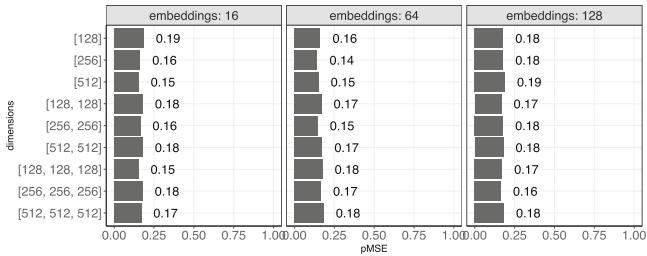
### 4.3 CTGAN

GANs are designed to work well with continuous variables, but GANs can struggle modeling relationships in micro data because cells in micro data are not informative of neighbors in the same way that pixel cells are in a photograph [2]. CTGAN contains a number of hyperparameters that one can use to tune the model. For this study, we tune the hyperparameters in three main ways. First, we maintain a constant number of iterations (3,000), but allow the batch size to vary, as shown in Fig. 6. Second, we maintain a constant batch size (500), but allow the number of iterations to vary, as shown in Fig. 7. Third and finally, we vary the dimensionality of the generator/discriminator networks and the embedding dimension, as shown in Fig. 8. In our data, tuning these hyperparameters makes little difference ( $pMSE \approx 0.16$ ). For reference,  $pMSE$  from CTGAN is higher than DataSynthesizer (0.07) and synthpop (0.02).

We are aware of other research where tuning hyperparameters does make a difference in the quality of synthetic data output (as yet unpublished). One possible explanation for the fact that hyperparameters do not affect the quality of the synthetic data output is that our data are too low dimensional for the parameters to make a difference. While CTGAN does not produce data with



**Fig. 7.** The relationship between the number of update steps and utility of the synthetic data generated by CTGAN (measured by the pMSE) with the batch size held constant (500).



**Fig. 8.** The relationship between the architecture of the CTGAN and utility of the synthetic data (measured by the pMSE).

high levels of utility with the data we use here, we do not mean to suggest that GANs are bad SDGs in general. CTGAN is not the only GAN in Synthetic Data Vault and multiple other GANs exist. Based on our own experience, it is possible to create a GAN that provides higher levels of utility. More generally, one must distinguish between the package and the method.

## 5 Conclusion

While generating synthetic data is easier than ever, generating high quality synthetic data remains complicated. In this article, we make two points. First, one must know the data. Most research examining SDGs uses clean data from Census or Machine Learning libraries, especially in the computer science literature. Unlike clean data, real data contain variables with messy values that can affect the utility of SDGs. It is not possible to simply input real data into a SDG and expect high quality synthetic data output without carefully pre-processing the data. The problem is that producing synthetic data with high levels of utility requires knowledge of the data that may not be understood by those with knowledge of the generator.

Second, one must know the generator. Synthpop produces synthetic data with high levels of utility, but the CART method struggles with computational

efficiency on datasets that contain variables with many categories. DataSynthesizer uses a Bayesian network model to produce synthetic data with high levels of computational efficiency, but the algorithm assumes all variables are discrete, which reduces utility of synthetic continuous variables. CTGAN uses a GAN architecture to generate synthetic data which may not produce synthetic data with high levels of utility. There is no one size fits all solution and choosing the right SDG is the result of different trade-offs.

**Acknowledgments.** This work was supported by a grant from the German Federal Ministry of Education and Research (grant number 16KISA096) with funding from the European Union-NextGenerationEU.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## A Appendix: Utility Measures

The **propensity score (pMSE)** is an utility measure that estimates how well one can discriminate between the original and synthetic data based on a classifier [15, 16] and is implemented in R from the synthpop package [9]. This is sometimes called a ‘broad’ [15] or ‘general’ [3] measure of utility or ‘statistical fidelity’ [5]. The main steps are to append or stack the original and the synthetic data, add an indicator (1/0) to distinguish between the two, use a classifier to estimate the propensity of each record in the combined dataset being ‘assigned’ to the original data. The pMSE is the mean squared error of these estimated propensities:

$$pMSE = \frac{1}{N} \sum_{i=1}^N [\hat{p}_i - c]^2 \quad (1)$$

where  $N$  is the number of records in the combined dataset,  $\hat{p}_i$  is the estimated propensity score for record  $i$ , and  $c$  is the proportion of data in the merged dataset that is synthetic (in many cases  $c = 0.5$ ). The pMSE can be estimated using all the variables in the dataset, but it can also be computed using subsets of the variables, e.g., all pairwise combinations of variables to evaluate specifically how well the distribution of these variables is preserved. The smaller the pMSE, the higher the analytical validity of the synthetic data.

**Computational efficiency** is the run time (in seconds) required to create one single synthetic dataset from a given SDG.<sup>9</sup> This is sometimes referred to as ‘efficiency’ [5] or ‘output scalability’ [19]. The basic idea is that the algorithms used by SDGs can suffer from the curse of dimensionality.

---

<sup>9</sup> In terms of computing power, SDGs were run on a 2022 Macbook Air with 16GB of RAM and an M2 Chip with 8-Core CPU, 8-Core GPU, and a 16-Core Neural Engine. All SDGs were run one at a time in order to minimize computational power problems from parallelization.

## References

1. Dankar, F.K., Ibrahim, M.: Fake it till you make it: guidelines for effective synthetic data generation. *Appl. Sci.* **11**(5), 21–58 (2021)
2. Drechsler, J., Haensch, A.C.: 30 years of synthetic data. arXiv preprint [arXiv:2304.02107](https://arxiv.org/abs/2304.02107) (2023)
3. Drechsler, J., Reiter, J.: Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB establishment survey. *J. Official Stat.* **25**(4), 589–603 (2009)
4. Goodfellow, I., et al.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, vol. 27 (2014)
5. Jordon, J., et al.: Synthetic data – what, why and how? (2022)
6. Liew, C.K., Choi, U.J., Liew, C.J.: A data distortion by probability distribution. *ACM Trans. Database Syst. (TODS)* **10**(3), 395–411 (1985)
7. Little, C., Elliot, M., Allmendinger, R.: Comparing the utility and disclosure risk of synthetic data with samples of microdata. In: Domingo-Ferrer, J., Laurent, M. (eds.) *PSD 2022. LNCS*, vol. 13463, pp. 234–249. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_17](https://doi.org/10.1007/978-3-031-13945-1_17)
8. Little, R.J., et al.: Statistical analysis of masked data. *J. Official Stat.* **9**, 407–407 (1993)
9. Nowok, B., Raab, G.M., Dibben, C.: synthpop: bespoke creation of synthetic data in R. *J. Stat. Softw.* **74**, 1–26 (2016)
10. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on generative adversarial networks. arXiv preprint [arXiv:1806.03384](https://arxiv.org/abs/1806.03384) (2018)
11. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410 (2016). <https://doi.org/10.1109/DSAA.2016.49>
12. Ping, H., Stoyanovich, J., Howe, B.: Datasynthesizer: privacy-preserving synthetic datasets. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pp. 1–5 (2017)
13. Raab, G.M., Nowok, B., Dibben, C.: Guidelines for producing useful synthetic data. arXiv preprint [arXiv:1712.04078](https://arxiv.org/abs/1712.04078) (2017)
14. Rubin, D.B.: Statistical disclosure limitation. *J. Official Stat.* **9**(2), 461–468 (1993)
15. Snock, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. *J. R. Stat. Soc. Ser. A Stat. Soc.* **181**(3), 663–688 (2018)
16. Woo, M.J., Reiter, J.P., Oganian, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confidentiality* **1**(1) (2009)
17. Xu, L., Skouliaridou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: *Advances in Neural Information Processing Systems* (2019)
18. Young, J., Graham, P., Penny, R.: Using Bayesian networks to create synthetic data. *J. Official Stat.* **25**(4), 549–567 (2009)
19. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: PrivBayes: private data release via Bayesian networks. *ACM Trans. Database Syst. (TODS)* **42**(4), 1–41 (2017)



# Evaluating the Pseudo Likelihood Approach for Synthesizing Surveys Under Informative Sampling

Anna Oganian<sup>1</sup>(✉), Jörg Drechsler<sup>2</sup>, and Mehtab Iqbal<sup>3</sup>

<sup>1</sup> National Center for Health Statistics, Centers for Disease Control and Prevention,  
3311 Toledo Rd, Hyattsville, MD 20782, USA  
[aoganyan@cdc.gov](mailto:aoganyan@cdc.gov)

<sup>2</sup> Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany  
[joerg.drechsler@iab.de](mailto:joerg.drechsler@iab.de)

<sup>3</sup> School of Computing, Clemson University, 821 McMillan Rd, Clemson, SC 29631, USA  
[mehtabi@g.clemson.edu](mailto:mehtabi@g.clemson.edu)

**Abstract.** In recent years, national statistical organizations have increasingly relied on synthetic data when releasing microdata containing sensitive personal or establishment information. This paper deals with the challenges of using synthetic data to protect the privacy of survey respondents. For this type of data it is often important to consider the survey design information when creating the synthesis models.

The paper discusses two techniques that can be used for generating survey microdata under informative sampling. Specifically, it examines an approach that combines design-based and model-based methods through the use of the pseudo-likelihood approach within the sequential regression framework. As far as we are aware, the pseudo-likelihood method has not been used in the context of sequential regression synthesis before.

This method is compared with another approach in which design variables are included as predictors in the regression models. In the latter approach, the survey weights have to be synthesized and included in the final data product, while the former generates synthetic simple random samples that are representative of the original population without weights.

**Keywords:** Statistical disclosure limitation · survey data · pseudo likelihood · synthetic data · probability proportional to size sampling · sequential regression modeling · survey weights

## 1 Introduction

It is the legal responsibility of statistical agencies to safeguard the privacy and confidentiality of individuals or enterprises participating in national surveys. The Privacy Act of 1974 [23], the Public Health Service Act [24], and the Confidential Information Protection and Statistical Efficiency Act [1] are examples of laws in the USA that address this requirement.

At the same time, the public release of the data in a disaggregated format presents significant benefits for research and policy-making. Releasing synthetic survey data can be beneficial when the data is considered to be highly sensitive. Disseminating synthetic data instead of the original data might allow for the inclusion of sensitive variables that might otherwise have to be suppressed and offering finer granulation for categorical variables. This level of detail might otherwise only be available through research data centers (RDCs). See [3,4,9] for further details on the synthetic data approach. Still, synthesizing survey data can be challenging. Most existing methods for data synthesis assume that the original data covers the entire population or can be considered a simple random sample from the population [3,4]. However, this is not the case for most population survey data, which is typically collected using complex sampling designs. Using synthesis models that do not account for the survey design may result in biases in design-based inferences, especially for finite population totals and means if the sampling design is informative [10]. Informative in this context means that the distribution in the sample differs from the distribution in the population.

Thus, just as complex survey data analysis necessitates methods tailored to the sampling design through weighting and specialized variance estimation techniques, the synthesis of survey data may benefit from the use of special synthesis techniques to generate synthetic survey data. To our knowledge, there is no consensus in the statistical disclosure limitation (SDL) community on how to account for the survey design in the synthetic data context.

In the survey statistics literature, it is common practice to include all design variables, i.e., those variables that govern the sampling design, in the models to make the sampling design ignorable when estimating the model parameters. However, it is not always possible to fully incorporate all features of the sampling design through this strategy. Besides, when disseminating sensitive data, releasing the original weights might be problematic [5] and thus, the weights will also have to be synthesized. The major challenge with this approach is that it requires the relationship between the survey weights and the variables to be modeled correctly for synthesis. One early strategy for dealing with complex sampling designs in data synthesis was presented in [12]. The authors propose two strategies that both built on the idea of using the synthetic versions of the design variables to recalculate the sampling weights. We did not include this strategy in our evaluations as the procedure would only allow to account for the sampling design, but not for further adjustments to the design weights that are typically necessary in practice to account for nonresponse or coverage errors. Another approach discussed in the synthetic data literature is to use the finite population Bayesian bootstrap that accounts for the complex design when generating synthetic populations [2]. This approach has the disadvantage that it generates the synthetic populations by sampling actual records from the original data and thus offers very limited disclosure protection. More recently, [10] proposed a synthesis strategy that relies on a pseudo-likelihood approach [6,7,14]. The authors show in the context of business surveys that their approach is superior to including the design variables into the synthesis models. However, they rely on a joint modeling strategy, when implementing their approach. Joint modeling requires specifying one joint distribution for the entire dataset and generating synthetic data by drawing from this model. In practice, survey data are typically

multidimensional and consist of dozens if not hundreds of variables of different types. Specifying one joint distribution for all these variables is typically infeasible in practice. Therefore, a sequential regression approach is commonly employed. As the name suggests, datasets are synthesized sequentially with this approach, using regression models that condition only on variables that have already been synthesized in earlier steps. This data synthesis strategy is available in various software packages such as *synthpop* in R [13] and *IVEware* [16], which can be called from SAS but is also available as a standalone product. Since the implementation of this approach is available in free software and given that it provides a flexible tool to account for different data structures, it is more attractive for data custodians as it reduces the modeling burden for a broad range of data synthesis scenarios.

This paper focuses on generating fully synthetic data based on the sequential regression approach for survey data with an informative sampling design. Some examples of such designs are probability proportional to size sampling (PPS), stratified designs with unequal probabilities of selection, and others [10, 21]. We extend the pseudo-likelihood approach presented in [10] to the sequential regression context and use simulations to compare its performance to the standard approach that incorporates the design variables as predictors in the synthesis models. We note that for full synthesis this implies that the survey weights need to be synthesized as well.

## 2 Synthesis Strategies to Account for Complex Sampling Designs

Let us first introduce the notation that will be used throughout the paper. Consider a microdata set with  $p$  variables. The sequential regression framework, which is the focus of this paper, is a well-established approach that can be used to generate synthetic data and/or to impute missing data. It is based on using a series of regression models to approximate the joint distribution of a  $p$ -variate data set, where every variable is treated separately using regression models suitable for that specific variable. Continuous variables can be generated using a normal model; binary variables can be generated using logistic regression, and so on. Let  $Y = \{Y_1, \dots, Y_p\}$  denote the variables in the dataset. The first model generates synthetic data by drawing new values from the marginal distribution of  $Y_1$ . The second model generates synthetic values for  $Y_2$  based on  $f(Y_2|Y_1)$ . Subsequent models utilize synthetic values from prior steps to predict the next variable in the sequence. There are parametric and nonparametric approaches to building the individual models in the sequence; see [3] for a comprehensive introduction. A review of the inferential procedures necessary for obtaining valid inferences based on standard synthesis as well as for the pseudo-likelihood approach is provided in the Appendix.

### 2.1 The Pseudo-likelihood Approach for Data Synthesis

The pseudo-likelihood approach is based on a construct called a pseudo-population. The pseudo-population can be created by replicating each record in the sample  $w_i - 1$  times, where  $w_i$  is the survey weight of unit  $i$ . The pseudo-population is a useful tool that approximates the original population based on the sample and mimics its structure. This idea is closely related to the Horwitz-Thompson estimator [8], which is widely used in

survey sampling. The pseudo-likelihood approach builds on this idea, by weighting the contribution of each sample unit by its survey weight when setting up the pseudo-likelihood function:

$$PL(\Theta; Y) = \prod_{i=1}^n f(\mathbf{y}_i | \Theta)^{w_i},$$

where  $\mathbf{y}_i$  is the  $(1 \times p)$  vector of observed values for unit  $i$ ,  $i = 1, \dots, n$ ,  $n$  is the sample size, and  $\Theta$  contains the unknown distribution parameters to be estimated. When adopting the pseudo-likelihood approach for the sequential regression context, we assume that for each regression model in the chain of regressions  $Y \sim X$ , where  $X$  is the matrix of predictors, the pseudo-likelihood function is given by:

$$PL(\Theta; Y|X) = \prod_{i=1}^n f(y_i | x_1^{(i)} \dots x_k^{(i)}, \Theta)^{w_i} \quad (2.1)$$

where  $k$  is the number of variables that have been synthesized in previous steps of the sequential regression approach.

It is easy to verify that this assumption is equivalent to using the pseudo-likelihood to model the entire dataset.

$$\begin{aligned} PL(\Theta; Y_1, Y_2, \dots, Y_p) &= \prod_{i=1}^n f(y_1^{(i)}, \dots, y_p^{(i)} | \Theta)^{w_i} \\ &= \prod_{i=1}^n f(y_1^{(i)} | \Theta)^{w_i} \prod_{i=1}^n f(y_2^{(i)} | y_1^{(i)}, \Theta)^{w_i} \dots \prod_{i=1}^n f(y_p^{(i)} | y_1^{(i)}, \dots, y_{p-1}^{(i)}, \Theta)^{w_i} \end{aligned} \quad (2.2)$$

For continuous variables, Eq. 2.1 can be written as:

$$\prod_{i=1}^n f(y_i | \mathbf{x}_i; \vec{\beta}, \sigma) = \prod_{i=1}^n \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^{w_i} \exp \left( -\frac{w_i(y_i - \mathbf{x}_i \vec{\beta})^2}{2\sigma^2} \right) \quad (2.3)$$

By maximizing the pseudo-likelihood above with respect to the parameters  $\beta$  and  $\sigma$  we can get the estimates of these parameters necessary for the synthesis of  $y$ :

$$\hat{\beta} = (XWX')^{-1}XWY \quad (2.4)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n w_i(y_i - \mathbf{x}_i \vec{\beta})^2}{\sum_{i=1}^n (w_i)} \quad (2.5)$$

where  $X$  is the  $k \times n$  matrix of predictors,  $W$  is the  $n \times n$  diagonal matrix of weights, and  $Y$  is the  $n \times 1$  vector representing the predicted variable.

These are expressions of regression coefficients and regression standard error weighted using survey weights. Note that expression (2.4) for the regression coefficients obtained using the pseudo-likelihood approach is equivalent to the one that would be obtained by using the Horwitz-Thompson estimators of the population totals  $X_{pop}$  and  $Y_{pop}$  in the standard least squares estimator for the regression coefficients:  $\beta_{pop} = (X_{pop}X'_{pop})^{-1}X_{pop}Y_{pop}$ . These weighted expressions have been suggested

to be used for estimating population regression parameters for probability samples [11, 20]. According to [11], estimator 2.4 is asymptotically normal and consistent for the corresponding population parameter. Thus, to approximate the posterior draws of the parameters in a Bayesian synthesis approach, new values for the regression coefficients can be drawn from the normal distribution.

It's important to note that in the context of synthetic data, using the pseudo-likelihood approach would result in synthetic data that could be treated as simple random samples from the population since the parameter estimates obtained from 2.4 and 2.5 already account for the sampling design and are thus unbiased estimates of the true parameters in the population. For binary variables, the pseudo-likelihood function can be written as follows:

$$PL(\vec{\beta}; Y|X) = \prod_{i=1}^n \pi_i(\mathbf{x}_i)^{w_i y_i} (1 - \pi_i(\mathbf{x}_i))^{w_i(1-y_i)} \quad (2.6)$$

where  $\pi_i(\mathbf{x}_i) = \frac{1}{1+exp(\mathbf{x}_i'\vec{\beta})}$  is the probability of success for the binary variable  $y_i$ . By setting the partial derivatives of the pseudo-log-likelihood function with respect to  $\vec{\beta}$  equal to zero, we obtain the following estimating equations:

$$\sum_{i=1}^n w_i(1 - y_i)\mathbf{x}_i' = \sum_{i=1}^n w_i \frac{exp(-\mathbf{x}_i'\vec{\beta})}{1 + exp(-\mathbf{x}_i'\vec{\beta})}. \quad (2.7)$$

The regression coefficients have no closed-form solution in this case. Therefore, optimization techniques such as iteratively reweighted least squares should be used to find the numerical solution in this context. These are implemented in many software packages, including R. For categorical variables with  $J > 2$  categories, the pseudo-likelihood function is an extension of the binary expression from two to  $J$  components under the product sign of (2.6). The  $j$ -th component, for  $j \in (1 : J - 1)$  is

$$\pi_{ij}(\mathbf{x}_i) = \frac{exp(\mathbf{x}_i'\vec{\beta})}{1 + \sum_{l=1}^{J-1} exp(\mathbf{x}_i'\vec{\beta})} \quad (2.8)$$

where  $J$  is the reference category. Thus, a numerical solution for multinomial logistic regression based on the pseudo-likelihood function can be obtained using optimization methods as well.

## 2.2 Alternative Strategies to Account for Complex Sampling Designs

An alternative approach to account for the sampling design is to include the design information in the models and to synthesize the weights similarly to other numerical variables. This method is easy to implement using available software. We use the R package *synthpop* to synthesize all variables including the design information and the weights. However, this approach may not always be optimal as the type of relationship between the weights and outcome variables can affect the quality of synthesis. To evaluate the impact of potential model mis-specification we use both parametric and non-parametric approaches for synthesis. The latter one is based on using Classification And

Regression Trees (CART) [18]. The advantage of using CART compared to parametric synthesis is that it does not require to specify the functional form of the regression model, which can be especially helpful when trying to model the relationship between the survey variables and the survey weights.

### 3 Design of the Simulation Study

We conducted a simulation study using a repeated simulation design to explore the pseudo-likelihood approach within the sequential regression framework and compare it with other methods to account for informative sampling designs. The simulation consists of four steps: (i) generating the true population (ii) taking stratified random samples from this population using PPS sampling, (iii) generating synthetic data using the different modeling techniques described in the previous section, (iv) analyzing the synthetic datasets. Steps (ii) to (iv) were repeated  $s = 10,000$  times to be able to assess whether valid inferences can be obtained from the synthetic data. In the following we describe each of the simulation steps in detail.

#### 3.1 Generating the Population

We generated two populations that differed in how the weights  $W$  were related to the other variables. For both populations, the data consisted of two strata. The first stratum contained 1,000,000 records while the second stratum contained 10,000 records. Both populations consisted of four variables: two continuous variables ( $C1$  and  $C2$ ) and two binary variables ( $B1$  and  $B2$ ). Additionally, we generated a measure of size variable  $M$  that was used to compute the probabilities of selection for the PPS sampling design.

For the first population, denoted “Linear”, the relationship between  $W$  and the other variables was approximately linear. For the second population, denoted “Nonlinear”, the relationship between  $W$  and the other variables was nonlinear.

The variables in the “Linear” population were generated according to the following models:

$$\begin{aligned} M &\sim \text{Unif}(500, 600) \\ C1 &\sim N(\mu = 0.8 * M, \sigma = 15) \\ C2 &\sim N(\mu = 0.8 * C1, \sigma = 15) \end{aligned}$$

$$\begin{aligned} p(B1 = 1) &= 1/(1 + \exp(16 - 0.03 * M)) \\ p(B2 = 1) &= 1/(1 + \exp(3 - 7 * B1)) \end{aligned}$$

The variables in the “Nonlinear” population were generated according to the following models:

$$\begin{aligned} M &\sim \text{Unif}(100, 900) \\ C1 &\sim N(\mu = 0.8 * M, \sigma = 15) \\ C2 &\sim N(\mu = 0.8 * C1, \sigma = 15) \\ p(B1 = 1) &= 1/(1 + \exp(10 - 0.02 * M)) \\ p(B2 = 1) &= 1/(1 + \exp(3 - 7 * B1)) \end{aligned}$$

We note that the PPS sampling design generally introduces a nonlinear relationship between the weights and the variables (see the next section). However, in order to generate the Linear population, we deliberately picked a range of  $M$  in such a way that the relationship between  $W$  and the variables was still approximately linear. Increasing the range of  $M$  when generating the Nonlinear population ensured that this was no longer the case. Figure 3 in the appendix shows the relationships between the survey weights and  $C1$  for the two populations demonstrating the approximately linear relationship for the first population and the nonlinear relationship for the second population. Similar relationships were observed for the other variables (results omitted for brevity). For both designs, the variables  $C1, C2$ , and  $B1, B2$  were moderately to highly correlated with correlation coefficients (squared canonical correlations for binary variables) in the range of 0.5–0.9. We note that in our simulation, we always used the same data generating process for both strata. In practice, this would imply that there is little benefit from using a stratified sampling design unless the goal of stratification is to ensure sufficient sample sizes for each of the strata. We introduced the stratification only for methodological reasons as it allows us to vary the probability of selection between the two strata (see Sect. 3.2) to ensure that the sampling design matters.

### 3.2 Drawing the Sample

A sample of  $n = 1,000$  records was selected from each stratum using a stratified PPS sampling design, i.e., for records  $i, \dots, N_s$ , where  $N_s$  is the size of stratum  $s$  in the population, the probability of selection was set to  $\pi_i^s = nM_i/(N_s\bar{M}_s)$ , where  $\bar{M}_s = \sum_{N_s} M_i/N_s$ . Keeping  $n$  fixed for both strata implied that stratum two was oversampled relative to stratum one.

### 3.3 Generating the Synthetic Data

All three synthesis approaches, i.e., the pseudo-likelihood approach (**PL**), the approach that includes weights as linear predictors based on parametric models (**PWeights**), and the approach that uses CART models for the synthesis (**CARTWeights**) were implemented as described in Sects. 2.1 and 2.2. The weights were computed as the inverse of the probability of selection,  $w_i = 1/\pi_i^s$ . For each synthesizer, the synthesis models exactly matched the data generating process, i.e., the first model generated synthetic values for  $M$  by sampling with replacement from its marginal distribution. All other variables were generated based on the models described in Sect. 3.1. For **PWeights** and **CARTWeights** we synthesized the weights in the last step using all other variables as predictors. We decided to always rely on the correct model specification to ensure that differences in the results for the three synthesizers can be attributed solely to the different treatment of the survey weights. For each synthesis approach, we generated  $m = 10$  synthetic datasets.

### 3.4 Analyzing the Synthetic Data

We evaluated several statistics computed on the synthetic data. These included point and interval estimates for the means and proportions as well as for regression coefficients

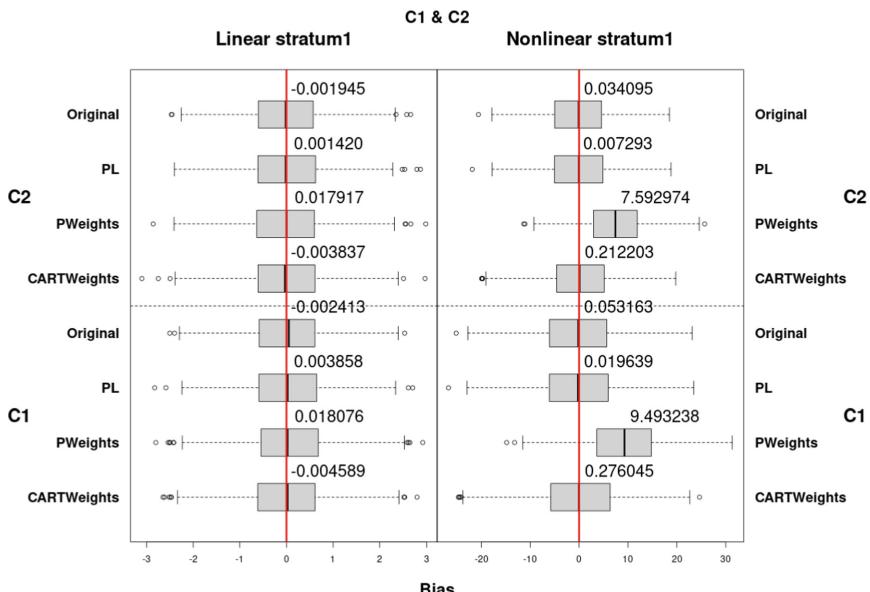
for different regression models. Specifically, we looked at biases in the point estimates compared to the true population values and coverage rates for 95% confidence intervals. These coverage rates are computed by calculating the confidence intervals for each simulation run and then evaluating how often the true population parameter lies inside these confidence intervals across the simulation runs. Assuming unbiased point and variance estimates, the confidence intervals should cover the true population parameters in approximately 95% of the simulation runs.

## 4 Simulation Results

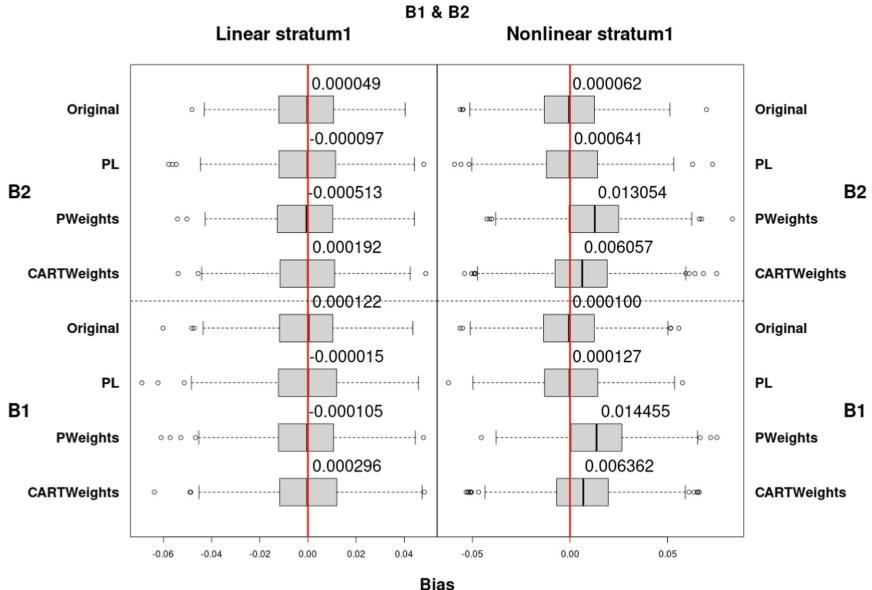
Due to space limitations, we include only the results for the means and proportions. Similar results were obtained for the regression coefficients. These results are excluded for brevity.

Figures 1 and 2 show boxplots of the differences between the estimated stratum means and proportions and their true values for each simulation run. If the estimators were unbiased, the boxplots should be centered around zero. The figures contain the results for the different synthesis approaches and for the original samples for Linear and Nonlinear populations for Stratum 1. Similar plots for Stratum 2 are presented in Figs. 4 and 5 in the Appendix.

The results in Figs. 1 and 2 show that the pseudo-likelihood approach has smaller biases for all variables compared to the PWeights and CARTWeights approach. However, biases tend to be negligible for all methods for the Linear scenario.



**Fig. 1.** Differences between the estimated means and true population means for continuous variables C1 and C2 from Linear and Nonlinear populations for Stratum 1. The numbers on top of each boxplot are the empirical biases.



**Fig. 2.** Differences between the estimated proportions and true population proportions for binary variables B1 and B2 from linear and nonlinear populations for Stratum 1. The numbers on top of each boxplot are the empirical biases.

In the Nonlinear scenario, biases increase for all the synthesizers, but also for the original data. However, the bias for the pseudo-likelihood approach is still comparable to the bias for the original sample, while biases increase substantially for the other methods. While **CARTWeights** still gives almost unbiased results for the continuous variables included in Fig. 1, the bias is more severe for the binary variables as shown in Fig. 2. The **PWeights** approach exhibits substantial bias for all variables. Figures 4 and 5 in the Appendix show similar results for Stratum 2.

These results are not surprising, because the parameters of the regression models for the pseudo-likelihood approach are unbiased estimates of the original population parameters, based on robust Horwitz-Thompson estimates as mentioned in Sect. 2.1. On the other hand, **PWeights** and **CARTWeights** create synthetic weights based on a regression model that only aims at preserving the relationships between the weights and the other variables. The methodology does not model the population parameters explicitly and one cannot expect to obtain unbiased estimates based on this strategy especially, if the relationship is mis-specified as is the case for the parametric modeling approach in the Nonlinear scenario. We also note that using CART models instead of parametric models reduced the bias as expected. However, considerable biases remain for the binary variables.

Coverage rates for the different synthesis strategies are reported in Table 1. The coverage rates are closer to the nominal coverage rates for the pseudo-likelihood approach

**Table 1.** Coverage rates for the means and proportions

Method	M	C1	C2	B1	B2
Linear Stratum1					
PL	0.95	0.97	0.98	0.98	0.98
PWeights	0.99	1.00	1.00	0.99	0.99
CARTWeights	0.99	0.99	0.99	0.99	0.99
Original	0.95	0.95	0.95	0.95	0.95
Linear Stratum2					
PL	0.95	0.97	0.97	0.98	0.98
PWeights	0.99	0.99	0.99	1.00	0.99
CARTWeights	0.99	0.99	0.99	0.99	0.99
Original	0.95	0.95	0.95	0.95	0.95
Nonlinear Stratum1					
PL	0.88	0.88	0.89	0.94	0.95
PWeights	0.95	0.95	0.96	0.98	0.99
CARTWeights	0.99	0.99	0.99	0.99	0.99
Original	0.95	0.95	0.95	0.95	0.95
Nonlinear Stratum2					
PL	0.88	0.88	0.89	0.93	0.94
PWeights	0.94	0.94	0.94	0.97	0.98
CARTWeights	0.99	0.99	0.99	0.99	0.99
Original	0.95	0.95	0.95	0.95	0.95

comparative to other methods for the Linear case in both strata, while both PWeights and CARTWeights show substantial overcoverage for all variables.

For the Nonlinear scenario, characterized by increased variability of the weights and nonlinear relationships between the weights and the other variables, the pseudo-likelihood approach shows some undercoverage for continuous variables, while PWeights leads to overcoverage for the binary variables. Finally, CARTWeights still exhibits overcoverage for all the variables. Note that both PWeights and CARTWeights show overcoverage for the binary variables despite the fact that Fig. 2 revealed substantial biases in the point estimates. This implies that the uncertainty in the estimates must be substantially overestimated based on these methods. One reason for this overestimation could be that the variance estimator of [15] has large variability and we obtained negative variance estimates in 48.26% (47.65%) of the simulation runs for PWeights (CARTWeights). For those runs we used the ad-hoc adjustment of [17], which ensures positive variance estimates, but overestimates the true variance. We also noted that the degrees of freedom for the  $t$ -distribution were less than four in 83.74% (84.23%) of the simulation runs for PWeights (CARTWeights) leading to excessively wide confidence intervals.

## 5 Concluding Remarks

This paper evaluates different strategies for synthesizing survey data under informative sampling designs. The pseudo-likelihood approach showed promising results in terms of data utility. All estimates were approximately unbiased and coverages were close to their nominal level although we saw some undercoverage for the nonlinear scenario. Additional research may be helpful to further validate this approach. Results for approaches that treat the survey weights as just another variable in the data show that this strategy is generally inferior to the pseudo-likelihood approach. The strategy crucially depends on the correct specification of the relationship between the survey weights and the variables of interest. As we saw for the nonlinear scenario, misspecification introduces bias in the estimates obtained from the synthetic data. A disadvantage of the pseudo-likelihood approach is that it is currently not implemented in data synthesis software such as synthpop. Implementing these procedures would be a fruitful endeavour for the future.

The paper only focuses on data utility and not on disclosure risk. Our primary goal was to evaluate, which strategies are most promising when generating synthetic data from survey data collected using an informative sampling design. Besides, for fully synthetic methods, assessing disclosure risk is challenging, as re-identification risk is usually not a concern. Furthermore, given the substantial differences in utility for the different approaches, we expect that the results for the risk of disclosure would be a mirror image of the utility results. The pseudo-likelihood approach would most likely show higher risks than the other methods because of the inverse relationship between risk and utility. Whether this is really the case would be an interesting area for future research.

All simulation studies presented in this paper ensured that the parametric models match the true data generating process. The motivation was to disentangle potential biases from not properly accounting for the survey design from biases that might arise due to model mis-specification. In practice, data synthesis will always be subject to model mis-specification and it would be interesting to evaluate the relative importance of these two bias components. A downside of the pseudo-likelihood approach is that it requires parametric modeling. Since CART models tend to be robust against model mis-specification, these models might still offer higher utility when considering both sources of bias. In this context, it would be interesting to explore whether proposals of CART models that account for the survey design [22] could also be used for data synthesis.

## Disclaimer

The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

**Acknowledgement.** The work of JD was supported by a grant from the German Federal Ministry of Education and Research (grant number 16KISA096) with funding from the European Union-NextGenerationEU.

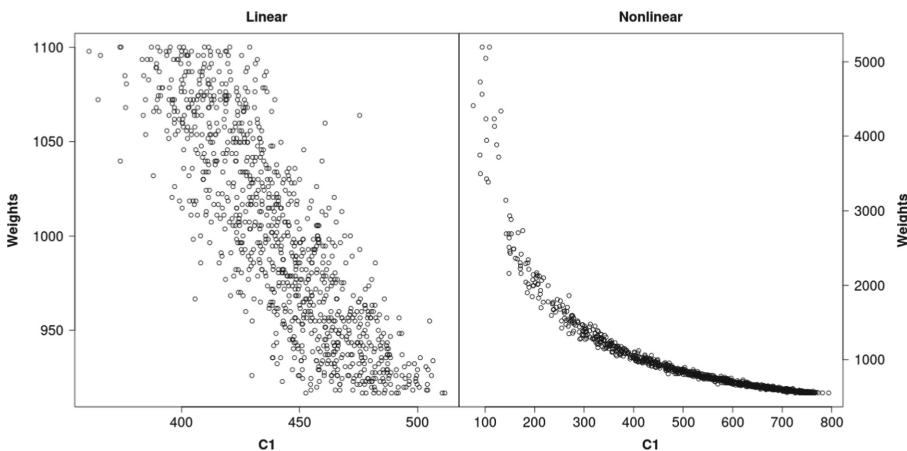
**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## Appendix

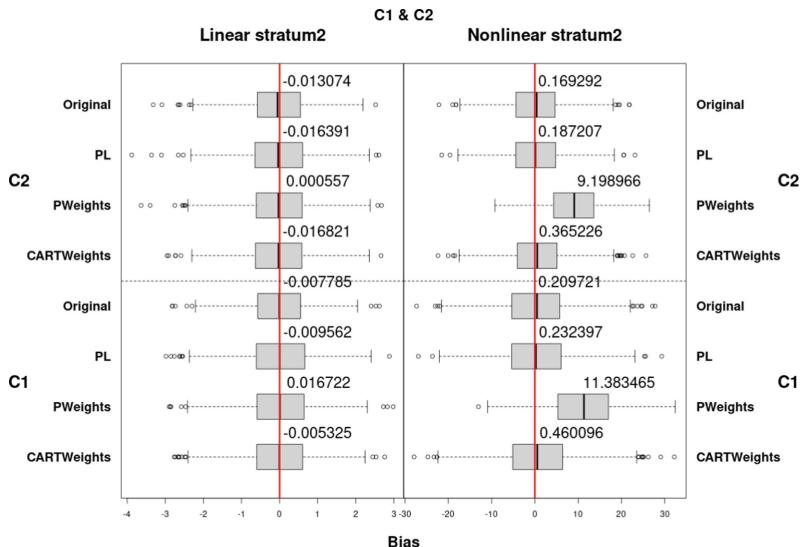
### A. Review of Inferential Procedures for Fully Synthetic Data.

To be able to account for the uncertainty associated with the synthesis (drawing random synthetic samples from  $P(Y|X)$ ),  $m$  synthetic sets are usually released. Each of them can be analyzed by standard methods, and then the results of  $m$  analyses are combined to produce estimates, confidence intervals, and test statistics that properly reflect the uncertainty of synthesis. With the original data, inference regarding the unknown parameter  $\Theta$  is typically based on some point estimate  $\theta$  and an estimate for its variance  $u$ . For the analysis of the imputed datasets, let  $\theta_i$  and  $u_i$  be the point and the variance estimates for each synthetic dataset. The combining rules for fully synthetic data for  $\Theta$  are based on the following components [15, 19]:  $\bar{\theta}_m = \sum_{i=1}^m \theta_i/m$ ,  $b_m = \sum_{i=1}^m (\theta_i - \bar{\theta}_m)^2/(m-1)$ , and  $\bar{u}_m = \sum_{i=1}^m u_i/m$ .

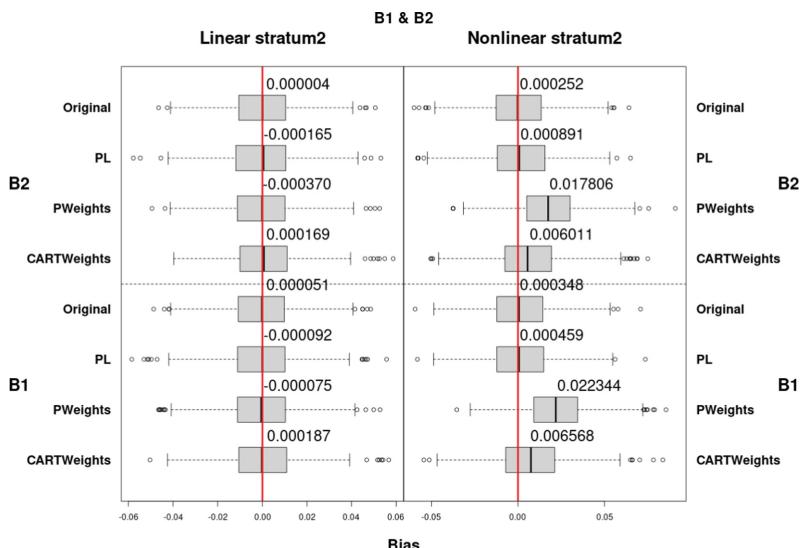
When the pseudo-likelihood approach is used to generate synthetic data, the sampling process is already accounted for by including the weights in the pseudo-likelihood. Thus, the uncertainty introduced by the sampling process is already reflected in the variability of the  $\theta_i$  between the different synthetic datasets. The variance estimator under this scenario is  $V_f^{pseudo} = (1 + \frac{1}{m})b_m$  ([10, 15]).



**Fig. 3.** Scatterplot of Weights vs C1 for the linear (left) and the nonlinear (right) population.



**Fig. 4.** Differences between the estimated means and true population means for continuous variables C1 and C2 from linear and nonlinear populations for Stratum 2. The numbers on top of each boxplot are the empirical biases.



**Fig. 5.** Differences between the estimated proportions and true population proportions for binary variables B1 and B2 from linear and nonlinear populations for Stratum 2. The numbers on top of each boxplot are the empirical biases.

## References

1. Confidential Information Protection and Statistical Efficiency Act Title V of the E-Governmental Act of 2002, Public Law 107-347. [www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/cipsea/cipsea\\_statute.pdf](http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/cipsea/cipsea_statute.pdf)
2. Dong, Q., Elliott, M.R., Raghunathan, T.E.: A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Pract.* **40**(1), 29 (2014)
3. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control. Theory and Implementation. No. 201 in Lecture Notes in Statistics. Springer, New York (2011). <https://doi.org/10.1007/978-1-4614-0326-5>
4. Drechsler, J., Haensch, A.C.: 30 years of synthetic data. *Stat. Sci.* **39**(2), 221–242 (2024)
5. Fienberg, S.E.: The relevance or irrelevance of weights for confidentiality and statistical analyses. *J. Priv. Confidentiality* **1**(2) (2010)
6. Fuller, W.A.: Sampling Statistics. Wiley, Hoboken (2009)
7. Godambe, V., Thompson, M.E.: Parameters of superpopulation and survey population: their relationships and estimation. *Int. Stat. Rev./Revue Internationale de Statistique* 127–138 (1986)
8. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **47**(260), 663–685 (1952)
9. Hu, J., Bowen, C.M.: Advancing microdata privacy protection: a review of synthetic data methods. *Wiley Interdisc. Rev.: Comput. Stat.* **16**(1), e1636 (2024)
10. Kim, H., Drechsler, J., Thompson, K.: Synthetic microdata for establishment surveys under informative sampling. *J. Roy. Stat. Soc. Ser. A* **184**, 255–281 (2021). <https://doi.org/10.1111/rssa.12622>
11. Lumley, T., Scott, A.: Fitting regression models to survey data. *Stat. Sci.* **32**(2) (2017)
12. Mitra, R., Reiter, J.P.: Adjusting survey weights when altering identifying design variables via synthetic data. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 177–188. Springer, Heidelberg (2006). [https://doi.org/10.1007/11930242\\_16](https://doi.org/10.1007/11930242_16)
13. Nowok, B., Raab, G.M., Dibben, C.: Synthpop: bespoke creation of synthetic data in R. *J. Stat. Softw.* **74**(11), 1–26 (2016). <https://www.jstatsoft.org/index.php/jss/article/view/v074i11>
14. Pfeffermann, D.: The role of sampling weights when modeling survey data. *Int. Stat. Rev./Revue Internationale de Statistique* 317–337 (1993)
15. Raghunathan, T., Reiter, J.P., Rubin, D.: Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* (2003)
16. Raghunathan, T.E., Solenberger, P.W., Van Hoewyk, J.: IVEware: imputation and variance estimation software. Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor (2002)
17. Reiter, J.P.: Satisfying disclosure restrictions with synthetic data sets. *J. Off. Stat.-Stockholm* **18**(4), 531–544 (2002)
18. Reiter, J.P.: Using cart to generate partially synthetic, public use microdata. *J. Off. Stat.* **21**(3), 441–462 (2005)
19. Rubin, D.: Statistical disclosure limitation. *J. Off. Stat.* **9**, 461–468 (1993)
20. Särndal, C., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer Series in Statistics. Springer, Heidelberg (2003)
21. Sugden, R., Smith, T.: Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495–506 (1984)
22. Toth, D., Eltinge, J.L.: Building consistent regression trees from complex sample data. *J. Am. Stat. Assoc.* **106**(496), 1626–1636 (2011)

23. United States Congress: The privacy act of 1974 (5 U.S.C. 552a). <http://www.justice.gov/opcl/privstat.htm>
24. United States Congress: Public health service act (42 U.S.C. 242m(d) section 308(d)) (1944). [http://www.law.cornell.edu/escode/html/uscode42/usc\\_sup\\_01\\_42\\_10\\_6A.html](http://www.law.cornell.edu/escode/html/uscode42/usc_sup_01_42_10_6A.html)



# The Production of Bespoke Synthetic Teaching Datasets Without Access to the Original Data

Mark Elliot<sup>1</sup>(✉) , Claire Little<sup>1</sup> , and Richard Allmendinger<sup>2</sup>

<sup>1</sup> School of Social Sciences, University of Manchester, Manchester M13 9PL, UK  
{mark.elliot,claire.little}@manchester.ac.uk

<sup>2</sup> Alliance Manchester Business School, University of Manchester,  
Manchester M13 9PL, UK  
richard.allmendinger@manchester.ac.uk

**Abstract.** Teaching datasets are a pivotal component of the data discovery pipeline. These datasets often serve as the initial point of interaction for data users, allowing them to explore the contents of a dataset and assess its relevance to their needs. However, there are instances where their viability is limited, particularly where source data is only accessible within restricted settings, such as trusted research environments (TREs). In response to this challenge, this paper proposes the production of synthetic datasets tailored for specific teaching purposes by utilising already cleared (and published) analyses as the basis for the synthesis. Unlike generic synthetic datasets, the datasets created are designed to solely reproduce the specific analyses. Crucially, the datasets can be generated without access to the original data. Two experiments with census data demonstrate the viability of the method and a live use case is described. Issues arising such as marginal disclosure risk are then discussed.

**Keywords:** Data Synthesis · Evolutionary Algorithms · Data Utility · Disclosure Risk

## 1 Introduction

Data Synthesis (DS) [13, 17] is a methodology within statistics and machine learning that produces an artificial dataset, that does not contain any real records but approximates the underlying data structure of the original (real) data whilst (hopefully) having low disclosure risk. DS can be used as a data confidentiality method to prevent leakage of confidential information about data subjects whilst delivering analytical utility equivalent to that of the original data. Data synthesis can therefore allow better access to information that might otherwise be safeguarded or restricted since it presents a lower disclosure risk than the original data. For a broad review of DS see [4].

Teaching datasets are a pivotal component of the data discovery pipeline. These datasets are often the initial exposure of users to the data, allowing them

to explore a dataset's contents and assess its relevance to their needs. Typically constituting compact subsets of the complete dataset, traditional teaching datasets employ data minimisation techniques to control disclosure risks.

Despite this crucial role in the data discovery process, there are instances where the viability of orthodox teaching datasets is limited, particularly in scenarios where the source data is only accessible within restricted settings, such as trusted research environments (TREs). Some TREs have attempted to address this limitation by creating generic synthetic datasets for teaching purposes. However, this approach encounters a significant challenge inherent in all general synthetic data – it may fail to accurately replicate the analytical results desired by trainers for their students. TREs can also be uncomfortable with assessing disclosure risk for synthetic data as it does not map onto standard output disclosure control rules (see e.g. [8])

In response to this challenge, this paper proposes an alternative: the production of bespoke synthetic datasets tailored for specific teaching purposes. The envisioned approach utilises already cleared and published analyses as the basis for the synthesis. This could be the trainer's own analyses or those published by a third party. Unlike generic synthetic datasets, the bespoke synthetic datasets are designed to solely reproduce the required analyses. The ultimate vision is to enable trainers to produce their own bespoke synthetic teaching datasets that look and feel like real datasets and can reproduce the required outputs faithfully. Crucially, the datasets can be generated without access to the original data.

Access to such bespoke synthetic datasets offers an opportunity for users to undergo training using realistic data before applying for access to the real restricted dataset. This may not only enhance the training experience, but also reduce the time required within a TRE. By introducing this innovative approach to DS, this work aims to redefine the landscape of data accessibility, offering enhanced training experiences and expanded opportunities for data exploration within the research data community.

## 2 Background

The primary objective of this paper is to investigate the feasibility of generating teaching datasets tailored to restricted data access scenarios. Building on earlier proof-of-concept work [5, 10, 12], we implement methods from population-based search, and in particular Evolutionary Algorithms (EAs) [16] to achieve this.

DS is a new application for the EA community. Initial proof-of-concept work [5, 10, 12] using census data has provided a formal definition of the problem (i.e., decision variables, constraints, and objective functions) and then investigated the suitability and robustness of off-the-shelf methods to tackle the problem. A key difference between our proposed EA method and existing data synthesis methods (both the statistical and machine learning variants) is that no access to the original dataset is required to create the synthetic data, making it suitable for our use case. We are not aware of other work that does this; however, we acknowledge that as early as 2003, Burridge [1] proposed a method called Information Preserving Statistical Obfuscation (IPSO), which generates

a dataset that reproduces the original analysis – but this was performed with access to the original data, unlike our approach.

EAs perform iterative optimisation using three main, biologically inspired operators: selection, crossover, and mutation. Briefly: an initial population of candidate solutions is specified. In our case, a candidate solution is a synthetic dataset (using synthetic datasets as candidates differs from existing EA methods that tend to use strings or vectors of data of the same type), and the fitness (quality) of the candidates is calculated using the objective function.

The parental selection operator is used to select candidates (parents) to reproduce for a new population, with fitter candidates more likely to be selected. The crossover operator combines some of the parents to produce new candidate solutions (children). The mutation operator then mutates some of the candidates (i.e., randomly changes some of the decision variables). The children or a combination of children and parents form the population of the next generation (this step is called environmental selection). This process is repeated multiple times (generations), using the fitness to guide it, with fitter solutions produced with each generation. Typically, the process terminates when a specified number of generations has been produced or a particular fitness level has been reached.

A useful feature of EAs is their flexibility - there are many parameters that can be changed or set, and the objective function can be designed for the specific purpose (one can also optimize multiple competing objective functions, also known as multi-objective optimization). In this study, the fitness function is designed to optimise the synthetic data such that it matches the desired analytical output. Previous work [2,3] has shown the feasibility of using EAs to generate synthetic microdata.

### 3 Research Design

#### 3.1 General Approach

The study reported here consists of two experiments (with a live use case described in Sect. 6).

In the case of two experiments, an output or set of outputs from a dataset are defined. The dataset is split into two (original and holdout) and the analyses are then run on the original dataset to produce the *reference outputs*.

Those reference outputs are then converted into objective functions for the EA, which produces the synthetic dataset. The EA is then run until it converges. The analytical output is then produced for the synthetic data and compared to the reference output and the equivalent output of the analyses applied to the holdout data. The holdout data simulates another sample from the same population as the original.

Since the method does not access the original dataset, it requires analytical output from the original data that has already been cleared as safe for publication. This could be coefficients of a regression model or summary statistics, for example. It also requires basic metadata; the names of each of the variables and the possible values they can take.

**Data.** The data used were a subset of the 1991 UK Census Sample of Anonymised Records [15]. Analytic output is based on a paper by Gardiner and Hill [7] - which examined car ownership of the elderly (aged  $\geq 50$ ) population of Sheffield. We have expanded this (to allow for a larger dataset) and include all of the UK county of South Yorkshire (which includes Sheffield). The base dataset consists of 8054 individuals aged 50 or over in the South Yorkshire area who had answered the car question (how many cars, if any, they had access to). This base dataset was split into an experimental dataset (henceforth referred to as the “original dataset”)<sup>1</sup> and a holdout dataset both containing 4027 records, with the holdout dataset not used at all in the creation of the synthetic data. The variables used were: AREAP (a geocode), AGE (in single years), SEX, ETHGROUP (ethnic group), LTILL (the presence of long-term illness), TENURE, CARS (number of cars that the respondent has access to). The univariate distributions of these variables are shown in Table 1.

**Table 1.** Univariate distributions of the variables used in the study derived from the original dataset.

Name	Value	Label	Proportion
AREAP	Barnsley	48	0.1716
	Doncaster	49	0.2210
	Rotherham	50	0.1838
	Sheffield	51	0.4236
AGE	50–95		
SEX	Male	1	0.4527
	Female	2	0.5473
ETHGROUP	White	1	0.9878
	Other	2	0.0122
LTILL	Yes	1	0.3700
	No	2	0.6300
TENURE	Own occ. outright	1	0.3650
	Own occ. Buying	2	0.2131
	Rented priv. furn.	3	0.0055
	Rented job/business	5	0.0117
	Rented housing assoc.	6	0.0204
	Rented local authority	7	0.3486
CARS	No access	0	0.4726
	Access to $\geq 1$	1	0.5274

<sup>1</sup> we use the term original here rather than the more orthodox “training” to make it clear that we are not training a model on the data.

## 4 Experiment 1: A Single Output

For experiment 1 we focus on just one analytical output - a logistic regression model. To generate the reference output, the logistic regression was performed on the original dataset, using a binary version of CARS as the target variable (where 0 = no access to cars and 1 = access to 1 or more cars). The reference level for the explanatory variables are: from the Sheffield (AREAP = 51) area, white (ETHGROUP = 1), male (SEX = 1), no long-term illness (LTILL = 2) and living in a home that is owned outright (TENURE = 1). The Model coefficients for the reference output are shown in Table 2.

**Table 2.** Logistic regression model coefficients generated from the original data in Experiment 1, where CARS (access to cars) is the target variable.

Parameter	Coefficient
Constant	6.4034
AGE	-0.0787
AREAP_48	-0.0239
AREAP_49	-0.0398
AREAP_50	-0.0650
SEX_2	-0.5501
ETHGROUP_2	-0.8254
LTILL_1	-0.2033
TENURE_2	0.1788
TENURE_3	-1.6399
TENURE_4	-1.7013
TENURE_5	-0.7950
TENURE_6	-1.8219
TENURE_7	-1.9224

### 4.1 Method

The EA used the coefficients from the logistic regression as its objectives and calculated the mean squared error (MSE) between the original coefficients and coefficients generated using the synthetic data to form the fitness measure.<sup>2</sup> The EA was run for 2000 generations, with no crossover and a gradually decreasing

---

<sup>2</sup> For these experiments we used simple unweighted MSEs (within each output). We acknowledge that other alternatives could have been used and in a separate of experiments we have used standardised coefficients instead - the results were broadly similar. The optimum weighting methodology will be decided by intensive benchmarking work - the goal here was to examine what could be achieved with a very simple algorithm.

mutation rate (this has been found to be optimal in other experiments), starting at 0.01 and decreasing every 250 generations (to 0.005, 0.001, 0.0005, 0.0001, …). The population size was 24 (i.e. after the selection process for each generation there are 24 candidate synthetic datasets) and the initial population was generated using the uniform distribution. When a value is mutated, its replacement is drawn from the univariate distribution (from Table 1). An elitist selection strategy was used in which the best children/parents of each generation form the next generation. Five randomly initialised (with a different random seed each time) runs were performed.

## 4.2 Results

Table 3 shows the regression coefficients for the model produced by the synthetic data in each of the five runs. As can be seen in the table, in each run of the EA the data have converged on a solution very close to the original dataset.

In Table 4 we compare the results obtained by the EA synthesised dataset (from run 1) with those of the holdout dataset. The coefficients for the EA synthesised dataset are much closer to those of the original dataset than those of the holdout dataset are. To be clear about what this implies: *this method produces more accurate reproduction of the required analytical properties than a second sample drawn from the same population (using the same sampling mechanism)*. For reference, Table 4 also shows the results obtained by a general synthetic dataset produced from the original (using CART in synthpop [14] with the default settings).

## 5 Experiment 2: Multiple Outputs

In the second experiment, we added two additional outputs: cross-tabulations between the CARS variable and two other variables (TENURE and ETH-GROUP); these can be found in Tables 5 and 6. This mimics exploratory analysis that might be conducted as part of the training exercise before the model itself is run.

### 5.1 Method

As in Experiment 1, the EA was run for 2000 generations, without crossover and a gradually decreasing mutation rate, starting at 0.01 and decreasing every 250 generations (to 0.005, 0.001, 0.0005, 0.0001, …). There were 24 synthetic datasets in the population, and the initial population was generated using the uniform distribution. When mutating a value, its replacement is drawn from the univariate distribution (from Table 1). An elitist strategy was used, in which the optimal children/parents of each generation form the next.

The EA used the three outputs as objectives and calculated the mean squared error (MSE) between those and the outputs generated using the synthetic data. A weighted sum of the three outputs drove the EA. Ten different weightings were

**Table 3.** A comparison of model coefficients for the original and synthetic data for five runs of the synthesiser.

Parameter	Original output coefficient	Synthetic data coefficient				
		Run 1	Run 2	Run 3	Run 4	Run 5
Constant	6.4034	6.4034	6.4032	6.4035	6.4036	6.4037
AGE	-0.0787	-0.0787	-0.0752	-0.0745	-0.0731	-0.0743
AREAP_48	-0.0239	-0.0239	-0.0239	-0.0241	-0.0238	-0.0239
AREAP_49	-0.0398	-0.0398	-0.0397	-0.0398	-0.0400	-0.0398
AREAP_50	-0.0650	-0.0650	-0.0649	-0.0646	-0.0646	-0.0650
SEX_2	-0.5501	-0.5501	-0.5500	-0.5498	-0.5500	-0.5504
ETHGROUP_2	-0.8254	-0.8254	-0.8253	-0.8252	-0.8256	-0.8254
LTILL_1	-0.2033	-0.2033	-0.2032	-0.2033	-0.2034	-0.2035
TENURE_2	0.1788	0.1788	0.1789	0.1786	0.1787	0.1789
TENURE_3	-1.6399	-1.6399	-1.6404	-1.6394	-1.6394	-1.6407
TENURE_4	-1.7013	-1.7031	-1.7033	-1.7033	-1.7030	-1.7025
TENURE_5	-0.7950	-0.7950	-0.7949	-0.7948	-0.7950	-0.7949
TENURE_6	-1.8219	-1.8219	-1.8217	-1.8219	-1.8217	-1.8217
TENURE_7	-1.9224	-1.9224	-1.9225	-1.9226	-1.9227	-1.9220
Mean absolute error between run and original output:	0.00037	0.00049	0.00059	0.00053	0.00053	

**Table 4.** A comparison of model coefficients for the original, EA synthetic, a general synthetic dataset and the holdout dataset.

	Regression coefficients			
	Original	EA synthetic	General synthetic	Holdout
Constant	6.4034	6.4034	6.6741	5.7832
AGE	-0.0787	-0.0787	-0.0813	-0.0707
AREAP_48	-0.0239	-0.0239	-0.1057	-0.0345
AREAP_49	-0.0398	-0.0398	-0.1767	-0.1533
AREAP_50	-0.0650	-0.0650	-0.2197	0.1968
SEX_2	-0.5501	-0.5501	-0.5176	-0.6311
ETHGROUP_2	-0.8254	-0.8254	0.8265	-0.8528
LTILL_1	-0.2033	-0.2033	-0.2442	-0.0914
TENURE_2	0.1788	0.1788	-0.0391	0.3443
TENURE_3	-1.6399	-1.6399	-1.7678	-1.8901
TENURE_4	-1.7013	-1.7031	-1.6062	-1.5852
TENURE_5	-0.7950	-0.7950	-0.6090	-0.5415
TENURE_6	-1.8219	-1.8219	-2.0131	-1.8200
TENURE_7	-1.9224	-1.9224	-1.9847	-1.8996
Mean absolute error between dataset and original:	0.0004	0.2323	0.1460	

**Table 5.** Cross-tabulation of tenure and access to car derived from original dataset.

TENURE	Frequency		Proportion	
	No access to car	Access to car	No access to car	Access to car
Own occ. outright	470	1000	0.117	0.248
Own occ. Buying	162	696	0.04	0.173
Rented priv. furn.	15	7	0.004	0.002
Rented priv. unfurn.	105	39	0.026	0.01
Rented job/business	22	25	0.005	0.006
Rented housing assoc.	64	18	0.016	0.004
Rented local authority	1065	339	0.264	0.084

**Table 6.** Cross-tabulation of ethnic group and access to car derived from original dataset.

ETHGROUP	Frequency		Proportion	
	No access	Access to car	No access to car	Access to car
White	1880	2098	0.467	0.521
Other	23	26	0.006	0.006

tried: all equal and then slowly decreasing the weighting given to the logistic regression output and increasing the weighting for the two table outputs (they are labelled as run 1–10), the weightings are listed in Appendix A.<sup>3</sup>.

## 5.2 Results

Unlike the first experiment, the runs of the EA differ by varying by the weights of the different objectives. In Table 7 we compare the results obtained by the EA synthetic dataset for runs, 1, 3, 5, 7 and 10 with each other, the holdout dataset and a general synthetic dataset. As can be seen, regardless of the setting of the weight parameter, the EA synthetic data are closer to the original data than the holdout dataset in reproducing the model. Although they are further from the original than in experiment 1 - particularly with run 1 - the differences are still markedly less than those that would be produced by a second equivalent sample and would therefore we assume, be acceptable (for the use case).

Equivalent results for the cross-tabulations can be found in Table 8 and 9. Here run 10 outperforms the others - unsurprisingly, given it was weighted to optimising these outputs.

---

<sup>3</sup> Since the MSE of the regression results was higher than those for the table outputs, weighting it lower had the effect of avoiding the regression dominating the outcome.

**Table 7.** Comparison of logistic model output of the original data with runs 1, 3, 5, 7 and 10 of the EA synthetic data and the holdout data.

	Original	EA Run 1	EA Run 3	EA Run 5	EA Run 7	EA Run 10	Holdout	General
Constant	6.4034	6.4016	6.4074	6.3980	6.3944	6.3215	5.7832	6.6741
AGE	-0.0787	-0.0729	-0.0745	-0.0754	-0.0760	-0.0773	-0.0707	-0.0813
AREAP_48	-0.0239	-0.0201	-0.0236	-0.0293	-0.0222	-0.0650	-0.0345	-0.1057
AREAP_49	-0.0398	-0.0379	-0.0371	-0.0388	-0.0339	-0.0779	-0.1533	-0.1767
AREAP_50	-0.065	-0.0650	-0.0631	-0.0662	-0.0655	-0.0922	0.1968	-0.2197
SEX_2	-0.5501	-0.5539	-0.5496	-0.5491	-0.5477	-0.3885	-0.6311	-0.5176
ETHGROUP_2	-0.8254	-0.8201	-0.8114	-0.8223	-0.8228	-0.8061	-0.8528	0.8265
LTILL_1	-0.2033	-0.2043	-0.2019	-0.2052	-0.2086	-0.2708	-0.0914	-0.2442
TENURE_2	0.1788	0.1739	0.1790	0.1795	0.2029	0.6984	0.3443	-0.0391
TENURE_3	-1.6399	-1.6366	-1.6330	-1.6359	-1.6308	-1.6154	-1.8901	-1.7678
TENURE_4	-1.7013	-1.7023	-1.6993	-1.7092	-1.7009	-1.7676	-1.5852	-1.6062
TENURE_5	-0.795	-0.7934	-0.7914	-0.7876	-0.7984	-0.8221	-0.5415	-0.609
TENURE_6	-1.8219	-1.8145	-1.8208	-1.8200	-1.8252	-1.9509	-1.82	-2.0131
TENURE_7	-1.9224	-1.9276	-1.9308	-1.9452	-1.9704	-2.2264	-1.8996	-1.9847
Mean absolute error between dataset & original:	0.0034	0.0037	0.0048	0.0085	0.1078	0.1460	0.2323	

**Table 8.** A comparison of the proportion of respondents with access to a car conditioned on tenure in the original data, runs 1, 3, 5, 7 and 10 of the EA synthetic data, general synthetic and the holdout data.

TENURE	Original	EA Run 1	EA Run 3	EA Run 5	EA Run 7	EA Run 10	Holdout	General Synth.
Own occ. outright	0.6795	0.6985	0.6923	0.6805	0.6785	0.6796	0.6582	0.6929
Own occ. Buying	0.8122	0.7436	0.7291	0.7300	0.7164	0.8028	0.8145	0.7952
Rent priv. furn.	0.3333	0.4444	0.3871	0.3846	0.3846	0.3750	0.2500	0.2500
Rent priv. unfurn.	0.2778	0.4035	0.3800	0.3462	0.3617	0.3056	0.2703	0.2903
Rent job/business	0.5455	0.5758	0.5313	0.5517	0.5357	0.5385	0.5625	0.5455
Rent housing assoc.	0.2000	0.3571	0.3830	0.3333	0.3429	0.2857	0.2273	0.1739
Rent local authority	0.2414	0.3458	0.3205	0.3079	0.2914	0.2399	0.2281	0.2443
Mean absolute error between dataset & original:	0.0880	0.0755	0.0584	0.0621	0.0247	0.0246	0.0222	

## 6 Discussion

The results described above are compelling. They show how it is possible to produce usable synthetic data that is able to replicate multiple analytical outputs at a level better than that produced by another sample from the same population. This then provides an in principle mechanism for trainers to produce their own teaching datasets with access to the original data.

**Table 9.** A comparison of the proportion of respondents with access to car conditioned on tenure in the original data, to runs 1, 3, 5, 7 and 10 of the EA synthetic data, general synthetic and the holdout data.

ETHGROUP	Original	EA Run 1	EA Run 3	EA Run 5	EA Run 7	EA Run 10	Holdout	General Synth.
White	0.5273	0.5613	0.5480	0.5359	0.5273	0.5273	0.5162	0.5254
Other	0.5000	0.4430	0.4247	0.4394	0.4286	0.4615	0.5385	0.7143
Mean absolute error:	0.0455	0.0480	0.0346	0.0357	0.0192	0.0248	0.1081	

We were also interested in testing how easy it would be to produce an actual teaching dataset of sufficient quality to be used by a trainer in practice. One of the requirements this threw up was the addition of other statistical properties beyond a single model (further descriptive statistics and weights). Working with Administrative Data Research UK we developed a teaching dataset of the UK linked ASHE-Census dataset<sup>4</sup>. The dataset produced was generated on the basis of outputs in the methodology report produced by the team that developed the ASHE-Census dataset [6]. The requirements were that the data should be able to reproduce a linear regression model predicting income, the univariate frequencies of five variables (sex, ethnicity, education level and disability, UK born) and the conditional means and medians of the income variable on those five variables. One additional step was required; the addition of weights. For the purposes of providing a realistic training experience, a weight variable needed to be added to mimic what the trainees would find when they eventually had access to the real data in the TRE. The weight was created by calibrating the synthetic dataset to census microdata.

The synthetic dataset was successfully produced to the satisfaction of the trainer and was first used in a training course run by the UK's National Centre for Research Methods in April 2024<sup>5</sup>.

## 6.1 Disclosure Risk

One question that might arise when considering this method is what about the disclosure risk? Returning to experiment 1, we carried out some analyses which were helpful indicators.

Firstly, the most obvious place for an adversary to attack this data would be to use the explanatory variables in an attempt to disclose the values of the response variable in the model used to construct the data. We used the TCAP statistic [9–11] to assess this. See Table 10.

<sup>4</sup> See: <https://www.adruk.org/news-publications/news-blogs/new-linked-dataset-available-to-provide-insights-into-earnings-and-employment-in-england-and-wales/>.

<sup>5</sup> <https://www.ncrm.ac.uk/training/show.php?article=13306>.

**Table 10.** TCAP values for the various datasets used in experiment 1 given the target is the response variable CARS and the keys are the explanatory variables.

Dataset comparison	Raw TCAP	Calibrated TCAP
Original -> Original	1.000	1.000
Original -> EA run 1	0.733	0.464
Holdout -> EA run 1	0.719	0.437
Original -> General Synthetic	0.777	0.552
CAP Baseline	0.502	0.000

TCAP assesses the probability of an adversary that links a known population unit to some data correctly inferring the target attribute for that population given that the l-diversity = 1, given the values of key variables. The baseline value is that provided when the adversary makes a random draw from the univariate distribution of the target (essentially guessing). Unsurprisingly, the synthetic data provides a better inferability than guessing. However, there are two things to note here. The adversary is almost as likely to be able to make correct inferences about population units that are **not** in the data that have been used to produce the analytical output (as represented by the holdout data) than about those that are represented in that data. The TCAP value here therefore represents the inference risk arising from the model itself. This is intuitive. Since we have used the model and no other information to generate the synthetic data, it makes sense that any risk identified is the risk arising from the model itself. However, we can dig a little deeper here. It is possible that the nature of an evolutionary algorithm means that secondary analytical properties may emerge from the process. We can examine this by using TCAP to treat every other variable in the dataset as a possible target. This analysis is shown in Table 11.

The key take away from this is that the synthetic data is barely more informative than the baseline and in this case is slightly more informative about the holdout data than the original data. Essentially, there is no emergent risk in the data beyond that which is already present in the model itself. Since in our scenario here the model will have already cleared and published, we can say that in this case the synthetic data would present negligible marginal risk. Note that this is not a general conclusion about this mechanism - data would have to be assessed on a case by case basis. Particularly where there are multiple outputs, it is possible to imagine cases where unanticipated marginal risk might emerge.

**Table 11.** TCAP values for the various datasets used in experiment 1 given the target is each of the explanatory variables and the keys are the remaining variables in the dataset.

	Target variable:						Mean
	AREAP	AGE	SEX	ETHGROUP	LTILL	TENURE	
Original -> Original	1	1	1	1	1	1	1
Original -> EA run 1	0.296	0	0.543	0.953	0.583	0.367	0.457
Holdout -> EA run 1	0.299	0	0.516	0.949	0.576	0.377	0.453
Original -> General Synth.	0.481	0.194	0.682	0.984	0.744	0.569	0.609
Baseline	0.292	0.029	0.504	0.976	0.534	0.302	0.440

## 6.2 Concluding Remarks

The paper has presented a new approach to generating synthetic data based on analytical output without requiring access to real data. The particular use case here is the production of teaching dataset, and for this it appears very promising.

In terms of the experiments reported here, the engaged reader will have noted that the outputs did not include standard errors, fit measures, deviance scores and so on that one would typically expect in standard statistical output. Understanding the impact of multiple outputs on the data quality and how to manage complexity of that with adaptive mutation, diversity management and other more advanced algorithms is the focus of our current work.

However, the method we have demonstrated also opens up another, perhaps more intriguing, possibility. By embedding an analytical output in a synthetic dataset the methods opens up the possibility of formalising the assessment of disclosure risk for analytical outputs from safe settings. At present, this process is typically managed through the combination of the application of rules and the evaluation of a human output checker, and so moving this on to more formal grounds (perhaps in the form of a toolkit for output checker to use) would be an advance. But the potential is that by embedding the output in synthetic data it is possible to assess the risk using standard microdata disclosure control methods as we have done here. This is further element of our current work.

A second possible extension of this work is the synthesising of data from unlinked sources. It is a modest extension of the experiment 2 above to imagine multiple datasets (perhaps drawn from the same population) from which analytical outputs have been produced being quasi-linked through the production of a joint synthetic dataset without anyone even having sight of both datasets.

In future work, we will be exploring both of these possibilities as well as considering the application of the method reported here to the production of a wider range of teaching datasets.

**Acknowledgments.** This study was funded by the Economic and Social Research Council (grant numbers ES/Z502984/1 and ES/T000066/1). We also thank Lucy Stokes of the National Institute of Economic and Social Research for her input into the ASHE-Census live test.

**Code.** The code for the experiments reported here can be found at: <https://github.com/clairelittle/psd2024-bespoke-synthetic-datasets>

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## Appendices

**Table 12 in Appendix A**

**Table 12.** Objective function weights for the outputs used in Experiment 2.

Run	Output 1 weight	Output 2 weight	Output 3
1	0.33	0.33	0.33
2	0.25	0.375	0.375
3	0.2	0.4	0.4
4	0.15	0.425	0.425
5	0.1	0.45	0.45
6	0.05	0.475	0.475
7	0.025	0.4875	0.4875
8	0.01	0.495	0.495
9	0.001	0.4995	0.4995
10	0.0001	0.49995	0.49995

## References

- Burridge, J.: Information preserving statistical obfuscation. *Stat. Comput.* **13**, 321–327 (2003). <https://doi.org/10.1023/A:1025658621216>
- Chen, Y., Elliot, M., Sakshaug, J.: Genetic algorithms in matrix representation and its application in synthetic data. In: UNECE Worksession on Statistical Confidentiality 2017 (2017). [https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/2.Genetic\\_algorithms.pdf](https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/2.Genetic_algorithms.pdf)
- Chen, Y., Elliot, M., Smith, D.: The application of genetic algorithms to data synthesis: a comparison of three crossover methods. In: Domingo-Ferrer, J., Montes, F. (eds.) PSD 2018. LNCS, vol. 11126, pp. 160–171. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-99771-1\\_11](https://doi.org/10.1007/978-3-319-99771-1_11)
- Drechsler, J., Haensch, A.C.: 30 years of synthetic data (2023). <https://doi.org/10.48550/arXiv.2304.02107>
- Elliot, M., Little, C., Allmendinger, R.: Do samples taken from a synthetic microdata population replicate the relationship between samples taken from an original population? In: UNECE Expert Meeting on Statistical Data Confidentiality 2023 (2023). [https://unece.org/sites/default/files/2023-08/SDC2023\\_S4\\_5\\_UnivManchester\\_Elliot\\_D.pdf](https://unece.org/sites/default/files/2023-08/SDC2023_S4_5_UnivManchester_Elliot_D.pdf)

6. Forth, J., Phan, V., Ritchie, F., Whittard, D., Stokes, L., Bryson, A., Singleton, C.: ASHE - census 2011 data linkage: user Guide for Drop 2 of the ASHE-Census 2011 Dataset (2022). <https://www.wagedynamics.com/wp-content/uploads/2023/01/ASHE-CEW11-User-Guide-Version-2.1-Drop-2.pdf>
7. Gardiner, C., Hill, R.: Analysis of access to cars from the 1991 UK census samples of anonymised records: a case study of the elderly population of sheffield. *Urban Stud.* **33**(2), 269–281 (1996)
8. Griffiths, E., C., G., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A., Woods, C.: Handbook on Statistical Disclosure Control for Outputs (2019). [https://ukdataservice.ac.uk/app/uploads/thf\\_datareport\\_aw\\_web.pdf](https://ukdataservice.ac.uk/app/uploads/thf_datareport_aw_web.pdf)
9. Jennifer, T., Mark, E.: The synthetic data challenge. In: Conference of European Statisticians (2019). [https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019\\_S3\\_UK\\_Synthetic\\_Data\\_Challenge\\_Elliott\\_AD.pdf](https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthetic_Data_Challenge_Elliott_AD.pdf)
10. Little, C., Elliot, M., Allmendinger, R.: Comparing the utility and disclosure risk of synthetic data with samples of microdata. In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 234–249. Springer, Cham (2022)
11. Little, C., Elliot, M., Allmendinger, R.: Synthetic census microdata generation: a comparative study of synthesis methods examining the trade-off between disclosure risk and utility. *J. Official Stat.* (2024)
12. Little, C., Elliot, M., Allmendinger, R., Samani, S.S.: Generative adversarial networks for synthetic data generation: a comparative study. In: Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality (2021). [https://unece.org/sites/default/files/2021-12/SDC2021\\_Day2\\_Little\\_AD.pdf](https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Little_AD.pdf)
13. Little, R.J.A.: Statistical analysis of masked data. *J. Official Stat.* **9**(2), 407–426 (1993)
14. Nowok, B., Raab, G., Dibben, C.: Synthpop: bespoke creation of synthetic data in R. *J. Stat. Softw.* **74**(11), 1–26 (2016). <https://doi.org/10.18637/jss.v074.i11>
15. Office for National Statistics, Census Division, University of Manchester, Cathie Marsh Centre for Census and Survey Research: Census 1991: Individual Sample of Anonymised Records for Great Britain (SARs) (2013). <https://doi.org/10.5255/UKDA-SN-7210-1>
16. Reeves, C., Rowe, J.E.: Genetic Algorithms: Principles and Perspectives: A Guide to GA Theory, vol. 20. Springer, New York (2002)
17. Rubin, D.B.: Statistical disclosure limitation. *J. Official Stat.* **9**(2), 461–468 (1993)

# **Synthetic Data Generation Software**



# A Comparison of SynDiffix Multi-table Versus Single-table Synthetic Data

Paul Francis<sup>(✉)</sup>

Max Planck Institute for Software Systems (MPI-SWS), Saarbrücken, Germany  
[francis@mpi-sws.org](mailto:francis@mpi-sws.org)

**Abstract.** SynDiffix is a new open-source tool for structured data synthesis. It has anonymization features that allow it to generate multiple synthetic tables while maintaining strong anonymity. Compared to the more common single-table approach, multi-table leads to more accurate data, since only the features of interest for a given analysis need be synthesized. This paper compares SynDiffix with 15 other commercial and academic synthetic data techniques using the SDNIST analysis framework, modified by us to accommodate multi-table synthetic data. The results show that SynDiffix is many times more accurate than other approaches for low-dimension tables, but somewhat worse than the best single-table techniques for high-dimension tables.

## 1 Introduction

In recent years there has been a lot of interest in synthetic structured data for statistical disclosure control. The US National Institute of Standards and Technology (NIST) has a research program [5] to better understand the utility and privacy of data deidentification mechanisms. They have developed the SDNIST software tool and several sample datasets to analyze and compare the utility and privacy of synthetic data mechanisms<sup>1</sup>. NIST has so far archived the test results of nearly 30 different techniques from more than a dozen organizations.

There are a variety of use cases for synthetic data, including for instance data enhancement and generating test data. These use cases do not necessarily require that the synthetic data very accurately mimics the original data. Indeed sometimes the goal is to modify the statistics of the original data, for instance to remove bias or increase certain profiles.

The primary use case for SDNIST, however, is statistical disclosure. The goal is to replicate the original data as accurately as possible while preserving anonymity. The original datasets supplied by SDNIST come from US Census American Community Survey (ACS) data, and the SDNIST utility metrics measure how closely the synthetic data matches the original data.

Most synthetic data techniques are designed to produce *single-table* datasets—all columns of the original dataset are synthesized to produce a single

<sup>1</sup> <https://github.com/usnistgov/sdnist>.

synthesized dataset. An advantage of single-table approaches is ease of use. The single table can be pulled into any data analysis tool and manipulated exactly as with the original data. An important disadvantage of single-table approaches, however, is that data accuracy degrades with increased columns.

An alternate approach is to make *multi-table* datasets. Each table synthesizes different column combinations, ranging from single-column tables to all columns. For instance, suppose there are 50 columns in the data, and the analyst is interested in the correlation between two columns. In the single-table approach, nominally all 50 columns would be synthesized, and the two columns of interest taken from the 50-column synthesized table. In a multi-table approach, the analyst would use the table with only those two columns.

In the multi-table approach thousands of tables can easily be generated. Multi-table is therefore only viable if it is strongly anonymous despite multiple tables. Prior synthetic data mechanisms do not have this characteristic. The more tables that are generated from the same data, the more privacy is lost. It is unknown how many tables can be built before anonymity is dangerously degraded, because existing systems are not designed to be used in multi-table mode, and therefore such testing is unnecessary.

Although multi-table is certainly less convenient than single-table, it is common practice among statistics offices to release data as multiple tables, each with a relatively small number of columns.

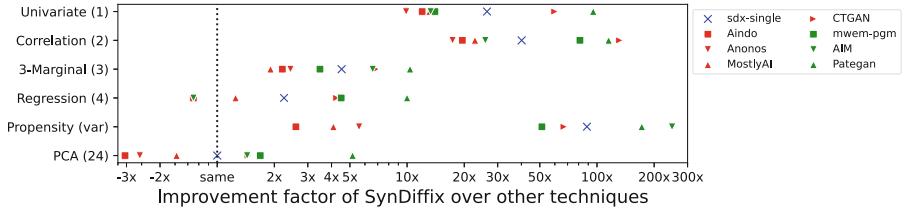
SynDiffix [8] is a new structured synthetic data generator that remains strongly anonymous no matter how many tables are generated (Sect. 2). This paper uses SDNIST to compare the utility and privacy of SynDiffix against 15 other mechanisms, each of which have been submitted to SDNIST by its respective developers. We show that SynDiffix is many times more accurate than the best alternative techniques for low-dimensional tables (see Fig. 1). Compared to the next best techniques, which are proprietary commercial products, SynDiffix' median measure is  $10\times$  more accurate for single-column measures,  $17\times$  more accurate for 2-column measures, and  $2\times$  more accurate for 3-column measures.

That advantage degrades for higher-dimension measures. For a 4-column measure (linear regression), SynDiffix is more accurate than most techniques, but  $30\%$  less accurate than the best techniques. For the 24-column measure (PCA), SynDiffix is again more accurate than most techniques, but  $3\times$  less accurate than the best technique.

We also show that, under SDNIST's privacy metric, SynDiffix has very strong anonymity. While most of the mechanisms also have strong anonymity, SynDiffix has even stronger anonymity than generative approaches, and only slightly weaker anonymity than the differential privacy mechanisms.

The contributions of this paper are:

- A broad comparison of utility for SynDiffix' multi-table synthesis with 15 single-table synthetic data techniques using SDNIST.
- An improved interpretation of SDNIST's privacy metric, taking into account the statistical baseline of the data.
- An enhancement of the SDNIST tool to accommodate multi-table approaches.



**Fig. 1.** Improvement factor of SynDiffix over other techniques for each measure (number of measured columns). Techniques with insufficient anonymity or fewer than 24 columns in synthetic table are not comparable and are therefore excluded. Measures with a negative improvement factor (left of the dashed line) are more accurate than SynDiffix. Measures greater than 300× are not shown. Note log scale.

**Limitations:** While the measures in this paper do allow for a direct comparison of techniques, they do not measure utility for actual use cases. SynDiffix’ superior (or inferior) accuracy may or may not be important for any given use case. This paper only measures one dataset. Measures with other datasets, for a subset of the techniques in this paper, can be found in [8], and reinforce the findings of this paper.

## 2 Overview of SynDiffix

Here we give a brief overview of SynDiffix. A full description can be found at [8].

SynDiffix operates by building multi-dimensional search trees from the original data, and then assigning synthetic data from the nodes of the search trees. SynDiffix builds a family of trees with different dimensions: all one-dimension trees, all two-dimension trees, up to a single tree with all columns. Lower-dimension trees have better precision, while higher-dimension trees better capture the relationship between attributes. SynDiffix uses information from all trees to synthesize data. If the original data has too many columns to build all combinations of all dimensions, SynDiffix partitions the original table into multiple lower-dimension tables, synthesizes each table, and then merges them back together.

SynDiffix’ anonymity derives from how it assigns nodes in the trees. The nodes themselves incorporate three core anonymization features, *range snapping*, *sticky noise*, and *aggregation*.

**Range Snapping:** Unlike most search trees, which partition ranges by splitting them in half or in a way that balances the tree, SynDiffix forces ranges to conform to a fixed set of sizes and offsets. Specifically, ranges conform to a power-of-two sequence ( $\dots, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, \dots$ ), and likewise for offsets (a range of 2 can fall on offsets  $\dots, -2, 0, 2, 4, \dots$ ). Text and datetime values are converted to numbers prior to tree building, and converted back when values are assigned from the tree nodes.

**Aggregation:** All nodes in all trees are aggregates of multiple individuals. Nodes without enough individuals are pruned from the trees. The corresponding rows appear higher in the tree. For time-series data, SynDiffix must be configured to identify which rows belong to which individuals.

**Sticky Noise:** Noise is added to each node’s row count. The noise is sticky in that for any given node, as defined by the ranges and offsets of each attribute, the amount of noise is always the same.

Taken together, these three mechanisms lead to strong anonymization. Aggregation ensures that no individual data is released. Range snapping and sticky noise ensures that any given data value will always appear in the same limited set of aggregates with the same noise, no matter how many synthetic tables are generated. In this sense, SynDiffix is similar to the Cell Key Method used by the statistics offices of Australia [19], Germany [10], and the UK [1]. SynDiffix, however, generalizes the approach to work automatically, including with time-series data, and to produce synthetic data rather than tabular data.

### 3 Setup

The SDNIST synthetic data measurement tool works with three datasets. Each is composed of 24 attributes from the 44-question ACS [4]. For this study, we use the NATIONAL dataset, which consists of 27253 rows over 20 PUMA regions [2] across the USA.

The 16 techniques taken from SDNIST and used in this paper are listed in Table 1. Five of the techniques use differential privacy (DP), colored green. Several of these use generative modeling, and five other techniques use generative modeling without DP, colored red. Five techniques do not synthesize all 24 columns, and we regard four techniques as having excessively weak anonymity (Sect. 4.1). As these are not directly comparable, they have lighter color shades so that they can be distinguished in the various plots.

For SynDiffix, we generated 455 separate tables, one for each column combination required by SDNIST measures. These include 24 1-column tables, 181 2-column, 232 3-column, one 4-column, 16 9-column, and one 24-column table. This amounts to 1254 total columns, making the storage requirements for SynDiffix 52 times that of a single table. We used the default settings for SynDiffix’ average suppression threshold (5 rows), absolute suppression threshold (3 rows), and per-bucket noise standard deviation (1.4). It took just under 4 h to generate all 455 synthetic tables on a windows laptop with 32G RAM and an Intel i7-7820HQ CPU running at 2.9 GHz.

We additionally show the results for SynDiffix as a single table (‘sdx-single’). This allows us to directly see the benefits of a multi-table approach, as well as allows us to compare SynDiffix with other techniques on an apples-to-apples *usability* basis.

**Table 1.** Set of compared techniques, showing the number of columns synthesized (out of 24), and whether or not anonymization is weak. Technique labels link to the SDNIST report. Techniques without an epsilon do not use differential privacy. Techniques without both a repo and citation are proprietary.

Technique	Tech	Org	Cols	Weak	$\epsilon$	Cite	Repo
SynDiffix	K-dimension search trees	Open Diffix	24		[8]	<a href="#">link</a>	
sdx-single	K-dimension search trees	Open Diffix	24		[8]	<a href="#">link</a>	
Aindo	Generative model	Aindo	24				
Anonos	Generative model	Anonos	24				
MostlyAI	Generative model	MostlyAI	24				
CTGAN	Generative model	SDV	24		[20]	<a href="#">link</a>	
YData	Generative model	YData	24	X			<a href="#">link</a>
mwem-pgmn	Graphical models + DP	See pub	24		1	[14]	
AIM	Workload adaptive + DP	OpenDP	24		10	[13]	<a href="#">link</a>
Pategan	Generative model + DP	See pub	24		10	[11]	<a href="#">link</a>
Genetic	Approximate DP	See pub	21		10	[12]	<a href="#">link</a>
Sarus	Generative model + DP	Sarus	10		10	[6]	
CART	Decision trees	Synthpop	21			[16]	<a href="#">link</a>
K6-Anon	K-anonymity	sdcMicro	10			[18]	<a href="#">link</a>
PRAM	Random value changes	sdcMicro	10	X		[15]	<a href="#">link</a>
SMOTE	Minority oversampling	See pub	24	X		[21]	<a href="#">link</a>
Sample40	Simple sampling	US Census	24	X		[3]	

### 3.1 Changes to SDNIST

The SDNIST tool assumes a single synthetic table. In order to measure SynDiffix, we modified SDNIST to handle multiple tables. The Github repo with these modifications is <https://github.com/yoid2000/SDNist-multi>. Prior to each measurement operation, SDNIST-multi fetches the synthetic table that has the same columns required by the measurement. Other changes required by SDNIST-multi are discussed in the relevant sections.

## 4 SDNIST Measure Results

This section presents the results of the SDNIST measures. The code used to produce these results is <https://github.com/yoid2000/sdnist-summary>.

A summary of most of the utility measures is shown in Fig. 1, which give the *improvement factor* IF of SynDiffix (*sdx*) over the other techniques (*alt*). IF is computed as:

$$\text{IF} = \begin{cases} \delta_{alt}/\delta_{sdx}, & \text{if } \delta_{alt} \geq \delta_{sdx} \\ -\delta_{sdx}/\delta_{alt}, & \text{otherwise} \end{cases} \quad (1)$$

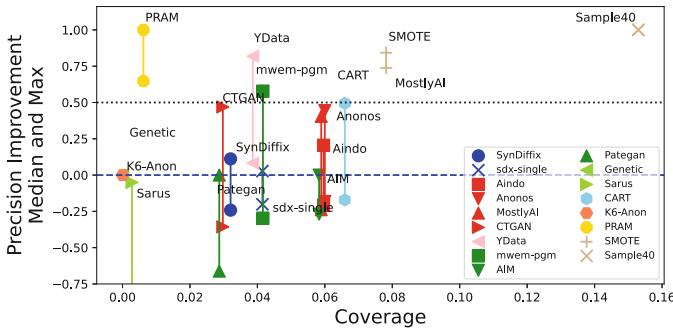
where  $\delta_x = |S_{perfect} - S_{actual}|$ , the absolute difference between a perfect score and the actual score for the technique  $x$ . The intuition here is that, compared to 10% error, 5% error represents an IF of  $2\times$ , and 2.5% error has an IF of  $4\times$ . A negative value is used to denote the case where the other technique have a better score than SynDiffix.

The measures in Fig. 1 are ordered by lowest to highest dimensionality, with univariate measuring only one column and PCA being a measure based on all 24 columns. Figure 1 excludes non-comparable techniques: those that either do not synthesize all 24 columns, or have weak anonymity. Those measures are however included in all other results.

Figure 1 shows that SynDiffix improves over other techniques by multiple factors for low-dimensional measures, but that the advantage degrades and is even reversed somewhat for high-dimensional measures.

#### 4.1 Privacy

To measure privacy, SDNIST emulates an attack whereby the attacker knows eight quasi-identifying attributes of an individual (EDU, SEX, RAC1P, PUMA, OWN\_RENT, INDP\_CAT, HISP, MSP), finds a record in the synthetic data that uniquely matches the quasi-identifier, and infers that the remaining attributes are those of the individual.



**Fig. 2.** Precision Improvement (PI) and Coverage where the attacker knows the quasi-identifiers of the target, finds a record with a unique and complete match of the quasi-identifiers, and infers an unknown attribute from that record. PI below 0.0 has no privacy loss whatsoever. PI below 0.5 has strong anonymity.

In a multi-table setting, this attack is most effective if the synthetic table contains the quasi-identifying columns plus only one additional column; the inferred column. Therefore, in measuring this attack for SynDiffix, we generated 16 separate 9-column tables, one for each inferred column.

Although SDNIST measures the precision of this attack for each column, it does not correctly interpret the privacy loss associated with the precision measure. In its summary reports (linked from Table 1), SDNIST claims that a 50%

average precision (“percentage of matched records”) represents strong privacy. The problem with this interpretation is that it does not take into account the statistical baseline of the data itself. 50% inference precision represents substantial privacy loss for an attribute that occurs in say only 1% of the population (i.e. total income **PINCP**), but no privacy loss for an attribute that occurs in 50% of the population (i.e. **SEX**).

In order to correctly account for the baseline, we measure *Precision Improvement* (PI). This is the improvement in precision above the statistical baseline [9]. PI is measured as:

$$\text{PI} = (P_{atk} - P_{base}) / (1 - P_{base}) \quad (2)$$

where  $P_{atk}$  and  $P_{base}$  are the precision measures for the attack and baseline respectively.

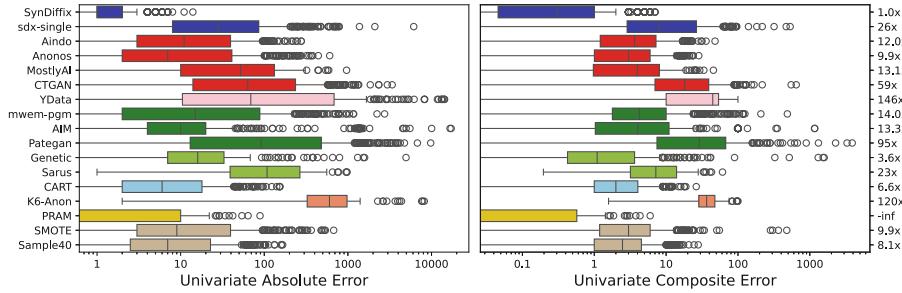
To compute  $P_{base}$ , we exploit the premise, taken from differential privacy, that a data release that does not include a given individual cannot be regarded as having compromised the privacy of that individual [7]. Given this, we can compute  $P_{base}$  by splitting the original dataset into *training* and *test* datasets, training an ML model using the quasi-identifier columns as input features, and predicting the value of each target feature in the test dataset. Because the test dataset is not part of the training dataset, any predictions made on the test dataset from the training dataset does not compromise the privacy of individuals in the test dataset. The resulting precision is therefore a privacy-neutral baseline  $P_{base}$ .

We computed  $P_{base}$  for each target column using `sklearn LogisticRegression` with `penalty='l1'`, `C=0.01`, `solver='saga'`, and `max_iter=100`. The mean  $P_{base}$  is 0.61 with a standard deviation of 0.37. The max  $P_{base}$  has perfect precision. This is for the column **DENSITY**, which gives the population density of a **PUMA** area, is non-personal, and is entirely predicted by **PUMA**. Given these relatively high baseline precisions, the suggestion of SDNIST that 50% average precision is safe turns out to be too conservative.

The results are shown in Table 2 under *QI Match*. Figure 2 plots the median and max PI for each technique. Coverage is the fraction of rows that are unique quasi-identifier matches.  $\text{PI} > 0$  leaks some amount of privacy. We believe that  $\text{PI} < 0.5$  can conservatively be regarded as anonymous (even at full Coverage).  $\text{PI} < 0.5$  gives an attacker substantial uncertainty, and gives individuals substantial deniability, relative to the baseline.

From Fig. 2, we see that most of the techniques fall below  $\text{PI} = 0.5$ , and so can be regarded as anonymous. Four techniques are well above  $\text{PI} = 0.5$ , and so we believe that these can be regarded as not adequately anonymous. These are labeled as ‘weak’ in Table 1.

In addition to the above inference precision measure, SDNIST counts the number of unique full records that are common to both the original and synthetic data. These results are shown in Table 2 as *Full Match*. This does not represent a very meaningful measure of privacy, because there is nothing special about all columns being unique matches versus some fraction of columns (i.e. quasi-



**Fig. 3.** Absolute and composite error for univariate (single feature) counts. The composite error is the minimum of the absolute error and the percent relative error. The values give the median composite error Improvement Factor (IF) for SynDiffix. Box plots show 0, 25, 50, 75, and 100 percentiles plus outliers. Note log scale.

identifiers as above or any other set of columns) being unique matches. One can find thousands of unique matches among subsets of columns. What is important is whether inferences well above a precision baseline can be made. Nevertheless, SDNIST tabulates the information, and so we repeat it for completeness.

## 4.2 Univariate Accuracy

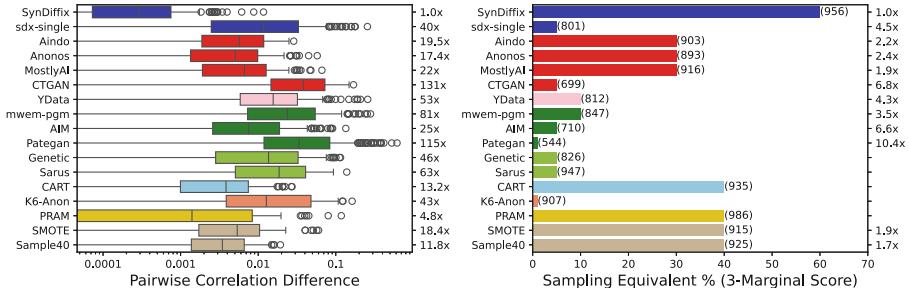
SDNIST measures the count of each feature value (the number of records in which each feature value appears). We define univariate absolute error for each feature value as  $E_{abs} = |C_o - C_s|$ , where  $C_o$  is the count in the original data, and  $C_s$  is the count in the synthetic data. We also define *composite error* as  $E_{comp} = \min(E_{abs}, E_{rel})$ , where  $E_{rel} = 100 * |C_o - C_s|/C_o$ , the percent relative error. Composite error reflects the fact that even large absolute errors can be small relative errors, and vice versa.

Figure 3 shows the univariate results, with selected corresponding values in Table 3. Compared to all other algorithms except PRAM, SynDiffix is multiple times more accurate. (PRAM unfortunately has unacceptable privacy, see Sect. 4.1.) The values on the right edge of the composite error plot gives the IF for the median composite error.

SynDiffix is a full order of magnitude more accurate than the best comparable model, the proprietary product Ananos.

## 4.3 Pairs Accuracy (Correlations)

SDNIST measures pairwise correlations and computes the difference between the original and synthetic data. The left plot of Fig. 4 displays this difference for every feature pair for the Kendall Tau correlation measure. It also shows the median IF. The corresponding numbers are given in Table 3. Here we see that SynDiffix is 17 $\times$  more accurate than the best comparable technique, the proprietary product Ananos.



**Fig. 4.** Accuracy of pairwise correlations and 3-marginals. The left plot gives the difference between the original and synthetic data for the Kendall Tau correlation coefficient, and corresponding improvement factors (right y-axis). The right plot gives the sampling rate over the original data that would be required to match the 3-marginal accuracy of the synthetic data. The right y-axis is the improvement factor of the 3-marginal accuracy.

#### 4.4 3-Marginal Accuracy

SDNIST measures the accuracy of a random selection of 3-feature combinations (3-marginals). SDNIST measures density difference of 232 3-marginals (out of a possible 2024, given 24 features), and derives an average across all measures which it calls the k-marginal score. The score ranges from 0 (no match) to 1000 (perfect match). SDNIST doesn't report the individual measures.

SDNIST also measures k-marginal scores for the original data given different sampling rates. By comparing the k-marginals for synthetic data and sampled original data, SDNIST establishes an equivalence data quality between synthesis and sampling.

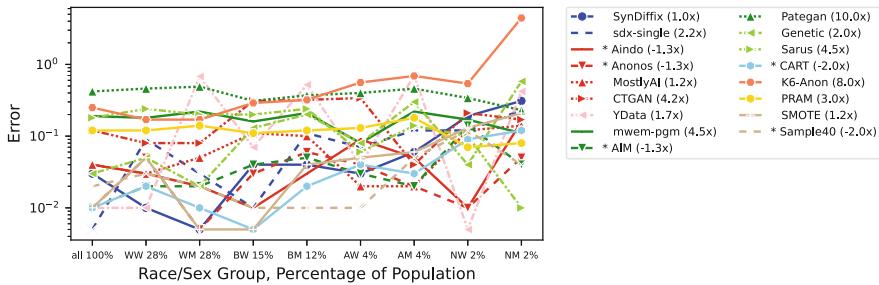
The right-hand plot of Fig. 4 shows the sampling equivalent and corresponding k-marginals score (number in parentheses). By this measure, SynDiffix is equivalent to a 60% sampling rate. The best generative models have a 30% equivalent sampling rate, and the best differential privacy techniques have a 10% equivalent sampling rate. Note that the k-marginal scores are not directly comparable between synthetic datasets with a different number of features. In particular, the k-marginal score for PRAM with 10 features (986) exceeds that of SynDiffix with 24 features (956) even though the equivalent sampling rate of PRAM is lower.

#### 4.5 Linear Regression (Four Features)

SDNIST measures the quality of a linear regression looking at the relationship between educational attainment and relative salary for different Race/Sex groups. An important attribute of a linear regression is the slope, which reflects the relationship between the two variables. We are therefore interested in the amount of error between the slope computed for the original data, and that of the synthetic data. We measure error for a single regression as  $E_{reg} = |S_o - S_s|$ ,

where  $S_o$  is the slope for the original data, and  $S_s$  is the slope for the synthetic data.

Figures 5 plots the error  $E_{reg}$  for each of the Race/Sex groups for each technique. It also gives the improvement factor of the median error for SynDiffix over the other techniques, where improvement factor is computed as in Sect. 4.2.



**Fig. 5.** Accuracy of the slope of a linear regression on educational attainment and salary. The two-letter code is Race/Sex, where Race is **W**hite, **B**lack, **A**sian, or **N**ative, and Sex is **M**en or **W**omen. The legend gives the improvement factor of SynDiffix over other techniques. Items marked with '\*' have less error than SynDiffix. Note the y axis is log scale.

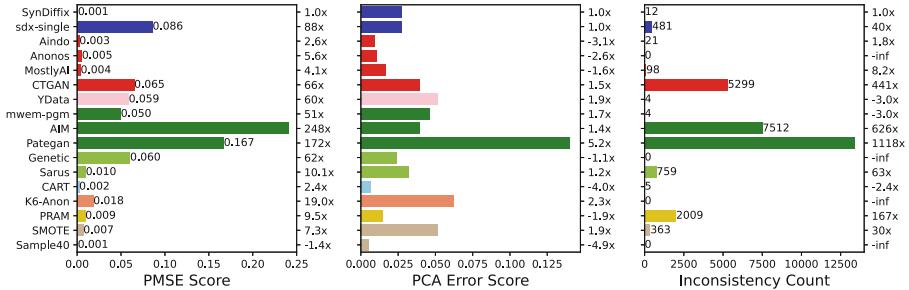
As Fig. 5 shows, three comparable techniques have better median accuracy than SynDiffix, by 30%. CART, which is not comparable because it synthesizes 21 rather than 24 columns, has better accuracy by 2 $\times$ .

While most of SynDiffix’ regressions have less than 6% error, the two “Native” regressions have substantially more error, roughly 10% and 20%. The fact that they perform less well is not surprising: the population is smaller (Native/Men NM has 395 out of total 23k records), and more fragmented (“Native” is actually three separate categories, Hawaiian, Alaskan, and Indian). These two regressions, however, also perform poorly relative to other techniques. As future work, we wish to explore why this might be.

#### 4.6 Propensity Mean Square Error

The Propensity Mean Square Error [17] (PMSE) measures data quality by training a classifier to distinguish between the original and synthetic data. The worse the classifier performance, the higher the synthetic data quality. A PMSE of 0 occurs when the synthetic data perfectly matches the original data.

Snoke et al. proposes that with multi-table synthetic data, a separate PMSE is computed for every table, and the average is taken as the PMSE [17]. We follow that approach here.



**Fig. 6.** Propensity MSE score, PCA error score, and Inconsistency Count. The right y-axes show improvement factor. Details in Table 4.

Figure 6 gives the results (with details in Table 4). SynDiffix has the lowest PMSE score of all techniques except for pure sampling (Sample40). Note that the PMSE score of sdx-single, which measures only the 24-column table, is worse than all of the generative models.

#### 4.7 Principle Component Analysis (All Features)

As a means of measuring synthetic data quality across all features, SDNIST runs a Principal Component Analysis (PCA) on the full dataset. SDNIST generates 5 PCs of 5 features each, and then generates a scatterplot visualization of each of 20 PC pairs. SDNIST zooms in on one particular PC pair (PC-0 and PC-1), highlighting individuals that satisfy a given constraint (marriage status of None versus not None). Note that this is the only SDNIST measure that uses the full 24-column SynDiffix table.

Unfortunately, SDNIST releases only images of the PCs, not the data itself or any accuracy measures. We therefore added a feature to our modified SDNIST implementation to compare the accuracy of the synthetic PCs with the original. Specifically, we compare each of the five original and synthetic PCs using the Kolmogorov-Smirnov score from `scipy.stats ks_2samp`, and present the average of these scores in Fig. 6 and Table 4.

We also show the original and synthetic scatterplots side-by-side for in Fig. 7. Both visually and from the scores we see that the comparable commercial generative techniques are more accurate than SynDiffix, with Aindo being 3x more accurate than SynDiffix. In particular, the best scatterplots effectively capture four distinct clusters, where SynDiffix merges two of the clusters together. SynDiffix is more accurate than the DP techniques as well as the open-source generative tool CTGAN.

#### 4.8 Inconsistencies

SDNIST detects 30 distinct data inconsistencies: combinations of two data values that are impossible. An example of an inconsistency is that a child (age less

than 15) can't be a disabled military veteran (columns `AGEP` and `DVET`). Each inconsistency is based on two columns, and we use the 2-column tables for this measure.

The results are shown in Fig. 6. SynDiffix has only 12 inconsistencies. Of the comparable techniques, only Ananos and mwem-pgm have fewer (0 and 4 respectively).

## 5 Discussion and Conclusion

This paper compares SynDiffix, a new open-source tool for generating structured synthetic data, with 15 other techniques (both commercial and open source), for both privacy and utility metrics. SynDiffix is different from other approaches in that it can safely generate multiple tables, each focused on a given analytic goal. This makes SynDiffix suitable for census data, which commonly releases data as multiple tables.

Using the SDNIST measurement tool modified by us to work with multi-table data, we show that SynDiffix is many times more accurate than other techniques for low-dimension measures, but somewhat worse than the best approaches for high-dimension measures. This suggests that one approach statistics offices may take is to use SynDiffix to release multiple accurate low-dimensional tables, and use a generative technique to release a single full table suitable for ML applications.

While the present study allows for a direct comparison between techniques, it does not evaluate the efficacy of SynDiffix for real analytic use cases. Doing so is an important step in determining whether SynDiffix can serve as a replacement for existing statistical disclosure methods. We hope that the present study motivates work in this direction.

## A Additional data

This appendix supplies additional data for the various measures. The software that produced these is at <https://github.com/yoid2000/sdnist-summary>.

For Table 2, *count* is the number of unique matches between the original and synthetic data (for full matches and quasi-identifying matches respectively). The percent of full matches relative to the total number of rows is show as %. *PI* is the inference precision improvement over the statistical baseline. Both median and max *PI* are shown. The coverage (*cov*) is the median fraction of unique quasi-identifying matches relative to all rows. The QI Match count for SynDiffix is the median count.

**Table 2.** Summary table for privacy measures.

	Full Match (Sect. 4.1)			QI Match (Sect. 4.1)		
	count	%	med PI	max PI	cov	count
SynDiffix	4	0.01	-0.31	0.11	0.031	840.0
sdx-single	4	0.01	-0.34	0.03	0.042	1131.0
Aindo	10	0.04	-0.34	0.21	0.060	1625.0
Anonos	5	0.02	-0.27	0.44	0.060	1635.0
MostlyAI	5	0.02	-0.30	0.40	0.059	1606.0
CTGAN	0	0.00	-0.59	0.47	0.030	811.0
YData	453	1.66	-1.17	0.82	0.039	1053.0
mwem-pgm	0	0.00	-0.26	0.58	0.042	1135.0
AIM	0	0.00	-0.38	0.00	0.058	1589.0
Pategan	0	0.00	-2.15	0.00	0.029	782.0
Genetic	5	0.02	0.00	0.00	0.000	0.0
Sarus	2087	7.66	-2.27	-0.05	0.003	77.0
CART	654	2.40	-0.31	0.49	0.066	1797.0
K6-Anon	7073	25.95	0.00	0.00	0.000	0.0
PRAM	10160	37.28	0.42	1.00	0.006	169.0
SMOTE	4219	15.48	0.62	0.84	0.078	2131.0
Sample40	10860	39.85	0.94	1.00	0.153	4169.0

For Table 3,  $N$  is the number of datapoints used for the corresponding boxplots, and  $med$  is the median error (composite in the case of Univariate).  $IF$  is the improvement factor of the median error. The equivalent sampled original table percentage is  $samp$ .

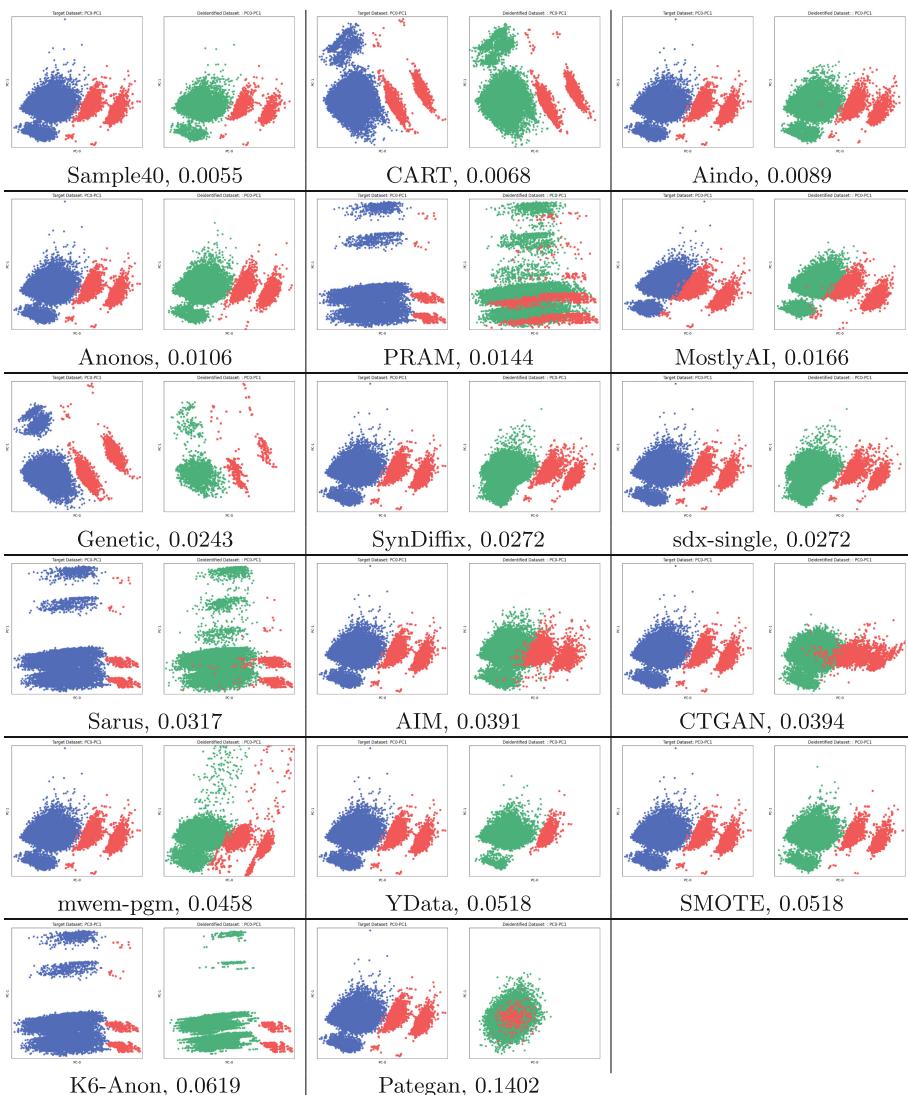
For Table 4,  $pmse$  and  $err$  are the PMSE and PCA errors respectively.  $count$  is the number of inconsistencies.  $IF$  are the corresponding improvement factors.

**Table 3.** Summary table for low-dimensional accuracy measures.

	Univariate (§4.2)			Correlation (§4.3)			3-marginals (§4.4)			
	N	med	IF	N	med	IF	score	IF	samp	IF
SynDiffix	499	0.3	1.0x	190	0.0003	1.0x	956	1.0x	60%	1.0x
sdx-single	510	8.0	26x	190	0.0117	40x	801	4.5x	5%	2.4x
Aindo	499	3.6	12.0x	190	0.0057	19.5x	903	2.2x	30%	1.8x
Anonos	499	3.0	9.9x	190	0.0051	17.4x	893	2.4x	30%	1.8x
MostlyAI	216	4.0	13.1x	190	0.0066	22x	916	1.9x	30%	1.8x
CTGAN	502	18.1	59x	190	0.0383	131x	699	6.8x	5%	2.4x
YData	499	44.6	146x	153	0.0156	53x	812	4.3x	10%	2.2x
mwem-pgm	499	4.3	14.0x	190	0.0237	81x	847	3.5x	10%	2.2x
AIM	499	4.0	13.3x	190	0.0075	25x	710	6.6x	5%	2.4x
Pategan	505	29.0	95x	190	0.0336	115x	544	10.4x	1%	2.5x
Genetic	202	1.1	3.6x	190	0.0136	46x	826		5%	2.4x
Sarus	77	7.2	23x	55	0.0186	63x	947		5%	2.4x
CART	451	2.0	6.6x	171	0.0038	13.2x	935		40%	1.5x
K6-Anon	77	36.7	120x	55	0.0127	43x	907		1%	2.5x
PRAM	77	0.0	-inf	55	0.0014	4.8x	986		40%	1.5x
SMOTE	499	3.0	9.9x	190	0.0054	18.4x	915	1.9x	40%	1.5x
Sample40	499	2.5	8.1x	190	0.0035	11.8x	925	1.7x	40%	1.5x

**Table 4.** Summary table for Propensity MSE, PCA Error, and Inconsistencies.

	PMSE (§4.6)		PCA Error (§4.7)		Inconsistencies (§4.8)	
	pmse	IF	ks-score	IF	count	IF
SynDiffix	0.0010	1.0x	0.0272	1.0x	12	1.0x
sdx-single	0.0859	88x	0.0272	1.0x	481	40x
Aindo	0.0025	2.6x	0.0089	-3.1x	21	1.8x
Anonos	0.0054	5.6x	0.0106	-2.6x	0	-inf
MostlyAI	0.0040	4.1x	0.0166	-1.6x	98	8.2x
CTGAN	0.0649	66x	0.0394	1.5x	5299	441x
YData	0.0588	60x	0.0518	1.9x	4	-3.0x
mwem-pgm	0.0498	51x	0.0458	1.7x	4	-3.0x
AIM	0.2412	248x	0.0391	1.4x	7512	626x
Pategan	0.1670	172x	0.1402	5.2x	13423	1118x
Genetic	0.0603	62x	0.0243	-1.1x	0	-inf
Sarus	0.0098	10.1x	0.0317	1.2x	759	63x
CART	0.0023	2.4x	0.0068	-4.0x	5	-2.4x
K6-Anon	0.0184	19.0x	0.0619	2.3x	0	-inf
PRAM	0.0093	9.5x	0.0144	-1.9x	2009	167x
SMOTE	0.0070	7.3x	0.0518	1.9x	363	30x
Sample40	0.0007	-1.4x	0.0055	-4.9x	0	-inf



**Fig. 7.** Original (left) and synthetic (right) scatterplot and average Kolmogorov-Smirnov score for all principle components, ordered by most-to-least accurate.

## References

1. Office of National Statistics (ONS) . Protecting personal data in Census 2021 results (2021). <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/protectingpersonaldataincensus2021results>

2. US Census Bureau . US Census Bureau Geographic Entities and Concepts . <https://www.census.gov/content/dam/Census/data/developers/geoareaconcepts.pdf>
3. US Census Bureau . ACS PUMS Files: The Basics (2020). [https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs\\_pums\\_handbook\\_2021\\_ch01.pdf](https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf)
4. US Census Bureau . The American Community Survey (2020). <https://www2.census.gov/programs-surveys/acs/methodology/questionnaires/2020/quest20.pdf>
5. US National Institute of Standards and Technology (NIST). Collaborative Research Cycle 2023 (2023). [https://pages.nist.gov/privacy\\_collaborative\\_research\\_cycle/](https://pages.nist.gov/privacy_collaborative_research_cycle/)
6. Canale, L., Grislain, N., Lothe, G., Leduc, J.: Generative modeling of complex data. arXiv preprint [arXiv:2202.02145](https://arxiv.org/abs/2202.02145) (2022)
7. Dwork, C.: Differential privacy. In: ICALP (2006)
8. Francis, P., Berneanu, C., Gashi, E.: Synduffix: more accurate synthetic structured data. arXiv preprint [arXiv:2311.09628](https://arxiv.org/abs/2311.09628) (2023)
9. Francis, P., Wagner, D.: Towards more accurate and useful data anonymity vulnerability measures. arXiv preprint [arXiv:2403.06595](https://arxiv.org/abs/2403.06595) (2024)
10. Geyer, F., Tent, R., Reiffert, M., Giessing, S.: Perspectives for tabular data protection-how about synthetic data? In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 77–91. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_6](https://doi.org/10.1007/978-3-031-13945-1_6)
11. Jordon, J., Yoon, J., Van Der Schaar, M.: PATE-GAN: generating synthetic data with differential privacy guarantees. In: International conference on learning representations (2018)
12. Liu, T., Tang, J., Vietri, G., Wu, S.: Generating private synthetic data with genetic algorithms. In: International Conference on Machine Learning, pp. 22009–22027. PMLR (2023)
13. McKenna, R., Mullins, B., Sheldon, D., Miklau, G.: Aim: an adaptive and iterative mechanism for differentially private synthetic data. Proc. VLDB Endow. **15**(11), 2599–2612 (2022)
14. McKenna, R., Sheldon, D., Miklau, G.: Graphical-model based estimation and inference for differential privacy. In: International Conference on Machine Learning, pp. 4435–4444. PMLR (2019)
15. Meindl, B., Templ, M.: Feedback-based integration of the whole process of data anonymization in a graphical interface. Algorithms **12**(9), 191 (2019)
16. Nowok, B., Raab, G.M., Dibben, C.: synthpop: bespoke creation of synthetic data in R. J. Stat. Softw. **74**, 1–26 (2016)
17. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. J. R. Stat. Soc. Ser. A Stat. Soc. **181**(3), 663–688 (2018)
18. Templ, M., Kowarik, A., Meindl, B.: Statistical disclosure control for micro-data using the r package sdcMicro. J. Stat. Softw. **67**(i04), 1–36 (2015)
19. Thompson, G., Broadfoot, S., Elazar, D.: Methodology for the automatic confidentialisation of statistical outputs from remote servers at the australian bureau of statistics. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Ottawa, Canada (2013)

20. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
21. Zhou, Y., Kantarcioglu, M., Clifton, C.: On improving fairness of AI models with synthetic minority oversampling techniques. In: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), pp. 874–882. SIAM (2023)



# An Evaluation of Synthetic Data Generators Implemented in the Python Library *Synthcity*

Emma Fössing<sup>1</sup> and Jörg Drechsler<sup>2,3,4</sup>

<sup>1</sup> Georg-August-Universität, Göttingen, Germany  
emmafoessing@gmail.com

<sup>2</sup> Institute for Employment Research, Nuremberg, Germany  
joerg.drechsler@iab.de

<sup>3</sup> Ludwig-Maximilians-Universität, Munich, Germany  
<sup>4</sup> University of Maryland, College Park, USA

**Abstract.** Generating synthetic data has never been so easy. With the increasing popularity of the approach more and more R packages and Python libraries offer ready-made synthesizers that promise generating synthetic data with almost no effort. These synthetic data generators rely on various modeling strategies, such as generative adversarial networks, Bayesian networks or variational autoencoders. Given the plethora of methods, users new to the approach have an increasingly hard time to decide where to even start when exploring the possibilities of synthetic data.

This paper aims at offering some guidance by empirically evaluating the analytical validity of 12 different synthesizers available in the Python library *synthcity*. While this comparison study offers only a small glimpse into the world of synthetic data (many more synthetic data generators exist and we also only rely on the default settings when training the various models), we still hope the evaluations offer some useful insights regarding the performance of the different synthesis strategies.

**Keywords:** Confidentiality · GAN · Neural Networks · Privacy · Synthetic

## 1 Introduction

The ever-rising demand for data has popularized synthetic data over the past decade. When generated using the appropriate tools, synthetic data holds much promise: it offers a dataset that preserves privacy while closely resembling the original dataset, enabling analysis providing results comparable to those obtained from the original dataset (see [9, 11, 15] for more details regarding the synthetic data approach). For many years, the generation of synthetic data has been perceived as a laborious task that required good programming and modeling skills. This perception changed in recent years with more and more packages offering off-the-shelf synthesizers based on various modeling strategies. Especially in the computer science literature, generating synthetic data seems to be

**Table 1.** Description of variables used in the empirical studies

Variable	Label	Range
Sex	<i>sex</i>	male, female
Race	<i>race</i>	4 categories
Marital status	<i>marital</i>	7 categories
Highest attained education level	<i>educ</i>	16 categories
Age (years)	<i>age</i>	15–90
Child support payments (\$)	<i>csp</i>	0, 1–23,917
Social security payments (\$)	<i>ss</i>	0, 1–50,000
Household property taxes (\$)	<i>tax</i>	0, 1–99,997
Household income (\$)	<i>income</i>	1–768,742

perceived as a straightforward task, accessible to anyone with access to pre-programmed models from open-access software packages like *synthpop* for R [21] or *synthcity* for Python [23].

While we generally do not agree with this oversimplification (tuning the hyperparameters of the synthesizers and knowing the data are critically important for obtaining good results), we take a different perspective here. We envision a data custodian with little experience in synthetic data and limited methodological background that tries to evaluate the potential of synthetic data for their data. With the plethora of synthetic data generators available now, the custodian is struggling which approach might be most suitable for this endeavor. Using this perspective, we compare various data synthesizers relying on their default values. Beyond looking at the performance of the synthesizers in terms of preserving analytical validity, we also comment on how easy it is to run the synthesizers, i.e., whether we experienced any technical difficulties implementing them. For our analysis we rely on the synthetic data generators available in the Python library *synthcity*. *Synthcity* considers itself as a platform for various synthetic data generators (SDGs). It also offers various metrics for measuring the analytical validity of the generated data. As a benchmark, we compare these SDGs with synthetic data generated using the default settings from the package *synthpop* that is popular in the statistical community (the default settings rely on CART models for generating the synthetic data). We note that similar comparisons have been performed in earlier papers [6, 20]. Our paper adds to the growing literature on comparison studies by adding several SDGs that have not been evaluated in these earlier papers.

## 2 The Data

For our evaluations, we use a subset of variables and records from the public use file of the March 2000 U.S. Current Population Survey (CPS). The data comprise nine variables measured on  $N = 49,436$  heads of households (see Table 1 for details). Similar data are used in [7, 8, 10, 25, 26] to illustrate and evaluate various aspects of synthetic data.

### 3 The Models

In this section, we provide a short description of the 12 different SDGs evaluated. Given the large number of synthesizers, these descriptions are necessarily very brief. We include references for each of the synthesizers for those interested in fully understanding the methodology. From the large number of synthesizers available in *synthcity* we only include general purpose synthesizers, i.e., we exclude SDGs that focus on specific data such as time series data, survival data, or images. We also exclude synthesizers whose primary motivation is not data protection, such as DECAF, which aims at generating fair data based on biased datasets. The remaining synthesizers can be grouped into five categories based on their modeling strategies: Generative Adversarial Networks, Variational Autoencoders, other neural network based methods, Bayesian networks, and tree-based methods. Some of the SDGs offer formal privacy guarantees based on the concept of differential privacy. Given that we only focus on utility in our paper, directly comparing these methods with other approaches that do not offer any formal guarantees seems unfair towards these SDGs. We therefore always run two versions of the formally private SDGs. In one run, we set the privacy parameter  $\varepsilon = 1$ . In the other run, we either turn off the privacy guarantee completely if the SDGs allows that or we set it to a very large value of  $\varepsilon = 10,000$  to ensure that the privacy enforcement is so weak that it does not sacrifice any utility beyond that resulting from the underlying modeling strategy (these implementations will be denoted by “(NP)” in the results section). In our evaluations we also use a benchmark method (labeled “MD” in the results section) that uniformly samples new values from the set of observed values for each variable, which implies that all generated variables follow a uniform distribution and any relationships between the variables are completely destroyed. Any useful SDG method should perform better than this benchmark.

#### 3.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs, [14]) are based on two competing neural networks, a generator and a discriminator. The generator  $G$  creates synthetic data. The discriminator  $D$  receives both the generated data and the original data as input and tries to distinguish them correctly. In the training process,  $G$  tries to maximize the loss of  $D$  by generating synthetic data that resembles the original data. Over the iterations of the training,  $G$  will generate more and more realistic synthetic datasets, while  $D$  gets better in discriminating between the original and synthetic records. The SDG output is the synthetic data generated by  $G$  at the last iteration of the training process.

*Synthcity* offers several GAN based SDGs: Both the *Anonymization through Data Synthesis-GAN* (AdsGAN, [32]) and the *Conditional Tabular GAN* (CTGAN, [31]) are conditional GANs with AdsGAN implementing explicit and tuneable privacy guarantees. These are not formal guarantees such as those offered through differential privacy. Instead, AdsGAN uses an identifiability constraint based on the weighted euclidean distance to prevent reidentification

based on the synthetic data [32]. The conditional framework for tabular data in CTGAN is implemented to properly model the relationships between continuous and discrete data even for imbalanced distributions [31].

The *Differentially Private GAN* (DPGAN) [30] offers formal privacy guarantees by adding noise to the gradients during the training process. A successor is the *Private Aggregation of Teacher Ensembles-GAN* (PATEGAN) [16] which uses a modified version of the Private Aggregation of Teacher Ensembles and ensures that  $G$  produces private data by training it on a differentially private  $D$  [16].

### 3.2 Variational Autoencoders (VAEs)

Another popular generative model employing neural networks is the Variational Autoencoder (VAE). VAEs are based on an encoder-decoder structure. The encoder condenses the original data into a low-dimensional latent space, while the decoder reconstructs the data based on the latent representation. The underlying probabilistic framework enables the user to sample from conditional distributions to generate synthetic data. The training is based on maximizing the tractable part of the likelihood (the intractable part can be shown to be a Kullback-Leibler divergence which is always positive, [18]).

*Synthcity* includes two of these frameworks which are the *Tabular VAE* (TVAE, [31]) and the *Robust Tabular VAE* (RTVAE) [1]. Akrami et al. [1] argue that using the beta-divergence as a distance measure instead of the Kullback-Leibler divergence is less outlier-sensitive and utilize it for their RTVAE model.

### 3.3 Other Neural Net Models

The *tabular Denoising Diffusion Probabilistic Model* (DDPM) [19] is based on neural networks that learn the inversion of Markov chains which start on parametric distributions. The goal is to approximate the end point of the target distribution of the Markov chains.

A disadvantage of GANs and VAEs is the lack of explicit density evaluations. *Normalizing flows* (NFlows) utilizes a series of one-to-one transformations and models complex distributions by inverting simple base distributions which ensures invertibility and differentiability [13, 22].

### 3.4 Bayesian Networks

The *synthcity* library also offers a variety of models that do not employ neural networks. One group of models offered is based on Bayesian networks. Bayesian networks (BN) rely on Directed Acyclic Graphs (DAGs) that establish the relationships between the different variables by setting up a graph structure that models the dependencies between the variables. *Synthcity* offers two SDGs based on Bayesian networks. The model plugin named *Bayesian Networks* is based on the Python library pgmpy (Probabilistic Graphical Models in Python) [2]. Zhang

et al. [33] introduced a Bayesian network synthesizer called *PrivBayes* that offers differential privacy guarantees. Our attempts at training the model were unsuccessful, largely due to constraints on computational resources, specifically the limitation of Random-Access Memory (RAM). Despite efforts to parameterize the model in a manner that mitigates extensive RAM usage, these limitations persisted. The maintainers of *synthcity* confirmed that this is a known problem with *PrivBayes* and suggested using alternative SDGs.

### 3.5 Tree-Based Methods

*Synthcity* also offers a model plugin called *Adversarial Random Forests* (ARF) [28]. ARF extends the classical random forest approach [3], which relies on an ensemble of independent classification and regression trees (CART) [4], by using unsupervised trees and rendering independent data within each leaf to generate the synthetic data [28]. The model is built based on the classical CART, which was suggested as a tool for synthesis many years before all of the other tools considered in this paper [26]. Over the years several studies have confirmed the high level of analytical validity achievable using CART based synthesizers [12, 20]. We therefore include this synthesizer as a benchmark to compare against. As *synthcity* does not include this SDG as a plugin, we resort to the package *synthpop* that also offers pre-programmed SDGs [21] and uses CART based models as the default.

## 4 Utility Metrics

Following Drechsler and Haensch [11] we evaluate three dimensions of utility. *Global (or broad) utility measures* try to assess the utility by directly comparing the similarity of the original and the synthetic data. *Outcome-specific (or narrow) utility measures* assess how well specific analysis tasks can be replicated using synthetic data instead of the original data. Finally, *Fit-for-purpose measures* evaluate whether basic features of the original data such as marginal distributions, correlations, or logical constraints between the variables are preserved. See [10] for a discussion of why it is important to always look at multiple dimensions of utility when assessing the quality of the generated data. In this section, we briefly summarize the utility measures that we employ for each of these dimensions.

### 4.1 Global Utility Measures

A popular global utility measure is the standardized propensity score mean squared error (S\_pMSE, [27, 29]). The procedure of computing the pmSE consists of the following steps:

1. Stack the  $n_{org}$  original records and the  $n_{syn}$  synthetic records adding an indicator, which equals one if the record is from the synthetic data and zero otherwise.

2. Fit a model to predict the data source (original/synthetic) using the information contained in the data. Let  $p_i$ ,  $i = 1, \dots, N$  (with  $N = n_{org} + n_{syn}$ ) denote the predicted value for record  $i$  obtained from the model.
3. Calculate the pMSE as  $1/N \sum_N (p_i - c)^2$ , with  $c = n_{syn}/N$ .

To overcome some of the shortcomings of the pMSE, [27] developed the standardized pMSE (S\_pMSE), which standardizes the pMSE by using the expected value and the standard error of the pMSE under the assumption of a correctly specified synthesizer. The smaller the S\_pMSE the higher the analytical validity of the synthetic data.

A related global utility measure is implemented in *synthcity*: The measure called “SyntheticDetectionXGB” also relies on the stacked data and uses XGBoost [5] to predict whether a record is from the original or synthetic data. To measure how well XGBoost can discriminate between the two data sources, the measure computes the area under the curve (AUC) for the receiver operator curve (ROC), which plots the false positive rate against the true positive rate. A value close to one indicates that XGBoost can perfectly discriminate between the original and synthetic data.

## 4.2 Outcome-Specific Utility

As our outcome-specific utility measure, we use the confidence interval overlap. The confidence interval overlap measure was first proposed by [17]. Paraphrasing from [9], its computation can be summarized as follows: For any estimand, we first compute the 95% confidence intervals for the estimand from the synthetic data,  $(L_s, U_s)$ , and from the original data,  $(L_o, U_o)$ . Then, we compute the intersection of these two intervals,  $(L_i, U_i)$ . The utility measure is

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)}. \quad (1)$$

When the intervals are nearly identical, corresponding to high utility,  $I \approx 1$ . When the intervals do not overlap, corresponding to low utility,  $I = 0$ . The second term in (1) is included to differentiate between intervals with  $(U_i - L_i)/(U_o - L_o) = 1$  but different lengths.

## 4.3 Fit-For-Purpose Measures

We look at two fit-for-purpose measures. The first measure takes a closer look at categorical variables by examining which categories are not represented in the synthetic data. Specifically, we identify which categories disappear in the synthetic data and measure their relative frequency in the original data. The measure reports the sum of these relative frequencies over all categories of all categorical variables that no longer exist in the synthetic data.

Our second measure uses the S\_pMSE again. However, instead of using all variables when training the classifier, we look at all possible combinations of pairs of variables, and always only use these variable pairs as predictors. This way,

**Table 2.** S\_pMSE and AUC-ROC scores for the different synthesizers.

SDG	S_pMSE	XGB-AUCROC
CART	0.43	0.4600487
DDPM	36.62	0.6292901
ARF	551.54	0.8487266
NFlows	859.36	0.9185467
BN	1630.35	0.9329831
PATEGAN (NP)	2087.13	0.9662546
CTGAN	2095.57	0.9675358
AdsGAN (NP)	2246.49	0.9681263
AdsGAN	2301.88	0.9598442
TVAE	2476.05	0.9692178
PATEGAN	3030.49	0.9781081
MD	3880.29	0.9999983
DPGAN (NP)	4077.25	0.99999686
DPGAN	4615.38	0.9999972
RTVAE	5712.05	1

the S\_pMSE measures how well the relationships between these variable pairs are preserved. This utility measure is implemented in the function *utility.table* in *synthpop* and we use this function for our evaluations. In the appendix we also report the S\_pMSE using each variable as the sole predictor when estimating the propensity score to see how well the marginal distributions are preserved.

## 5 Results

In this section, we present the results of our evaluation study. We group the results based on the different dimensions of utility discussed in the previous section.

### 5.1 Global Utility Measures

In Table 2 we present the results for S\_pMSE and the XG-Boost Detection measure. The SDGs are sorted in increasing order by the S\_pMSE. Although the ranking of the AUC-ROC scores is not identical, the ranking of the scores exhibit a high similarity.

Using the metrics from Table 2 to assess utility, we conclude that almost none of the synthesizers except for CART and maybe DDPM and ARF offer acceptable utility. We note that Raab et al. [24] suggest that the S\_pMSE should generally be less than 10 to be considered acceptable. Although we believe that deciding which number is acceptable will always depend on context, values above 1,000 are strong indicators for low utility. The fact that the AUC is above 0.9 for most

of the synthesizers also seems alarming. This implies that XGBoost is able to correctly predict for almost all records whether they stem from the original or the synthetic data. Some of the synthesizers even have lower utility than MD, which does not preserve any information from the original data at all.

## 5.2 Outcome-Specific Utility

To measure the outcome-specific utility, we run a linear regression of the logarithmic income on all other non-monetary variables for the original data and each of the synthetic datasets and compute the confidence interval overlaps. As some of the categorical variables have large numbers of categories (`educ`) or are highly imbalanced (`race` and `marital`), we aggregate these variables before fitting this model. This helps avoiding difficulties in comparing the results, as some of the sparsely populated categories are no longer present in some of the synthetic datasets (see the next section). We aggregate `educ` into four categories (“No high school degree”, “Finished high school”, “Associate or bachelor’s degree”, “Master’s degree or higher”), `race` into two categories (“white”, “non-white”), and combine the categories “spouse present and married” and “spouse absent and married” as well as “divorced” and “separated” for `marital`.

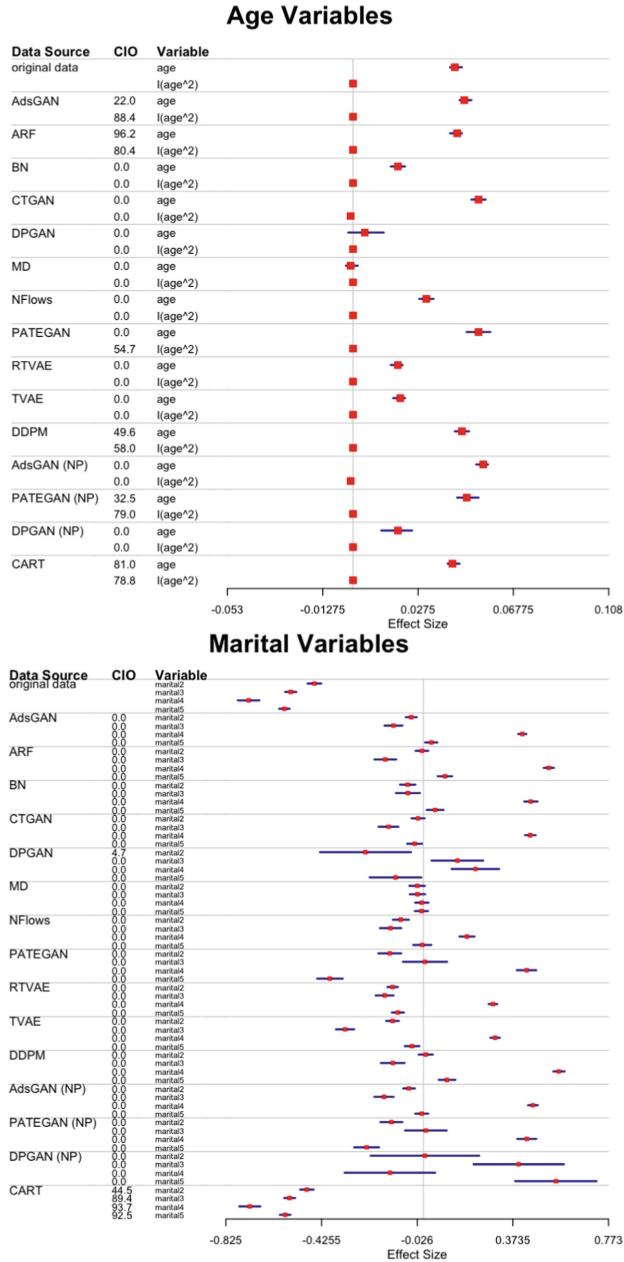
The effects are presented in Fig. 1 and Fig. 4 in the appendix. Although the categories have been merged, it is clear that most of the SDGs struggle to preserve the relationship between  $\log(\text{income})$  and the predictors. Figure 1 contains the results for `age` and `marital status`. The effect sizes for `age` are mostly similar to the original data, although the effect of age is often underestimated and the confidence interval overlaps are relatively small for most synthesizers. However, most synthesizers show poor results for marital status. While all categories have a strong negative effect in the original data, results for several synthetic datasets would imply a strong positive effect for some of the categories of marital status. In fact, CART is the only synthesizer for which confidence interval of the original and the synthetic data overlap for all categories of this variable.

Figure 4 in the appendix shows the results for `education` and the binary variables `race` and `sex`. Most synthesizers preserve the small effects of `race` and `sex` relatively well, although confidence interval overlaps tend to be small for most synthesizers. The effects of `education` are only preserved by some of the synthesizers with DDPM and CART being the only synthesizers that achieve a confidence interval overlap larger than zero for all three education categories.

The results for DPGAN stand out, with very large confidence intervals and, for multiple variables, effect sizes that are substantially different from the original data. In contrast, the CART-generated data produces a regression model that closely resembles the original data and outperforms all other SDGs, being the only synthetic data that achieves a confidence interval overlap of more than 44% for all regression coefficients (average overlap across all coefficients: 78.33%).

## 5.3 Fit-For-Purpose Measures

Table 3 lists all synthesizers that produced synthetic data in which at least one category of any of the categorical variables was no longer represented. For each



**Fig. 1.** Confidence interval overlap (CIO) in percent and effect sizes of age and marital status in a linear regression of  $\log(\text{income})$  on age, education, marital status, sex, and race. The horizontal lines represent the 95% confidence intervals.

**Table 3.** List of categories that are no longer present in the synthetic data and relative frequencies of these categories in the original data. Results are presented separately for each synthesizer. Synthesizers for which no category disappeared are omitted.

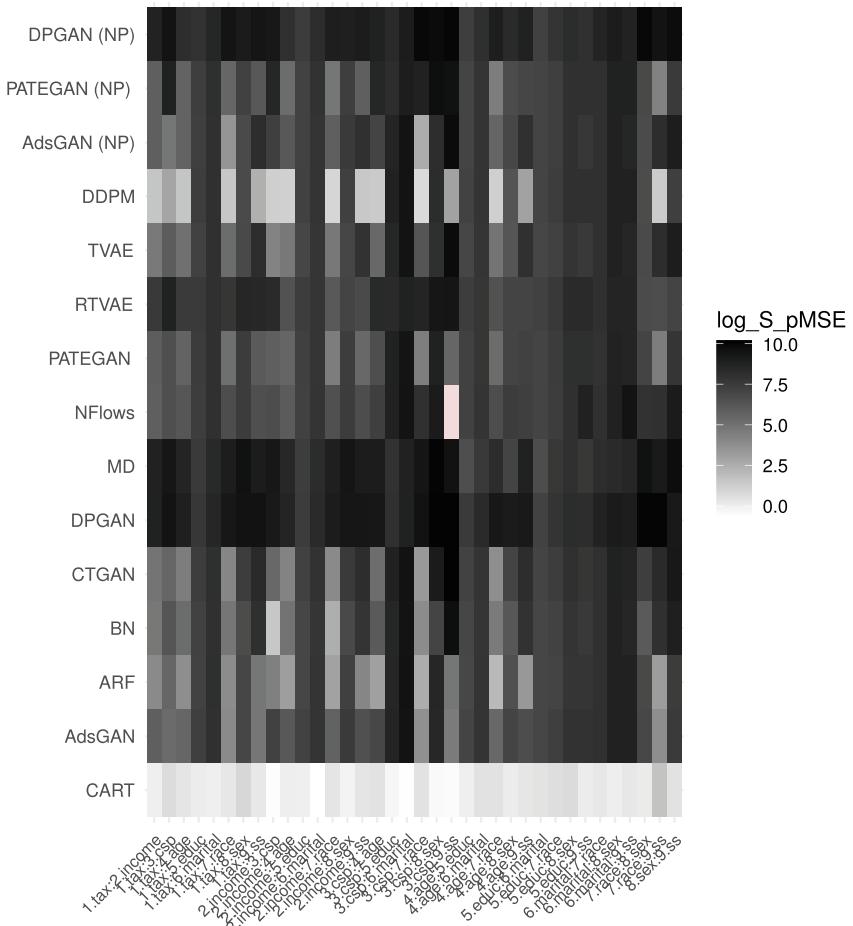
	Variable	Category	Rel. freq. in true data
CTGAN	educ	46	0.01
<b>Sum</b>			<b>0.01</b>
DPGAN	marital	1	0.55
	race	1	0.86
<b>Sum</b>			<b>1.41</b>
RTVAE	educ	31	0.00
	educ	33	0.02
	educ	36	0.03
	educ	38	0.01
	educ	41	0.04
	educ	44	0.06
	educ	46	0.01
	marital	2	0.00
	race	3	0.01
<b>Sum</b>			<b>0.18</b>
TVAE	race	3	0.01
<b>Sum</b>			<b>0.01</b>
DPGAN (NP)	marital	1	0.55
	race	1	0.86
<b>Sum</b>			<b>1.41</b>

of these synthesizers, it lists the categories that disappeared and their relative frequencies in the original data. The utility score is the sum of these frequencies (with higher values indicating lower utility).

For most of the synthesizers, only sparsely populated categories disappear. However, for DPGAN with and without formal privacy guarantees, two categories with high frequencies (above 50%) are no longer present in the synthetic data. The synthesizers listed in Table 3 are also among the worst performing synthesizers according to the global utility measures.

Additionally, we compare the synthetic data to the original data using the logarithmic S\_pmse computed for all pairwise combinations of variables. The values are log-scaled to account for the large disparities between the different synthetic datasets, ensuring that the results remain visually comparable.

The results are presented in the form of a heatmap in Fig. 2. Beyond the low performance of DPGAN and RTVAE that was also evident in the other utility metrics, we see that many of the synthesizers seem to struggle preserving the relationship of `csp` with `age` and `educ`. Most of the synthesizers also have lower utility for all pairwise comparisons that include `educ` or `marital`. We again find



**Fig. 2.** Logarithmic S\_pMSE scores using all pairwise combinations of variables as predictors when estimating the propensity scores. Note that for Nflows there is a missing value for the pair csp and ss due to computational reasons.

that CART substantially outperforms all other SDGs for all variables. DDPM provides higher utility in compared to other synthesizers (except CART) for many variables, despite its utility still being fairly low for the mentioned variables. Results for the S\_pMSE scores using each variable as the only predictor when estimating the propensity score are presented in Fig. 3 in the appendix. The results match those for the pairwise evaluations: CART performs best, while DPGAN gives the worst results. Most synthesizers struggle with the variables csp, marital, and educ.

## 6 Conclusion

In this paper, we compared various synthetic data generators implemented in the Python library *synthcity*. As a benchmark, we used CART based synthesizers as implemented in the *synthpop* package in R. Our simulations confirmed the findings of previous papers that CART models tend to offer much higher analytical validity than any of the other synthesizers. Among the other methods only DDPM and ARF produced results that might be considered acceptable in some circumstances. Another downside of the deep learning based approaches that we did not mention so far, is the high computational demand. Training these models can be time consuming and it can be useful to parallelize the computation process by using graphics processing units (GPUs). In theory, users could use cloud services to speed up the computations (we used Google Colaboratory to run our Python notebooks). However, in practice, this will typically not be an option as the data cannot be uploaded to the cloud for confidentiality reasons.

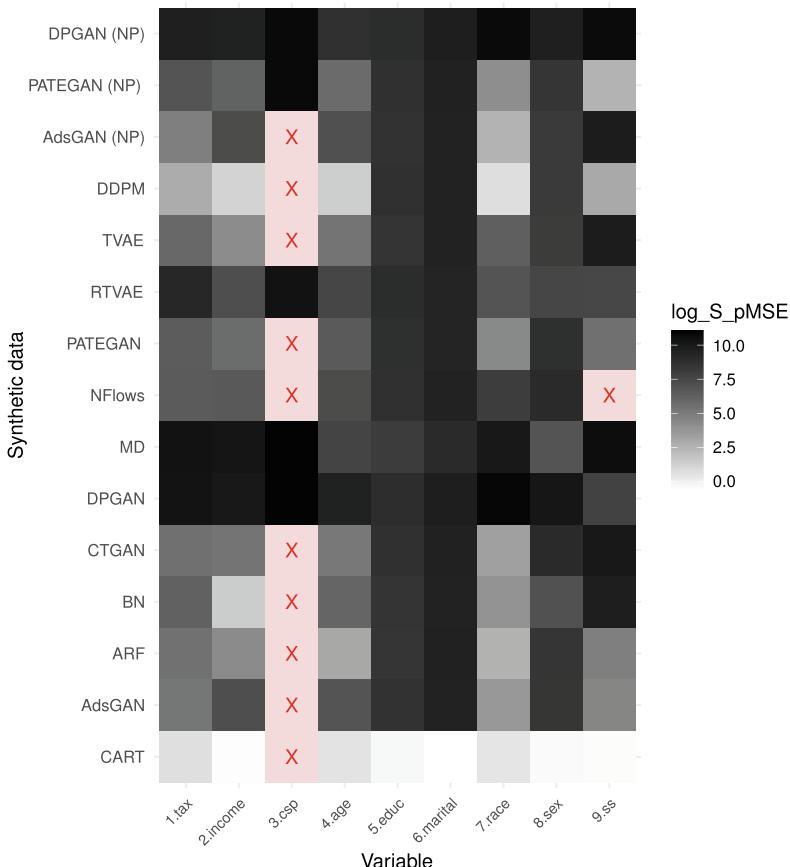
Our evaluation study has several important limitations. Most importantly, we always relied on the default settings of the synthesizers when generating the data. It is well known that, especially for deep learning models, careful training of the models is important to obtain optimal results. At the same time, deep learning models have a very large number of tuning parameters (the structure of the network, the number of epochs to run, the selection of the loss function, etc.). Given the high computational demand of these models, fully exploring the space of the hyperparameters is often infeasible in practice, especially when aiming to compare various synthesis strategies. This can be seen as another advantage of the CART synthesizers. They seem to be very robust and tend to generate high-quality results even without tuning. A second limitation of our evaluation is that we only used one relatively old dataset. We are less concerned regarding the timeliness of the data as this is a pure methodological study and we don't expect that the general findings would change much, if we would use the same variables from the current year. Still, the relative performance of the synthesizers might depend on the complexity and size of the data (in terms of sample size but also in terms of the number of variables). This might especially be true for the deep learning models, which tend to require large sample sizes to learn complex data structures. Another limitation of our study is that we looked at a limited set of utility metrics. Fully understanding the utility of a generated synthetic dataset is a difficult task [10]. However, given the substantial differences in performance, we expect that the big picture results would have been similar if we had run these comparisons on different datasets and extended the list of utility metrics. An obvious additional shortcoming of this paper is that it only focuses on measuring the utility. Clearly, the main motivation for generating synthetic data is to offer data protection. Thus, it will be important to understand which level of protection the different synthetic datasets offer. Furthermore, for those methods that offer formal privacy guarantees, looking only at utility is obviously an unfair comparison. Comparing the methods by the level of data protection they provide will be an interesting area for future research. Finally, the paper only offers an empirical evaluation of the different synthesizers. Digging deeper

to fully understanding the reasons behind the differences in performance would be an important contribution for the future.

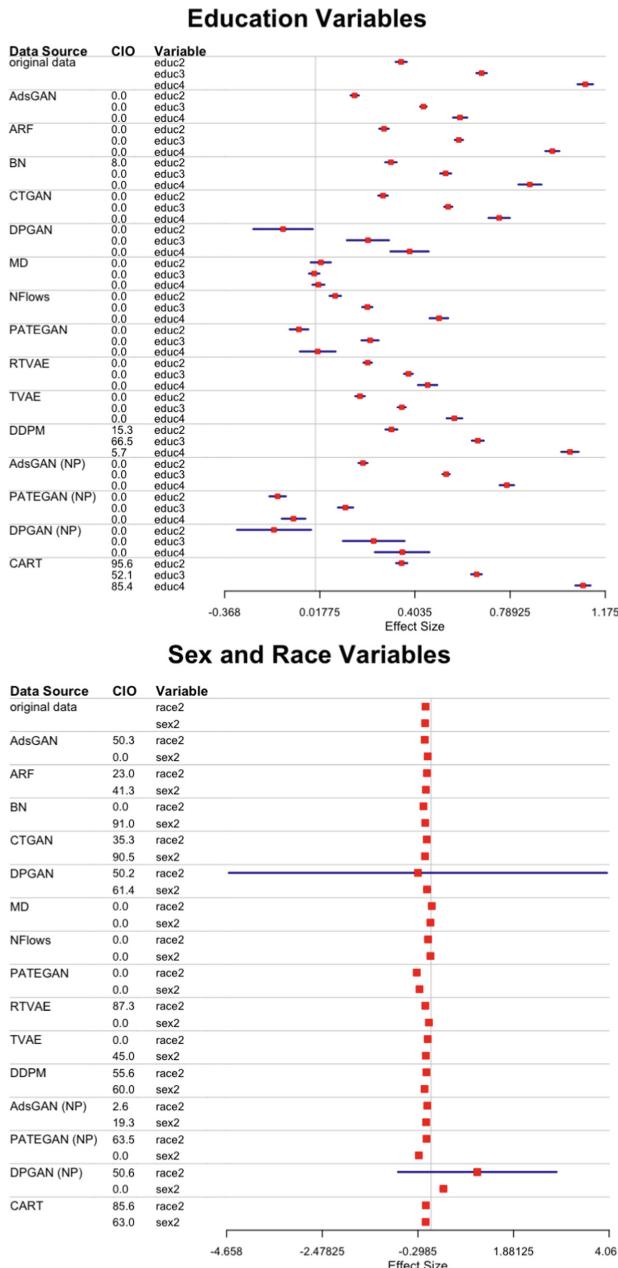
**Acknowledgments.** This work was supported by a grant from the German Federal Ministry of Education and Research (grant number 16KISA096) with funding from the European Union-NextGenerationEU.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## Appendix



**Fig. 3.** Logarithmic S\_pMSE scores for each variable separately as the only predictor when estimating the propensity scores. Note that in several instances the S\_pMSE could not be estimated due to computational reasons. These instances are marked with an X.



**Fig. 4.** Confidence interval overlap (CIO) in percent and effect sizes of education, sex, and race in a linear regression of  $\log(\text{income})$  on age, education, marital status, sex, and race. The horizontal lines represent the 95% confidence intervals.

## References

1. Akrami, H., Joshi, A.A., Li, J., Aydöre, S., Leahy, R.M.: A robust variational autoencoder using beta divergence. *Knowl.-Based Syst.* **238**, 107886 (2022)
2. Ankan, A., Panda, A.: pgmpy: probabilistic graphical models using python. In: SciPy, pp. 6–11. Citeseer (2015)
3. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
4. Breiman, L.: Classification and Regression Trees. Routledge, Milton Park (2017)
5. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
6. Dankar, F.K., Ibrahim, M.: Fake it till you make it: guidelines for effective synthetic data generation. *Appl. Sci.* **11**(5), 21–58 (2021)
7. Drechsler, J., Reiter, J.P.: Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In: Domingo-Ferrer, J., Saygin, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 227–238. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-87471-3\\_19](https://doi.org/10.1007/978-3-540-87471-3_19)
8. Drechsler, J., Reiter, J.P.: Sampling with synthesis: a new approach for releasing public use census microdata. *J. Am. Stat. Assoc.* **105**, 1347–1357 (2010)
9. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, vol. 201. Springer, New York (2011)
10. Drechsler, J.: Challenges in measuring utility for fully synthetic data. In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 220–233. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_16](https://doi.org/10.1007/978-3-031-13945-1_16)
11. Drechsler, J., Haensch, A.C.: 30 years of synthetic data. *Stat. Sci.* **39**(2), 221–242 (2024)
12. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. *Computat. Stat. Data Anal.* **55**(12), 3232–3243 (2011)
13. Durkan, C., Bekasov, A., Murray, I., Papamakarios, G.: Neural spline flows. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
14. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
15. Hu, J., Bowen, C.M.: Advancing microdata privacy protection: a review of synthetic data methods. *Wiley Interdisc. Rev. Comput. Stat.* **16**(1), e1636 (2024)
16. Jordon, J., Yoon, J., Van Der Schaar, M.: PATE-GAN: generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations (2018)
17. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A framework for evaluating the utility of data altered to protect confidentiality. *Am. Stat.* **60**(3), 224–232 (2006)
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114) (2013)
19. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: Tabddpm: modelling tabular data with diffusion models (2022). <https://arxiv.org/abs/2209.15421>
20. Little, C., Elliot, M., Allmendinger, R., Samani, S.S.: Generative adversarial networks for synthetic data generation: a comparative study. [arXiv:2112.01925](https://arxiv.org/abs/2112.01925) (2021)
21. Nowok, B., Raab, G.M., Dibben, C.: synthpop: bespoke creation of synthetic data in R. *J. Stat. Softw.* **74**, 1–26 (2016)

22. Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B.: Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.* **22**(57), 1–64 (2021)
23. Qian, Z., Cebere, B.C., van der Schaar, M.: Synthcity: facilitating innovative use cases of synthetic data in different data modalities. arXiv preprint [arXiv:2301.07573](https://arxiv.org/abs/2301.07573) (2023)
24. Raab, G.M., Nowok, B., Dibben, C.: Assessing, visualizing and improving the utility of synthetic data. arXiv preprint [arXiv:2109.12717](https://arxiv.org/abs/2109.12717) (2021)
25. Reiter, J.P.: Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. *J. R. Stat. Soc. Ser. A* **168**, 185–205 (2005)
26. Reiter, J.P.: Using CART to generate partially synthetic, public use microdata. *J. Official Stat.* **21**, 441–462 (2005)
27. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. *J. R. Stat. Soc. Ser. A Stat. Soc.* **181**(3), 663–688 (2018)
28. Watson, D.S., Blesch, K., Kapar, J., Wright, M.N.: Adversarial random forests for density estimation and generative modeling. In: International Conference on Artificial Intelligence and Statistics, pp. 5357–5375. PMLR (2023)
29. Woo, M.J., Reiter, J.P., Oganian, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confidentiality* **1**(1) (2009)
30. Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J.: Differentially private generative adversarial network. arXiv preprint [arXiv:1802.06739](https://arxiv.org/abs/1802.06739) (2018)
31. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
32. Yoon, J., Drumright, L.N., Van Der Schaar, M.: Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE J. Biomed. Health Inform.* **24**(8), 2378–2388 (2020)
33. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: Privbayes: private data release via Bayesian networks. *ACM Trans. Database Syst. (TODS)* **42**(4), 1–41 (2017)



# Evaluation of Synthetic Data Generators on Complex Tabular Data

Oscar Thees<sup>1,2,3</sup>(✉) , Jiří Novák<sup>1,2,4</sup> , and Matthias Templ<sup>1,2</sup>

<sup>1</sup> University of Northwestern Switzerland, Rickenbachstrasse 16,  
4600 Olten, Switzerland

`{oscar.thees,jiri.novak,matthias.templ}@fhnw.ch`

<sup>2</sup> Swiss Data Anonymization Competence Center (SwissAnon), Rickenbachstrasse 16,  
4600 Olten, Switzerland

<sup>3</sup> Technische Universität Wien, Karlsplatz 13, 1040 Wien, Austria

<sup>4</sup> University of Zurich, Rämistrasse 71, 8006 Zürich, Switzerland

<http://www.swissanon.ch>

**Abstract.** Synthetic data generators are widely utilized to produce synthetic data, serving as a complement or replacement for real data. However, the utility of data is often limited by its complexity. The aim of this paper is to show their performance using a complex data set that includes cluster structures and complex relationships. We compare different synthesizers such as synthpop, Synthetic Data Vault, simPop, Mostly AI, Gretel, Realtabformer, and arf, taking into account their different methodologies with (mostly) default settings, on two properties: syntactical accuracy and statistical accuracy. As a complex and popular data set, we used the European Statistics on Income and Living Conditions data set. Almost all synthesizers resulted in low data utility and low syntactical accuracy.

The results indicated that for such complex data, simPop, a computational and methodological framework for simulating complex data based on conditional modeling, emerged as the most effective approach for static tabular data and is superior compared to other conditional or joint modelling approaches.

**Keywords:** Tabular Data · Data Synthesis · Comparison of Synthesizers · Data Utility

## 1 Introduction

The awareness of the importance of personal data privacy is increasing, and legislative bodies have established precedents, such as the General Data Protection Regulation, which demonstrate their willingness to protect personal data in the context of the digital age. The vulnerability of complex data sets to privacy attacks is well documented. However, existing anonymization techniques have been found to be inadequate in providing adequate protection for open-data without borders while keeping the utility high (including outliers) [3, 4].

This effectively constrains the capacity to disseminate such data sets. In order to further make microdata available for various purposes, researchers like Rubin [25], or Jordon et al. [12] have therefore extensively proposed the usage of synthetic data.

The distinction between synthetic data and real data is evident. The former is generated by models, whereas the latter is derived from systems in the real world. However, real-world data are prerequisites for generating synthetic data, as they are the input that the models or synthetic data generators (SDGs) require to function.

Two primary approaches for synthetic data generation are joint modeling and conditional modeling. Joint modeling captures the entire data distribution simultaneously, making it powerful but computationally heavy. Naturally, it is difficult or even impossible to simulate data with high data utility when there are many variables and relationships using joint modeling. Conditional modeling generates data variable by variable based on conditional relationships, offering flexibility and scalability but sometimes missing intricate dependencies when they are not explicitly included in the model.

In the literature, many reference points are available for comparing synthetic data generators. An overview is given in Appendix B.

However, no contribution compared complex survey data that meets all challenges, such as handling missing values, diversity of column types (like continuous, semi-continuous, count, categorical and ordered categorical), complex distributions, and complex relationships between variables [15]. None of the packages can handle further sampling weights when samples are not drawn with simple random sampling without replacement, cluster structures (like persons in households) and hierarchical structures (like municipalities in regions). The aim of this contribution is thus to compare different generators on a more complex data set in order to highlight the deficits and advantages of various methods under a realistic setting.

The remainder of this article is structured as follows. In Sect. 2, we commence with an introduction and review of the various SDGs and the different methodological approaches within them. In Sect. 3, we present our simulation framework, introduce our data, and review the properties that define a good SDG. We also discuss how we can assess these properties in our data to rate the SDGs. In Sect. 4, we present the results of the comparison in terms of data utility between the different synthetic samples/synthetic populations (in the case of `simPop` [30]) and the true synthetic population. Finally, in Sect. 5, we offer an interpretation of these results and contextualize them within the relevant literature, providing a conclusion.

## 2 Data Synthesizers

With the growing interest in synthetic data, the last decade has seen the emergence of many tools that can be used to synthesise data. Methodologically, they can be very different. From classical statistical approaches, often based on statistical disclosure control or econometric models, to supervised machine learning

approaches such as tree-based methods, to unsupervised deep learning methods such as GANs, VAEs and LLMs.

**Table 1.** Overview of the used synthetic data generators.

Name	Availability	Used Method(s)	Software	Year	Last UPD
synthpop <sup>1</sup>	open source	CART	R-Package	2014	2022
Synthetic Data Vault <sup>2</sup>	open source	GaussianCopula, CTGAN, VAE	Python Package	2018	2024
simPop <sup>3</sup>	open source	Estimation of the conditional probabilities by multinomial log-linear models and random draws and random forest (alternative)	R-Package	2010	2024
Mostly AI <sup>4</sup>	proprietary	Combination of transformers, GANs, VAEs and autoregressive networks	Mostly AI (AT)	2017	2024
Gretel <sup>5</sup>	both	Synthetic ACTGAN	Gretel Labs (US)	2020	2024
REaLTabFormer <sup>6</sup>	open source	GPT-2	Python Package	2022	2024
arf <sup>7</sup>	open source	Adv. random forests	R-Package	2022	2024

<sup>1</sup><https://cran.r-project.org/web/packages/synthpop/> (accessed 26.06.2024)

<sup>2</sup><https://pypi.org/project/sdv/0.3.1/> (accessed 26.06.2024)

<sup>3</sup><https://cran.r-project.org/web/packages/simPop/> (accessed 26.06.2024)

<sup>4</sup><https://mostly.ai/> (accessed 26.06.2024)

<sup>5</sup><https://gretel.ai/> (accessed 26.06.2024)

<sup>6</sup><https://pypi.org/project/REaLTabFormer/> (accessed 26.06.2024)

<sup>7</sup><https://cran.r-project.org/web/packages/arf/> (accessed 26.06.2024)

Table 1 provides an overview of the tools utilized in the generation of synthetic data presented in this paper. The number of available tools has increased considerably in recent years, as evidenced by the comparison literature shown above. It is beyond the scope to compare all existing methods and toolboxes, so the selection of tools employed in this comparison was designed to encompass a diverse range of methods and software in order to provide a comprehensive overview of the available options.

The following sections describe the SDGs utilized. Unless otherwise specified, the default parameter settings are applied.

**synthpop** [18] is an R package created by Nowok et al. [18] to generate synthetic individual-level data without complex data structure using conditional modeling. Parametric and non-parametric methods for statistical disclosure control are available in this package. Among the parametric methods are included various types of regression, like normal linear, logistic, polytomous and ordered. Non-parametric methods are based on classification and regression trees models. The package also includes a log-linear model approach for categorical data implemented via an ipf procedure.

**Synthetic Data Vault** [21] is a Python library to create tabular synthetic data using various approaches. The Gaussian Copula Synthesizer employs a copula, a mathematical function that enables the description of a joint distribution of multiple random variables by analyzing the dependencies between their marginal distributions. The package’s CTGAN Synthesizer uses GAN-based methods and we used 500 epochs. The package also provides a variational autoencoder model (TVAE) synthesizer and we used 500 epochs.

**simPop** [30] is an R package designed to simulate synthetic populations (from where samples can be drawn if wished for). It utilizes tabular data and auxiliary information through model-based methods. Optionally, the user can define special structures of the data set, such as sampling weights, cluster structures and strata. Population totals are estimated in the first step using the Horvitz-Thompson estimator [11], considering only basic structural variables (later, we used “age”, “gender”, “federal state” and the cluster structure). For details on how the software deals with cluster structures and weighting, we refer to [2]. To simulate categorical variables, conditional probabilities were estimated using multinomial log-linear models and random draws from the resulting distributions. Note that more advanced methods, such as XGBoost and random forests, can be used. The method “multinom” was also used to simulate continuous variables combined with random draws from the resulting categories. Components of continuous variables of the data are simulated by resampling fractions (in our application: conditionally to “occupational code”, “citizenship” and “gender”) fitted from the original data.

**Mostly AI** [16] offers a service of data anonymization through data synthesis via a “Synthetic Data Platform,” [16]. It uses a combination of transformers, GANs, VAEs and autoregressive networks. The configuration accuracy preset, provided by the developers, was employed. This preset instructs the training model to reach the highest possible accuracy (and the lowest validation loss) within a specified time limit of 120 min and 100 training epochs. Additionally, the parameter model size, which defines the amount of internal parameters that the training model uses, was altered to the category large.

**Gretel** [10] offers a service for the generation of synthetic data using a variety of machine learning models. The service can be accessed via an API or data upload. The default method, *Synthetic ACTGAN*, was selected.

**arf** [31] is an acronym for “adversarial random forests”. The R package **arf** [31] employs an unsupervised random forest, which uses a recursive procedure in which the trees gradually learn the structural properties of the data through alternating rounds of generation and discrimination.

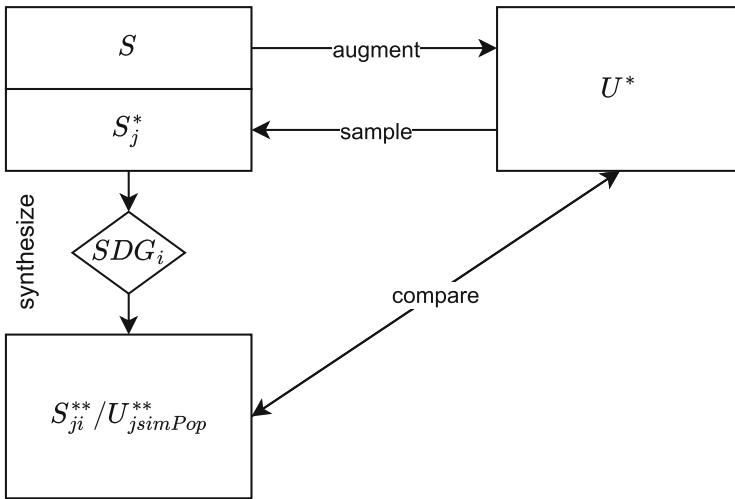
**REaLTabFormer** [28] is an acronym that stands for “realistic relational and tabular data using transformers”. It is a Python library that offers a sequence-to-sequence method for generating synthetic relational data sets. In this context, the term “relational” refers to the interrelatedness of observations that is typically established by a common identifier. Furthermore, the package employs a GPT-2 large language model for synthesising non-relational tabular data, where the observations are independent of each other. Our data could be divided into two distinct data sets, as the package demands.

However, for the purposes of this comparison, we worked with the method for non-relational tabular data. Consequently, it is acknowledged that the SDG is less adept at comprehending the intricacies of cluster structures.

### 3 Data Utility and Disclosure Risk

#### 3.1 Simulation Framework

Users generally want to make estimates for the whole population for survey data drawn with a complex sampling design from a finite population. In this setting, a design-based evaluation is recommended [1, 17]. Such a setting needs a finite population as the truth from which samples are drawn. These samples are synthesized and compared with the true population. In general, not all variables are known from the true population, and thus, the truth must be created first.



**Fig. 1.** This flowchart outlines the principal methodology employed in the paper.

The general workflow is visible in Fig. 1.

- $S$  is the public-use EU-SILC data set (2013) from Statistics Austria with  $n_S = 13'513$  observations.  $S$  is only used for creating a synthetic truth and is not considered elsewhere.
- $U^*$  serves as synthetic truth and is created through the application of data augmentation techniques, employing the weights and Horvitz-Thompson estimates of population counts from  $S$ . The code for simulating the population is given in [30], with minimal adaptations to the variable names.  $U^*$  serves as the synthetic truth in a finite population setting and design-based simulation. This population consists of  $N_{U^*} = 8'332'266$  observations.
- $S_j^*$ : Two samples are drawn with different sampling designs ( $j \in \{\text{equal, prop}\}$ ). *prop* means that the number of households selected from  $U^*$  is proportional to the size of categories in the variable region, and *equal* means

that the same number of households are drawn in each region. Thus, it represents two extreme approaches: one closely related to simple random sampling without replacement and one design with massive oversampling in small regions with respect to their size.  $S_j^*$  is calibrated using iterative proportional fitting on population totals from the cross-tabulation of `region`  $\times$  `gender`  $\times$  citizenship in  $U^*$ . Thus, the margins on these variables are equal to the synthetic truth. The sample sizes are  $n_{S_j^*} = 13'499$  and  $n_{S_j^{**}} = 13'838$ .

- $S_j^{**}$  represents the synthetic data sets obtained by applying synthesizers on  $S_j^*$ . It should be noted that when the `simPop` package is applied to  $S_j^*$ , a synthetic population designated as  $U_j^{**}$  is created, with the same number of households as  $U^*$  (from which a sample can be drawn at any time).

Population characteristics are estimated using the synthetic data obtained from the SDG's (from  $S_j^{**}$ ), thus the calculations on  $U^*$  (= the truth) are compared with the estimates in terms of syntactical accuracy and statistical accuracy, which will be discussed later.

### 3.2 Data

The European Union Statistics of Income and Living Conditions (EU-SILC) survey is an example of a real-world data set used in official statistics. This survey is well known for producing highly complex data sets. These data sets are used mainly to measure the risk of poverty and social cohesion in Europe and to monitor the Lisbon 2010 strategy and Europe 2020 goals of the European Union. Specifically, we used the Austrian EU-SILC public use data set from 2013.

The description of the variable and more information on this data set are provided in the manual of the R package `simPop` [30].

Table 4 in Appendix A lists and describes the EU-SILC variables used for our study. Some categories of economic status and citizenship, respectively, have been combined due to their low frequency of occurrence; the combined categories are marked with an asterisk (\*). A complete description of EU-SILC variables can be found in [9] or online through <https://www.gesis.org/en/missy/materials/EU-SILC/documents/codebooks> (accessed 25.06.2024).

### 3.3 Data Utility Measures

**Statistical Accuracy** is concerned with the statistical properties of the synthetic data, including univariate and multivariate distributions, and how these compare to the real data. The comparison of the univariate distribution involves evaluating the similarity between the distributions of individual variables in the synthetic and real datasets. This assessment ensures that essential statistical properties, such as means, variances, and skewness, are preserved in the synthetic data.

Estimating distributional distances in higher dimensions is a challenging and computationally expensive task to verify the utility of synthetic data. Therefore, we compare synthetic data and original data using the distribution of *propensity*

*scores* [24,32], which are employed to assess the accuracy of a logistic regression trained to differentiate between real and synthetic data<sup>1</sup>. Hereby, the synthetic data generated with an SDG is merged in a row with the original data, and a binary variable expresses the group membership (synthetic or real). A logistic regression model is fitted on the binary variable with region, “household size”, “citizenship”, “gender”, “occupational code”, and the logarithmized “personal gross income”. From the trained parameters, each observation of the synthetic and the real data is predicted with a membership score. If the synthetic data are not distinguishable from the original data, then these scores are very close to 0.5. In other words, if the distributions of these scores for the original and synthetic observations are similar and cannot be distinguished, it can be inferred that the distributions are closely aligned and the utility should be relatively high [27,32]. The two probability distributions can then be also compared by one estimate, the *pMSE*, which assesses the squared differences between the predicted probabilities and 0.5 (the value where the real and synthetic values are not distinguishable by the model).

**Syntactical Accuracy** refers to the plausibility of the data while preserving the logic of the real world, which requires preserving certain structural properties of the data. With respect to the analysed census data, we conducted a series of checks to verify the number of synthesized households, the number of synthesized only child households (synthetic data should not contain households consisting only of minors), and the number of synthetic mixed gender households. These values were then compared to their truth. Furthermore, we controlled for variables that may present natural constraints (e.g., age).

## 4 Results

A review of the literature, in conjunction with the application of the SDGs to the subject matter of this paper, revealed that in addition to the simPop [30] package, there might be a few other tools that meet all of the challenges that arise, such as handling missing values, the diversity of column types (including continuous, semi-continuous, count, categorical and ordered categorical), complex distributions, complex relationships between variables [15], as well as sampling weights when samples are not drawn with simple random sampling without replacement, cluster structures (such as a person in households) and hierarchical structures (such as municipalities in regions). Table 2 provides an overview of the capabilities of the employed SDGs, delineating their abilities and limitations in synthesising complex data.

---

<sup>1</sup> In addition to the logistic regression, other methods, such as decision trees or random forest, can be employed to calculate propensity scores.

**Table 2.** Synthetic data generators capabilities. \* except semi-continuous information (e.g. personal income in household structures, cf. Fig. 4) \*\* function syn.ipf for a limited number of categorical variables.

	synthpop	SDV	simPop	Mostly AI	Gretel	Realtabformer	arf
conditional (C) or joint (J) model	C/J**	C/J	C	J	J	C	C
sample or pop.	S	S	P	S	S	S	S
missing values	✓	✓	✓	✓	✓	✓	✗
mixed variables	✓	✓*	✓	✓*	✓*	✓*	✓*
sampling weights	✗	✗	✓	✗	✗	✗	✗
cluster structures by design	✗	✗	✓	✗	✗	✗	✗
hierarch. structures by design	✗	✗	✓	✗	✗	✗	✗
comp. costs	low	low/high	low	high	high	high	low

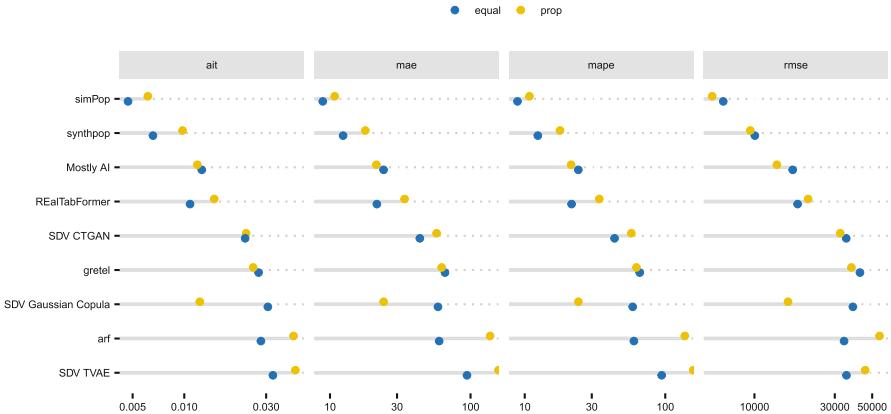
## 4.1 Data Utility

**Statistical Accuracy:** In the evaluation of utility, we start with the univariate evaluation, and then the multivariate distributions are assessed. A comparison was conducted between the synthetic distributions and the real data set for the variable “age.” When this variable was considered, the simPop population yielded the most accurate results, followed by synthpop and Mostly AI. The other approaches demonstrated significant discrepancies compared to the real dataset. Similar outcomes were observed when the other variables were evaluated.

Figure 2 displays the performance metrics Aitchison Distances (ait) [8], mean absolute error (mae), mean absolute percentage error (mape) and root mean squared error (rmse) of the different SDGs by the sample design. All of these metrics involve the comparison of a multi-way contingency table from synthetic and true data.

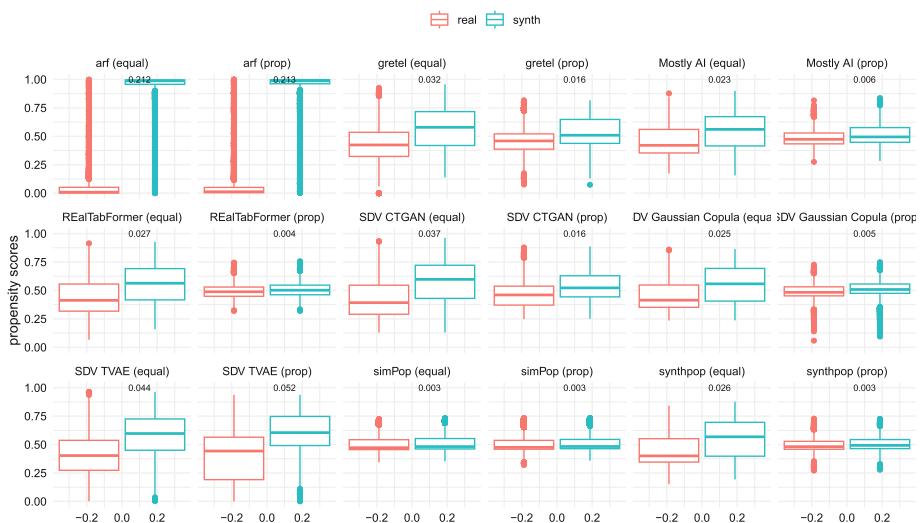
The simPop package demonstrated superior performance in both sampling designs in all observed measures, followed by synthpop.

For row-wise combined true and synthetic data, we fitted the logistic regression model described in Sect. 3.3. The distributions of the predicted propensity scores for all SDGs and both sampling designs are summarized in Fig. 3 as box-plots, which provide an overview of the similarity of their distributions. The pMSE [32] is also displayed in the figure as a metric. With regard to this metric, the synthetics were compared to the non-synthesized from  $U^*$ . One can observe a sample stratum effect in almost all SDGs. The propensity scores of the synthesized proportional sample  $S_{prop}^{**}$  are clearly more similar to those of  $U^*$  than the propensity scores of  $S_{equal}^{**}$ . The simPop and synthpop packages demonstrated the most favourable performance in terms of pMSE, but synthpop only provides good results for simple random sampling proportional to size, since it is not designed to use the information on sampling designs.

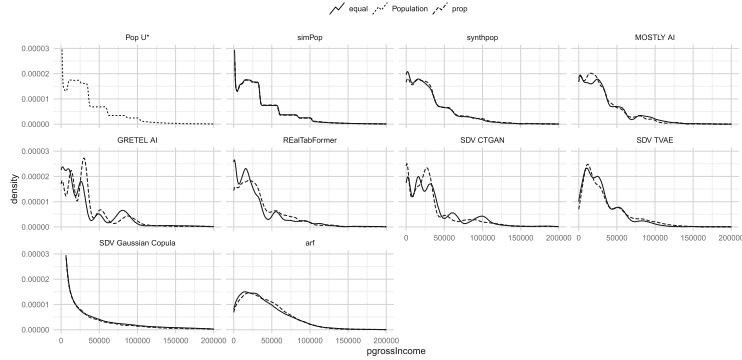


**Fig. 2.** Differences of synthetic data and true data from a three-dimensional contingency table on region, age and gender.

Figure 4 shows exemplarily the distribution of the truth and the SDGs compared. Note that personal gross income is semi-continuous, a lot of persons have an income, but also a particular part of persons have 0 income. simPop shows superior performance independently of the sampling design used, while others give rather unsatisfactory results.



**Fig. 3.** Propensity Score Boxplots (left: real, right: synthetic)



**Fig. 4.** Distribution of personal gross income for the true population and the synthetic data sets.

**Syntactical Accuracy:** Table 3 provides an overview of the results. The variable  $nhh$  represents the number of households in the synthesized sample  $S_{ji}^{**}$  and the variable  $\widehat{Nhh}$  represents the number in the estimated populations. Note that all methods provide household information (household ID, household-related information such as income components and household size) and household compositions provided by the age composition of people in households or their citizenship composition. The household ID and other household-related information are then simulated using all methods.

The number of households with only children is displayed on both the synthesized samples and the estimated population levels. The variables  $mghh$  and  $\widehat{Mghh}$  represent the number of mixed adult gender households, where at least one adult person in a common household is female, and one adult person in the same household is male.  $mghh$  represents the number in the synthesized samples  $S_{ji}^{**}$  and  $\widehat{Mghh}$  represents the number in the estimated populations.

It is evident that in all three observed metrics, `simPop` [30] outperforms its competitors. It is the only package that does not produce only children's households and produces the closest results to the synthetic truth  $U^*$  in terms of the number of households and the number of mixed-gender households. It should be noted that sample estimates are not provided for `simPop` because it generates a complete population (from which samples can be drawn if needed). SDGs such as `arf` [31] and `Gretel` [10] demonstrated limitations in their ability to comprehend the household structure of the data accurately.

**Table 3.** Synthactical Results

Tool and method	design	households			only children households			mixed gender households		
		nhh	$\widehat{Nhh}$	Diff	kidshh	$\widehat{Kidshh}$	Diff	mghh	$\widehat{Mghh}$	Diff
arf	equal	13,820	8,319,850	4,618,549	1,859	1,123,494	1,123,494	4	2,420	2,176,967
arf	prop.	13,474	8,316,345	4,615,044	1,912	1,179,966	1,179,966	9	5,561	2,173,826
GRETEL AI	equal	13,440	8,245,993	4,544,692	1,234	783,558	783,558	10	5,573	2,173,814
GRETEL AI	prop.	13,245	8,178,671	4,477,370	3,346	2,070,247	2,070,247	4	2,497	2,176,890
MOSTLY AI	equal	3,691	2,206,653	1,494,648	142	99,010	99,010	1,687	1,005,918	1,173,469
MOSTLY AI	prop.	3,566	2,201,512	1,499,789	129	96,990	96,990	1,556	960,981	1,218,406
REalTabFormer	equal	7,219	4,690,811	989,510	354	268,924	268,924	1,356	818,286	1,361,101
REalTabFormer	prop.	8,245	5,086,726	1,385,425	569	385,041	385,041	1,060	654,098	1,525,289
SDV CTGAN	equal	2,536	1,340,466	2,360,835	171	74,651	74,651	1,191	748,329	1,431,058
SDV GC	equal	4,899	3,047,787	653,514	246	174,165	174,165	1,794	1,103,632	1,075,755
SDV TVAE	equal	3,598	2,300,803	1,400,498	233	301,447	301,447	967	595,951	1,583,436
SDV CTGAN	prop.	2,477	1,529,983	2,171,318	203	129,206	129,206	1,117	688,953	1,490,434
SDV GC	prop.	4,820	2,975,426	725,875	243	170,246	170,246	1,646	1,015,925	1,163,462
SDV TVAE	prop.	3,209	1,978,211	1,723,090	209	277,848	277,848	729	448,999	1,730,388
synthpop	equal	4,918	3,006,013	695,288	106	143,276	143,276	1,810	1,080,365	1,099,022
synthpop	prop.	4,862	3,001,411	699,890	119	149,033	149,033	1,710	1,055,648	1,123,739
simPop	equal	-	3,682,556	18,745	-	0	0	-	2,205,779	26,392
simPop	prop.	-	3,686,672	14,629	-	0	0	-	2,165,390	13,997

Note:

Number of households (Nhh) in the synthetic truth ( $U^*$ ) = 3'701'301

Number of only children households (Kidshh) in the synthetic truth( $U^*$ ) = 0

Number of mixed gender adult households (Mghh) in the synthetic Truth( $U^*$ ) = 2'179'387

Regarding the problems of most methods to simulate household memberships. We note that it is rather impossible to learn from variables in the original data set (household ID, household size, sampling weights, age composition, citizenship composition, household income variables, and related household information in the data set, weights) which person belongs to the same household and how many households should be formed on which size. Note that in such real-world data sets, the correlation structure is weak, and there are only about 15.000 persons in about 6000 households. The number of households in a country is estimated using Horvitz-Thompson estimation using the simulated sampling weights. All the mentioned methods except simPop cannot deal with complex sampling designs and simulate sampling weights from the data and not by design. In other words, if you do not explicitly model the cluster structure, you get bad results on cluster structures.

## 5 Discussion and Conclusion

synthpop is more often compared in comparison studies in literature and achieved the second-best results in all metrics measured. synthpop consistently demonstrated favourable outcomes in SDG comparison literature, as evidenced by Little et al. [14], Pathare et al. [20], and Endres et al. [6], but it was never compared with simPop.

In contrast to the findings of Quian et al. [23], who observed that the adversarial random forest algorithm of their `synthcity` package outperformed algorithms such as CTGAN and TVAE on a static tabular data set, our study revealed that the adversarial random forest algorithm of the `arf` package [31] demonstrated sub-optimal performance in our complex data set, outperforming only the TVAE algorithm of the SDV [21] package and being worse than the CTGAN algorithm of the same package.

The SDG that showed the best performance was `simPop`, which exhibited superior results in all measured metrics. It provides consistent synthetic data, and it can deal with all challenges of real data, such as missing values, mixed type of variables, possible hierarchical and cluster structures of data (such as persons in households), and have solutions when data are not sampled with simple random sampling without replacement, widely used in practice when conducting surveys. It provides a set of machine learning methods that can be used for the generation of synthetic data.

Future work: A crucial aspect of comparing SDGs is efficiency, which encompasses the scalability of data dimensions and the speed at which SDGs can generate synthetic data, as well as memory usage. In this study, we presented our anecdotal evidence regarding efficiency, categorizing the methods as low, middle, and high without measuring it precisely.

In addition, improvements will be made to the SDGs used in this study. For instance, the `simPop` [30] package permits further calibration of auxiliary information when synthesizing data, necessitating verification. It is to be expected that the results of `simPop` will be even better with the optional built-in calibration.

The `RealTabFormer` package [28] offers a “relational” method using a sequence-to-sequence model, which must be tested and may yield superior results in complex structural data. In particular, given that the GPT-2 model was already one of the more effective models in our comparison and is part of this sequence-to-sequence model. In addition, the algorithms mentioned above for the relatively new `synthcity` [23] package require comparative analysis. Including additional complex data sets in the comparison is also necessary to form a more comprehensive and nuanced judgment.

Additionally, due to the page limit, we did not discuss the disclosure risk of synthetic data and fully drew attention to its utility and consistency.

However, the risk of disclosure for synthetic data should be relatively low, as discussed by Drechsler [5] and Templ [29]. The artificial nature of synthetic data provides high protection, but this is at the expense of lower utility compared to traditional methods [29]. Nevertheless, Drechsler [4] underscores that fully synthetic data are not free of risk, and the measurement of disclosure risks for fully synthetic data is so far still challenging. Consequently, further research into the risk-utility ratio of synthetic data is imperative.

Furthermore, the truth, the population  $U^*$ , could also be simulated by other methods. This is a highly non-trivial task, and further simulations may give different results.

**Acknowledgement.** This work was funded by the Swiss National Science Foundation with grant “*Harnessing event and longitudinal data in industry and health sector through privacy preserving technologies*” (grant number 211751).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## A Appendix A: Main Variables

**Table 4.** Main variables selected for the simulation of the Austrian EU-SILC data. Additional variables include personal and household income components and more categorical variables.

Variable	Name	Possible outcomes
Household ID	<i>db030</i>	1 to 3701301
Region	<i>db040</i>	9 regions
Household size	<i>hsize</i>	Number of persons in household
Age	<i>age</i>	Age in years
Gender	<i>rb090</i>	Female Male
Self-defined current economic status	<i>pl031</i>	work status
Citizenship	<i>pb220a</i>	Austria EU* Other*
Personal gross income	<i>pgrossIncome</i>	Sum of income components <sup>†</sup>
Household gross income	<i>hgrossIncome</i>	Sum of income components

\* combined categories

† contains structural zeros displayed as missing values for children’s personal gross income

## B Appendix B: Literature on the comparison of synthesizers.

Little [14] compared synthpop [18] with DataSynthesizer (Ping et al. [22]), CTGAN (Xu et al. [33]) and TableGAN (Park et al. [19]). synthpop outperformed all others in terms of data utility for twelve categorical variables from the 1991 Great Britain Individual Sample of Anonymized Records.

Pathare et al. [20] compared DataSynthesizer, Synthetic Data Vault (SDV) (Patki et al. [21]) and synthpop (using logistic regression), whereby DataSynthesizer and synthpop provided the best results based on 15 public data sets. A comparable outcome is observed by Endres et al. [6]. In their comparison study on tabular data, the authors demonstrated that SDGs, such as SMOTE and synthpop (non-parametric approach), exhibited greater performance in terms of proximity than GAN and VAE based SDGs, including SDV-GAN and SDV-VAE.

Kiran et al. [13] compared CTGAN of Synthetic Data Vault [21] package, PATE-CTGAN from the Smartnoise-synth package [26] and Data Synthesizer from the Data Synthesizer [22] package on 13 different tabular data sets. CTGAN and PATECTGAN showed the most effectiveness in mimicking the real data for all 13 datasets in their study.

Espinosa and Figueira [7] conducted a comparative analysis of DataSynthesizer [22] and CTGAN, in addition to other algorithms such as Borderline SMOTE, RealTabFormer [28] and CopulaGAN [21] on two tabular data sets. Their study revealed that Borderline SMOTE and RealTabFormer exhibited superior performance on utility measures.

Quian et al. [23] developed the Python library, *Synthcity*, with the objective of facilitating a comprehensive evaluation of the different SDG algorithms across multiple modalities and applications [23]. They compared the capabilities of their package in terms of data modalities and use cases to those of different existing SDG libraries, such as YData Synthetic, Gretel [10], SDV [21], Data-Synthesizer [22], SmartNoise [26] and nbsynthetic. The authors posit that their findings demonstrate the shortcomings of existing libraries, particularly in terms of censored features, composite data sets and irregular time series when it comes to data modalities. In terms of use cases, the competing libraries are missing features like cross-domain data augmentation and differential privacy. All of these areas' shortcomings can be addressed solely by the package *Synthcity* in their comparison. The authors further test the different algorithms in their package in case studies. With regard to static tabular data, 18 datasets from the OpenML benchmark were subjected to testing on the fidelity of synthetic data. The results indicated that the adversarial random forest algorithm, a tree-based generative model, emerged as the most effective approach, suggesting that tree-based generative models may offer a viable alternative to deep generative models for static tabular data.

## References

1. Alfons, A., et al.: The AMELI simulation study. Research Project Report WP6 – D6.1, FP7-SSH-2007-217322 AMELI (2011). <http://ameli.surveystatistics.net>
2. Alfons, A., Kraft, S., Templ, M., Filzmoser, P.: Simulation of close-to-reality population data for household surveys with application to EU-SILC. Stat. Methods Appl. **20**(3), 383–407 (2011). <https://doi.org/10.1007/s10260-011-0163-2>
3. Drechsler, J., Reiter, J.: Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB establishment survey. J. Off. Stat. **5**(4), 589–603 (2009)
4. Drechsler, J., Haensch, A.C.: 30 years of synthetic data. arXiv preprint <arXiv:2304.02107> (2023). <https://doi.org/10.48550/arXiv.2304.02107>
5. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. No. 201 in Lecture notes in statistics, Springer, New York (2011). <https://doi.org/10.1007/978-1-4614-0326-5>, OCLC: ocn733239576
6. Endres, M., Mannarapotta Venugopal, A., Tran, T.S.: Synthetic data generation: a comparative study. In: Proceedings of the 26th International Database Engineered Applications Symposium, pp. 94–102. IDEAS '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3548785.3548793>
7. Espinosa, E., Figueira, A.: On the quality of synthetic generated tabular data. Mathematics **11**(15) (2023). <https://doi.org/10.3390/math11153278>
8. Fačevicová, K., Hron, K., Todorov, V., Templ, M.: General approach to coordinate representation of compositional tables. Scand. J. Stat. **45**(4), 879–899 (2018). <https://doi.org/10.1111/sjos.12326>
9. Gesis: Series: European Union Statistics on Income and Living Conditions (EU-SILC) (2024). <https://www.gesis.org/en/missy/metadata/EU-SILC/>. Accessed 13 May 2024
10. GRETEL: GRETEL.AI: The synthetic data platform for developers (2024). <https://gretel.ai/>. Accessed 01 May 2024
11. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe **47**(260), 663–685 (1952). <https://doi.org/10.1080/01621459.1952.10483446>
12. Jordon, J., et al.: Synthetic data – what, why and how? arXiv preprint <arXiv:2205.03257> (2022). <https://doi.org/10.48550/arXiv.2205.03257>
13. Kiran, A., Kumar, S.S.: A methodology and an empirical analysis to determine the most suitable synthetic data generator. IEEE Access **12**, 12209–12228 (2024). <https://doi.org/10.1109/ACCESS.2024.3354277>
14. Little, C., Elliot, M.J., Allmendinger, R., Samani, S.S.: Generative adversarial networks for synthetic data generation: a comparative study. CoRR **abs/2112.01925** (2021). <https://doi.org/10.48550/arXiv.2112.01925>
15. Davila, M.F.R., Wolfram Wingerath, F.P.: Benchmarking tabular data synthesis for user guidance. In: Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference Co-located with the EDBT/ICDT 2024 Joint Conference, pp. 1–4, March 2024
16. MOSTLY.AI: MOSTLY.AI: Synthetic data generation and privacy-preserving analytics (2024). <https://mostly.ai/>. Accessed 01 May 2024
17. Münnich, R., Schürle, J.: On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No. 4, University of Tübingen (2003)

18. Nowok, B., Raab, G.M., Dibben, C.: Synthpop: bespoke creation of synthetic data in R. *J. Stat. Softw.* **74**(11), 1–26 (2016). <https://doi.org/10.18637/jss.v074.i11>
19. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.* **11**(10), 1071–1083 (2018). <https://doi.org/10.14778/3231751.3231757>
20. Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., Gawade, A.: Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *Int. J. Inf. Manag. Data Insights* **3**(2), 100177 (2023). <https://doi.org/10.1016/j.jjimei.2023.100177>
21. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410, October 2016. <https://doi.org/10.1109/DSAA.2016.49>
22. Ping, H., Stoyanovich, J., Howe, B.: DataSynthesizer: privacy-preserving synthetic datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management. SSDBM ’17, Association for Computing Machinery, New York, NY, USA (2017). <https://doi.org/10.1145/3085504.3091117>
23. Qian, Z., Davis, R., van der Schaar, M.: Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. arXiv preprint [arXiv:2301.07573](https://arxiv.org/abs/2301.07573) (2023)
24. Raab, G.M., Nowok, B., Dibben, C.: Assessing, visualizing and improving the utility of synthetic data (2021). <https://arxiv.org/abs/2109.12717>
25. Rubin, D.B.: Discussion of statistical disclosure limitation. *J. Off. Stat.* **9**(2), 461–468 (1993)
26. SmartNoise: smartnoise-sdk (2024). <https://docs.smartnoise.org/>. Accessed 17 May 2024
27. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data, June 2017. arXiv preprint [arXiv:1604.06651v2](https://arxiv.org/abs/1604.06651v2). <https://doi.org/10.48550/arXiv.1604.0665>
28. Solatorio, A.V., Dupriez, O.: Realtabformer: generating realistic relational and tabular data using transformers. arXiv preprint [arXiv:2302.02041](https://arxiv.org/abs/2302.02041) (2023). <https://doi.org/10.48550/arXiv.2302.0204>
29. Templ, M.: Statistical Disclosure Control for Microdata. Springer, Cham (2017). <https://doi.org/10.1007/978-3-319-50272-4>
30. Templ, M., Meindl, B., Kowarik, A., Dupriez, O.: Simulation of synthetic complex data: the R package simPop. *J. Stat. Softw.* **79**(10), 1–38 (2017). <https://doi.org/10.18637/jss.v079.i10>
31. Watson, D.S., Blesch, K., Kapar, J., Wright, M.N.: Adversarial random forests for density estimation and generative modeling. In: Ruiz, F., Dy, J., van de Meent, J.W. (eds.) Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 206, pp. 5357–5375. PMLR, 25–27 April 2023
32. Woo, M.J., Reiter, J.P., Oganian, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confid.* **1**(1), 111–124 (2009)
33. Xu, L., Veeramachaneni, K.: Synthesizing tabular data using generative adversarial networks. [arXiv:1811.11264](https://arxiv.org/abs/1811.11264) (2018). <https://doi.org/10.48550/arXiv.1811.11264>

# **Disclosure Risk Assessment**



# An Examination of the Alleged Privacy Threats of Confidence-Ranked Reconstruction of Census Microdata

David Sánchez<sup>1</sup>(✉) , Najeeb Jebreel<sup>1</sup> , Krishnamurty Muralidhar<sup>2</sup> , Josep Domingo-Ferrer<sup>1</sup>(✉) , and Alberto Blanco-Justicia<sup>1</sup>(✉)

<sup>1</sup> Department of Computer Science and Mathematics, CYBERCAT–Center for Cybersecurity Research of Catalonia, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain

{david.sanchez,najeeb.jebreel,josep.domingo,alberto.blanco}@urv.cat

<sup>2</sup> Department of Marketing and Supply Chain Management, Price College of Business, University of Oklahoma, 307 West Brooks, Adams Hall Room 10, Norman, OK 73019, USA  
krishm@ou.edu

**Abstract.** The threat of reconstruction attacks has led the U.S. Census Bureau (USCB) to replace in the Decennial Census 2020 the traditional statistical disclosure limitation based on rank swapping with one based on differential privacy (DP), leading to substantial accuracy loss of released statistics. Yet, it has been argued that, if many different reconstructions are compatible with the released statistics, most of them do not correspond to actual original data, which protects against respondent reidentification. Recently, a new attack has been proposed, which incorporates the confidence that a reconstructed record was in the original data. The alleged risk of disclosure entailed by such confidence-ranked reconstruction has renewed the interest of the USCB to use DP-based solutions. To forestall a potential accuracy loss in future releases, we show that the proposed reconstruction is neither effective as a reconstruction method nor conducive to disclosure as claimed by its authors. Specifically, we report empirical results showing the proposed ranking cannot guide reidentification or attribute disclosure attacks, and hence fails to warrant the utility sacrifice entailed by the use of DP to release census statistical data.

**Keywords:** U.S. Decennial Census 2020 · Statistical Disclosure Limitation · Reconstruction · Differential Privacy · Reidentification

## 1 Introduction

The U.S. Decennial Census is the world’s most prominent census data release, accounting for more than 330 million people. The policies implemented by the U.S. Census Bureau (USCB) —the statistical agency in charge of compiling,

managing, and releasing the U.S. Census data—carry therefore a great influence on the decisions made by many other statistical agencies.

For the Decennial Census 2020 release, the USCB decided to replace the statistical disclosure limitation (SDL) methodology they had been using in previous editions (based on data swapping) with a method based on differential privacy (DP). The reason for such a change was the alleged disclosure risks resulting from potential record reconstruction attacks on the prior 2010 Census release [1, 10]. These works claim it is possible to accurately reconstruct the individual records from the released census statistics and that this reconstruction entails serious privacy threats to citizens. To prevent accurate reconstruction, a DP-based method consisting in adding random noise to the released statistics was implemented in the 2020 release [2].

DP is a robust privacy-enhancing method when properly implemented [7], but it also introduces significant logical inconsistencies and inaccuracies in the protected data due to its random and perturbative nature [8, 21]. In fact, in many cases the 2020 Census DP-protected data have been manipulated to the point where relationships between variables are not possible on the ground: there are blocks with more households than household population, blocks with no population in households yet having occupied house units, or blocks with non-zero population but with no adults [16]. In this respect, the quality of block-level data in the 2020 release was so poor that the USCB acknowledged that block-level data should not be used for any meaningful analysis [24]. Given that census data are crucial for research and social decision-making, one can understand the consternation of potential users [11, 13, 19, 22].

Several authors have demonstrated that the claims of disclosure risk ensuing from the reconstruction attacks in [1] and [10] were vastly overstated [17, 19]. A major issue is that there are usually a lot of different reconstructions compatible with the released output, which makes it impossible for the attacker to know which is the good reconstruction (the one corresponding or closest to the original data) [17, 18].

After assessing the situation, the USCB decided to abstain from using DP to protect the American Community Survey for the foreseeable future since “it’s also not clear that differential privacy would ultimately be the best option” [4].

However, a recent work by Dick *et al.* claims it is possible to rank records reconstructed by an attacker to reflect how likely they are to belong to the original data [5]. The authors propose a *confidence-ranked reconstruction attack (CRR)* that reconstructs records from the published census statistics and ranks them by how frequently they appear in multiple reconstructions. The authors assert that their ranking can be used by an adversary to conduct a variety of targeted attacks on individuals because the highest-ranked reconstructed records have a high chance of appearing in the original data. As a conclusion, they raise “sober warnings on the privacy risks of releasing precise aggregate statistics of a dataset” and note that “the only defenses against [reconstruction attacks] are to introduce imprecision in the underlying statistics themselves, as techniques like

differential privacy do”. This is precisely what the USCB did in the 2020 Census release [12].

Despite the issues of the 2020 Census protection, CRR renewed the interest in using DP in census releases. Specifically, in a subsequent article entitled “Database Reconstruction Does Compromise Confidentiality”, [12]—by the current Chief Scientist at the USCB and her immediate predecessor—fully endorsed Dick *et al.*’s conclusions, both regarding the privacy threats of CRR, and the advice to use DP to protect the released statistics.

Due to the (observed) adverse consequences of using DP in Census releases (or in any data release, since that was not the scenario DP was designed for [3, 7]), the flaws of previous reconstruction attacks, and the influence that the USCB’s decisions may have on other statistical agencies or social science in general, the claims by Dick *et al.* deserve detailed scrutiny.

## Contributions and Plan

In this paper, we empirically demonstrate the inability of the CRR attack to threaten privacy in any meaningful manner. In particular, we show that: (i) the highest ranked records according to Dick *et al.*’s confidence ranking are also the most common and, hence, the most inherently protected against reidentification; (ii) rare or unique records, whose reconstruction might put the corresponding individuals at risk of disclosure, go unnoticed by the proposed confidence ranking; and (iii) the inaccuracy of the reconstruction and the large diversity of the non-existent records it generates (records that do not exist in the original data) render attribute disclosure attacks ineffective.

The rest of the paper is organized as follows. Section 2 reviews the CRR attack. In Sect. 3 we detail how we replicated Dick *et al.*’s experiments. Sections 4 and 5 assess the actual reidentification and attribute disclosure risks implied by the CRR attack. Finally, we present some conclusions.

## 2 Reviewing the Confidence-Ranked Reconstruction Attack

In and of itself, reconstruction poses no privacy risk unless *it is accurate*. In the following, we demonstrate that the confidence-ranked reconstruction that Dick *et al.* propose is ineffective both at accurately reconstructing Census records and at detecting, even approximately, records at risk.

The primary purpose of the CRR is to attach a confidence level to each reconstructed record that measures how likely it is for that record to appear also in the data they use as ground truth. This ground truth consists of synthetic microdata published by the USCB that closely resemble the real 2010 Census microdata. By “casting the reconstruction problem as an instance of large-scale, non-convex optimization, along with a subsequent step to convert non-continuous (*e.g.*, categorical) features back to their original schema” [5], they rank reconstructed record prototypes by how frequently they appear in multiple reconstructions.

A *record prototype* is a record type with some *multiplicity* (that is, the number of repetitions) in the microdata. We talk about prototypes rather than actual records because CRR is incapable of ascertaining the multiplicity of those prototypes and, therefore, *it cannot produce an accurate reconstruction of the synthetic microdata*.

Nevertheless, the authors interpret the rank of the reconstructed records as a measure of risk, because the empirical results they report show that the top  $k$  ranked (*i.e.*, most frequent) reconstructed prototypes were present in the synthetic data set ( $D$ ) in a large proportion. More specifically, they argue that the most confident/highest ranked records they reconstruct are at risk because they are those that are most likely to appear in the original data and, therefore, can be the target of a variety of privacy attacks, including “identity theft”.

Dick *et al.* measure the effectiveness of their CRR attack as the proportion of the top  $k$  ranked prototypes (*i.e.*, the most frequently reconstructed ones) that were present in  $D$ . According to their results, this proportion was near 1 for small values of  $k$ . However, since the measured proportion considers the number of record prototypes in  $D$  (without counting their repetitions), rather than the actual number of records (counting repetitions), the authors are neglecting the multiplicity of each prototype in  $D$ , which is key to privacy: a prototype appearing, say, 10 times in  $D$  means that 10 individuals share the same record values for the considered attributes and, therefore, those individuals are intrinsically protected against reidentification (they are 10-anonymous, in terms of the  $k$ -anonymity privacy model [20], and the probability of successfully reidentifying one of them is at most 1/10).

To better understand the importance of this aspect and to illustrate the practical ineffectiveness of the CCR attack, in the following, we exactly replicate the experiments done by Dick *et al.* on the same data, and report the number of repetitions in  $D$  of their reconstructed prototype (which the authors did not do).

### 3 Replicating Dick et al.’s Experiments

Our experiments have been done using the code and settings provided by Dick *et al.*<sup>1</sup>. The additional code we have written for our experiments is also available for reproducibility<sup>2</sup>. Specific details follow.

**Data Set:** We used the same subsets of synthetic U.S. Census microdata as ground truth. This dataset was released by the USCB and closely resembles the real 2010 Census microdata in terms of statistical characteristics. Specifically, we used the 2020-05-27 vintage Privacy-Protected Microdata File (PPMF) [23], which consists of 312,471,327 rows representing synthetic responses for individuals in the 2010 Decennial Census. The columns in the PPMF include attributes

---

<sup>1</sup> <https://github.com/terranceliu/rap-rank-reconstruction>.

<sup>2</sup> <https://github.com/NajeebJebreel/CRR-analysis>.

such as the respondent’s home location (state, county, census tract, and census block), housing type, sex, age, race, and Hispanic or Latino origin.

**Statistical Queries:** Dick *et al.*’s reconstruction was executed on the tables employed by the USCB for their internal reconstruction attack on the 2010 Census data. Each table defines a set of statistical queries that are performed on the (protected) Census microdata. These queries involve specifying column names, subsets of column domains, and census block or tract identifiers. The objective is to count the number of microdata rows that satisfy the specified criteria. The tables include P1 (total population), P6 (race), P7 (Hispanic or Latino origin by race), P9 (Hispanic or Latino and Not Hispanic or Latino by race), P11 (Hispanic or Latino and Not Hispanic or Latino by race for the population 18 years and over), P12 (sex by age for selected age categories), P12 A-I (sex by age for selected age categories iterated by race), PCT12 (sex by single year age), and PCT12 A-N (sex by single year age iterated by race). The P tables are released at the block level, while the PCT tables are released only at the census tract level.

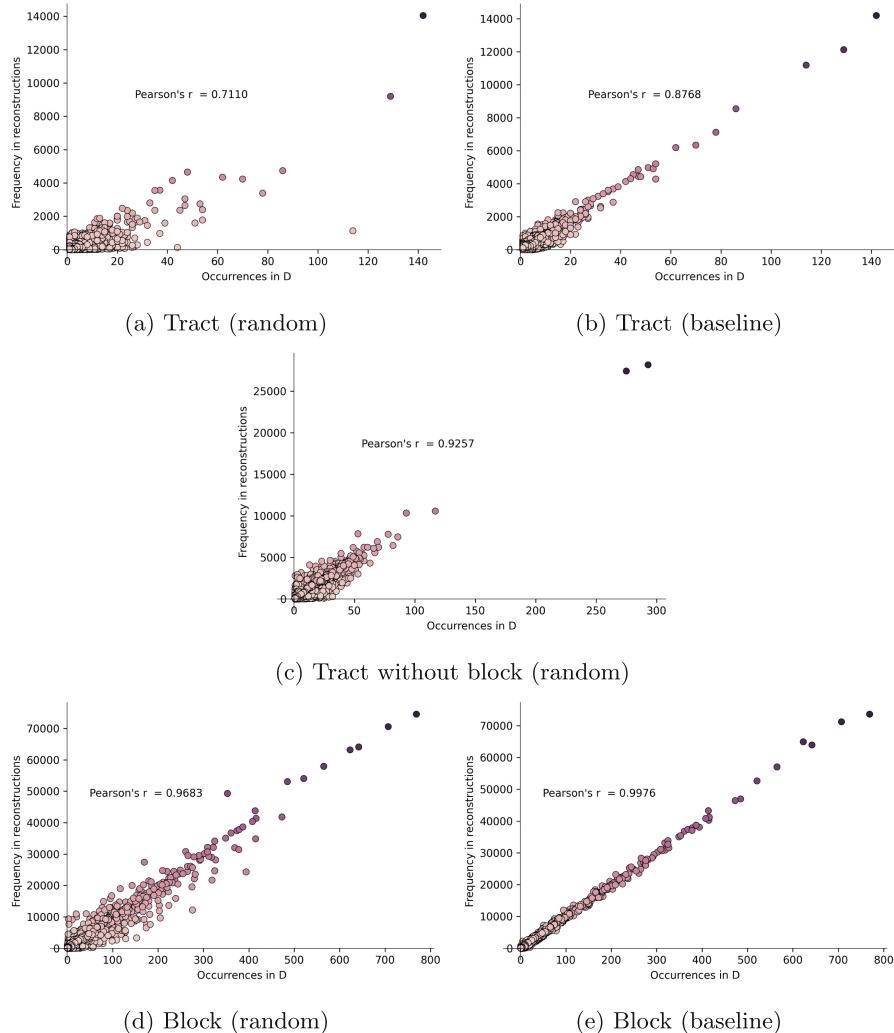
**Experiments:** We used subsets of the PPMF, which comprised all the rows belonging to specific census tracts or blocks. For tracts, we used the same random sample employed by Dick *et al.* On the other hand, blocks were selected according to the following (non-random) criteria described by the authors [5]: for each state, choose the block closest in size to the mean block size as well as the largest block; in addition, choose blocks closest in size to  $M/C$ , where  $M$  is the maximum block size in the state and  $C \in \{2, 4, 8, 16\}$ . This resulted in a total of 50 tracts and 300 blocks. Then, we reconstructed 100 data sets for each original data set, by employing both *baseline* and *random* initialization methods, while adhering to the same configuration and parameters. Baseline initialization follows the distribution of  $D$ —assumed to be publicly known—to set up the algorithm’s parameters, whereas random initialization leverages no prior information.

## 4 Results and Discussion

Having replicated Dick *et al.*’s experiments, we investigated the correlation between the multiplicities of records in  $D$  and their corresponding rank in the reconstructed data. Our aim was to determine whether highly ranked records threaten the privacy of individuals as Dick *et al.* claim.

To do so, we counted the number of occurrences of each record prototype in  $D$  and we compared it with its frequency in the reconstructions (which is the foundation of the confidence ranking). Figure 1 reports this comparison both at the tract and block levels and both with random and baseline initializations. We also report results for tracts without considering the block attribute (random initialization), which Dick *et al.* removed to ease tract reconstructions.

Our results clearly show that records with larger multiplicity in  $D$  have a higher frequency (rank) in the CRR. This dependency is very strong in all cases (Pearson’s correlation  $r$  between 0.71 and 0.99), but especially for block random



**Fig. 1.** Comparison between the multiplicities of record prototypes in  $D$  and their frequency in CRR.

or baseline initialization (where  $r > 0.96$ ). As a matter of fact, in all cases, the record with the largest multiplicity in  $D$  was also the most frequent/top-ranked prototype in the CRR. Since the confidence ranking is based on the frequency of record prototypes in multiple reconstructions, the top-ranked prototypes are the  $k$  most common prototypes in  $D$ ; that is, those with the greatest number of repetitions and, thus, those intrinsically most private. Therefore, all records top-ranked by CRR (*i.e.*, those claimed to be at risk) are intrinsically protected

against reidentification because “our privacy is protected to the extent that we blend in with the crowd” [9].

Conversely, rare or unique records in  $D$ , which are actually vulnerable to reidentification because of their rarity, do not appear among the top-ranked  $k$ . Hence, records really at risk stay paradoxically undetected by CRR.

By examining the distribution of  $D$ , we see that the percentage of unique records is 10.17% at the block level and 1.85% at the tract level. Hence, 89.83% are non-unique at the block level and 98.15% are non-unique at the tract level. Also, the multiplicity of the most common records is in the order of dozens at the tract level and in the order of hundreds at the block level. In terms of the well-known  $k$ -anonymity model [20], this means that unequivocal reidentification of any of those 89.83% and 98.15% records is not possible and that the probability of correct reidentification is as low as 1/140 at tract level (1/300 when dropping the block attribute), and 1/800 at the block level. From these results, it can be seen that CRR would still (wrongly) claim high rates if  $D$  was already protected via  $k$ -anonymity, *i.e.*, all prototypes in it had multiplicity  $k > 1$ . Therefore, it is clear that CRR does not capture the reidentification risk in  $D$ .

Notice that, since tracts are subdivided into blocks, one would expect the multiplicity in tracts to be larger than in blocks. However, the figures above show the opposite. These counterintuitive results are because we exactly replicated Dick *et al.*’s data sampling, where the sampling of blocks is independent from the sampling of tracts (both are done at the state level). More importantly, whereas tracts were randomly sampled, the largest and mean blocks per state were deterministically chosen. This yields blocks that tend to be larger than tracts. No reason is given in [5] for the different (non-random) criteria they employed to choose blocks. One motivation for choosing large blocks is that block-level data become more homogeneous because large blocks correspond to dense populations that tend to have inhabitants with similar profiles. This results in a greater amount of common records, which are those the confidence ranking scores higher.

Given the strong positive correlation between the multiplicity of prototypes and their confidence ranking, one might argue that the lowest-ranked record prototypes may be used to detect unique records in  $D$  and, therefore, indicate a privacy threat. In the following, we argue that this is not the case because of two reasons.

First, we must consider how Census data were protected in the 2010 release, on which all reconstruction attacks in the literature are based. Protection in that release was performed via *data swapping*. According to the USCB [25], the selection of the records to be swapped was highly targeted to the records with the most disclosure risk, that is, those that were unique in their block based on a set of key demographic variables. Also, the probability of being swapped had an inverse relationship with block size. Data from these households at risk were swapped with data from other households that had similar characteristics on a certain set of variables but were from different geographic locations.

According to the description above, the swapping methodology was targeted toward certain records that were defined as at risk of disclosure (*i.e.*, vulnerable records). Even though the exact details on how vulnerable records were determined are not public, the selection criterion was based on their uniqueness in their block [26]. Given that the statistics employed for CRR come from the swapped/protected data (and not the *original* data), we can deduce that reconstructed unique records are more likely to be protected (that is, swapped) records than original records.

Second, in order to be able to reconstruct (even protected) unique records in  $D$ , this should be done exhaustively and unequivocally. For this to be possible with CRR, records that are rare in the reconstructed data should also appear as rare in  $D$  with high certainty. To see how likely this is to happen, we calculated the following proportions:

- number of prototypes that occur once in  $D$  and occur once in the reconstructions over the total number of prototypes that occur once in the reconstructions;
- number of prototypes that occur twice in  $D$  and occur once or twice in the reconstructions over the total number of prototypes that occur once or twice in the reconstructions;
- number of prototypes that occur three times in  $D$  and occur once, twice, or three times in the reconstructions over the total number of prototypes that occur once, twice, or three times in the reconstructions.

Table 1 reports the above proportions as percentages, and they are extremely small. In other words, the certainty that a rare reconstructed prototype occurs in  $D$  is almost zero.

**Table 1.** Percentage of rare record prototypes in reconstructions that occur once, twice, or three times in  $D$ . Results are given for tracts, for tracts without block, and for blocks.

Level	Tract			Tract (w/o block)			Block		
	1/1	2/ $\leq 2$	3/ $\leq 3$	1/1	2/ $\leq 2$	3/ $\leq 3$	1/1	2/ $\leq 2$	3/ $\leq 3$
Occur. $D$ / freq. CRR	1/1	2/ $\leq 2$	3/ $\leq 3$	1/1	2/ $\leq 2$	3/ $\leq 3$	1/1	2/ $\leq 2$	3/ $\leq 3$
Random	0.09%	0.08%	0.05%	0.15%	0.16%	0.15%	0.20%	0.15%	0.12%
Baseline	0.09%	0.04%	0.02%	N/A	N/A	N/A	0.15%	0.06%	0.03%

To further illustrate the ineffectiveness of CRR as a reconstruction method, in Table 2 we report the percentage of reconstructed prototypes that did not occur in  $D$  (w.r.t. the total number of record prototypes in  $D$ ).

We can see that Dick *et al.*'s “optimization-based” method generates a very large number of non-existent record prototypes, which is 4 to 48 times larger than the number of prototypes in  $D$ . As discussed above, this results in very large uncertainty when reconstructing the rarest records in  $D$ , because there is a very high probability that the reconstructed prototypes do not exist in the original data.

**Table 2.** Percentage of reconstructed record prototypes that did not occur in  $D$ 

Level	Tract	Tract (w/o block)	Block
Random	4805.3%	703.2%	820.1%
Baseline	2223.6%	N/A	449.9%

On the other hand, by counting the percentage of records in  $D$  that did not appear in any of the reconstructions, we obtain the non-negligible figures reported in Table 3.

**Table 3.** Percentage of record prototypes in  $D$  that did not appear in the reconstructions

Level	Tract	Tract (w/o block)	Block
Random	8.16%	1.97%	3.31%
Baseline	3.07%	N/A	1.01%

## 5 On Attribute Disclosure Risk

In a yet more recent article [6], the authors argued that the main risk posed by CRR is actually *attribute disclosure*, that is, unequivocal inference by the adversary of confidential attribute values of known individuals, even without being able to reidentify them in the data set. This threat was not mentioned in their initial paper describing CRR [5], which only discussed privacy attacks related to identity disclosure. In that new paper, the authors fell short of supporting this new claim with any theoretical or empirical evidence. They just conjectured attribute disclosure through the following single fabricated example referred to a different dataset (the American Community Survey (ACS)):

Suppose that it is known that there are two 46-year-old married men with a particular racial designation and level of educational attainment within the [ACS] dataset. These features might be publicly known since they are not features the individuals intend to hide. Even if these two individuals match on all other features as well—that is, they share the same record prototype—if we are able to learn it, then we have learned facts about both of them (*e.g.*, their citizenship status, their income, etc.) that they may not have wanted to share. It does not matter that we cannot determine which record corresponds to which individual (the question does not even make sense, as the records are identical) since we have learned private information about both.

For the described privacy threat to lead to unequivocal inference of confidential attribute values on the known individuals, it is not enough to accurately generate the corresponding record prototype. The following must also hold:

1. The original data set must not contain any other records sharing the same values for the public attributes with the known individuals (in the above quotation, age, marital status, race, and education) but having different values for the confidential attributes (in the above quotation, citizenship status and income).
2. The original data set should be an exhaustive sample of the population to which the known individuals belong. Otherwise, we cannot be sure whether the known individuals are or are not present in the data set and, therefore, whether the potentially reconstructed records matching their public attributes correspond to them. Even though this holds for the Decennial Census (as long as the known individuals live in the U.S.), it is not the case for other non-exhaustive datasets, such as the American Community Survey (ACS) that Dick *et al.* used to illustrate the attribute disclosure threat.
3. CRR must not generate record prototypes sharing the same values for the public attributes but having diverse values for the confidential attributes.

Even though 1) and 2) may hold in practice, we have shown in Sect. 4 above that CRR was not only inaccurate but that it generated an enormous diversity of non-existent records (see Table 2). Therefore, 3) will not hold, and the diversity of the (erroneous) confidential values in the reconstructed records will prevent unequivocal attribute disclosure. In fact, it is well documented that diversity of confidential attributes effectively prevents attribute disclosure in  $k$ -anonymous-like data releases [14, 15].

## 6 Conclusions

We have shown that, despite the claims of [5] and [6], the proposed CRR attack is ineffective to i) detect the most privacy-sensitive records, ii) guide targeted reidentification attacks, or iii) guide attribute disclosure attacks. Moreover, we have shown that the “optimization-based” method by Dick *et al.* fails to reconstruct a significant proportion of original records while generating an enormous amount of records that do not exist in the original data set. The latter adds a very large uncertainty to disclosure inferences, whether aimed at reidentification or attribute disclosure. Having demonstrated that the CRR attack implies no privacy risks, we can conclude that it cannot be used to justify the use of (utility-damaging) DP methods —contrary to what [5] and [2] claim.

**Acknowledgments.** This research was funded by the European Commission (project H2020-871042 “SoBigData++”), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez and grant 2021SGR-00115), MCIN/AEI/ 10.13039/501100011033 and “ERDF A way of making Europe” under grants PID2021-123637NB-I00 “CURLING” and PRE2019-089210, and INCIBE and European Union NextGenerationEU/PRTR (project “HERMES” and INCIBE-URV Cybersecurity Chair).

## References

1. Abowd, J.: Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN (2021)
2. Abowd, J., Hawes, M.: Confidentiality protection in the 2020 U.S. Census of population and housing. *Annu. Rev. Stat. Appl.* **10**, 119–144 (2023)
3. Blanco-Justicia, A., Sánchez, D., Domingo-Ferrer, J., Muralidhar, K.: A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Comput. Surv.* **55**(8), 1–16 (2023)
4. Daily, D.: Disclosure avoidance protection for the American Community Survey (2022). <https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acss.html>. Accessed 3 May 2023
5. Dick, T., et al.: Confidence-ranked reconstruction of census microdata from published statistics. *Proc. Natl. Acad. Sci. U.S.A.* **120**(8), e2218605120 (2023)
6. Dick, T., et al.: Reply to Sánchez et al.: multiplicity does not protect privacy. *Proc. Natl. Acad. Sci. U.S.A.* **120**(8), e2304263120 (2023)
7. Domingo-Ferrer, J., Sánchez, D., Blanco-Justicia, A.: The limits of differential privacy (and its misuse in data release and machine learning). *Commun. ACM* **64**(7), 33–35 (2021)
8. Dove, I.: Applying differential privacy protection to ONS mortality data. Pilot study (2021). <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/applyingdifferentialprivacyprotectiontoonsmortalitydatapilotstudy>. Accessed 23 May 2023
9. Gehrke, J., Hay, M., Lui, E., Pass, R.: Crowd-blending privacy. In: Safavi-Naini, R., Canetti, R. (eds.) *Advances in Cryptology – CRYPTO 2012*. CRYPTO 2012. LNCS, vol. 7417, pp. 479–496. Springer, Berlin, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-32009-5\\_28](https://doi.org/10.1007/978-3-642-32009-5_28)
10. Hawes, M.: Reconstruction and reidentification of the Demographic and Housing Characteristics file (DHC) (2022). <https://www2.census.gov/about/partners/cac/sac/meetings/2022-09/presentation-reconstruction-and-re-identification-of-dhc-file.pdf>. Accessed 14 Mar 2023
11. Hotz, V., et al.: Balancing data privacy and usability in the federal statistical system. *Proc. Natl. Acad. Sci. U.S.A.* **119**(31), e2104906119 (2022)
12. Keller, S., Abowd, J.: Database reconstruction does compromise confidentiality. *Proc. Natl. Acad. Sci. U.S.A.* **120**(12), e2300976120 (2023)
13. Kenny, C., Kuriwaki, S., McCartan, C., Rosenman, E., Simko, T., Imai, K.: The use of differential privacy for census data and its impact on redistricting: the case of the 2020 U.S. Census. *Sci. Adv.* **7**(41) (2021)
14. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: privacy beyond k-anonymity and l-diversity. In: 23rd IEEE International Conference on Data Engineering (ICDE'07), pp. 106–115 (2007)
15. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkatasubramaniam, M.: L-diversity: privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1**(1), 3–es (2007)
16. Menger, G.: Using 2020 Census data (2021). <https://applyedgraphic.com/2021/09/using-2020-census-data/>. Accessed 9 May 2023
17. Muralidhar, K.: A Re-examination of the census bureau reconstruction and reidentification attack. In: Domingo-Ferrer, J., Laurent, M. (eds.) *Privacy in Statistical Databases*. PSD 2022. LNCS, vol. 13463, pp. 312–323. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_22](https://doi.org/10.1007/978-3-031-13945-1_22)

18. Muralidhar, K., Domingo-Ferrer, J.: Database reconstruction is not so easy and is different from reidentification. *J. Off. Stat.* **39**(3), 381–398 (2023)
19. Ruggles, S., Riper, D.V.: The role of chance in the Census Bureau database reconstruction experiment. *Popul. Res. Policy Rev.* **41**, 781–788 (2022)
20. Samarati, P.: Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001)
21. Santos-Lozada, A., Howard, J., Verdery, A.M.: How differential privacy will affect our understanding of health disparities in the united states. *Proc. Natl. Acad. Sci. U.S.A.* **117**(24), 13405–13412 (2020)
22. Schneider, M.: Researchers ask Census to stop controversial privacy method (2022). <https://www.usnews.com/news/business/articles/2022-08-08/researchers-ask-census-to-stop-controversial-privacy-method>. Accessed 15 May 2023
23. U.S. Census Bureau: Developing the DAS: Demonstration data and progress metrics (2020). <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planningmanagement/process/disclosure-avoidance/2020-das-development.html>
24. U.S. Census Bureau: Disclosure avoidance for the 2020 Census: An introduction (2021). <https://www2.census.gov/library/publications/decennial/2020/2020-census-disclosure-avoidance-handbook.pdf>. Accessed 15 May 2023
25. Zayatz, L., Lucero, J., Massell, P., Ramanayake, A.: Disclosure avoidance for Census 2010 and American Community Survey five-year tabular data products. Technical report. RRS2009-10, Census Bureau (2009). <https://www.census.gov/content/dam/Census/library/working-papers/2009/adrm/rrs2009-10.pdf>. Accessed 9 Mar 2023
26. Zayatz, L., Lucero, J., Massell, P., Ramanayake, A.: Disclosure avoidance for Census 2010 and American Community Survey five-year tabular data products (2009), <https://www.census.gov/content/dam/Census/library/working-papers/2009/adrm/rrs2009-10.pdf>, accessed 9 May 2023



# Synthetic Data: Comparing Utility and Risk in Microdata and Tables

Simon Xi Ning Kolb<sup>(✉)</sup>, Jui Andreas Tang<sup>(✉)</sup>, and Sarah Giessing

Federal Statistical Office of Germany, 65180 Wiesbaden, Germany  
[{simonxining.kolb, andreas.tang, sarah.giessing}@destatis.de](mailto:{simonxining.kolb, andreas.tang, sarah.giessing}@destatis.de)

**Abstract.** Synthetic data has begun to show potential as an alternative to traditional SDC methods in specific use cases. This development and the increasing research efforts further hint at an emerging role in future privacy protection. However, since data synthesis predominantly happens at microdata level, development of utility and risk metrics is also focused on this domain. Statistical agencies on the other hand limit data publication mostly to aggregates, by selecting various subsets of variables for cross tabulation. We analyze the correlations between microdata and tabular data metrics for assessing utility and risk. Using a large real life data set as an example for data synthesis, we show that certain global metrics may disproportionately represent small subsets of variables, making them an inappropriate estimator for the quality of aggregates. On the other hand, we show strong similarities between certain microdata level risk metrics and risks of group disclosure in aggregated data.

**Keywords:** Synthetic Data · SDC · Data Privacy · pMSE · TCAP · SUDA

## 1 Introduction

The thorough protection of official statistics, typically presented as multidimensional tables, is a challenging task. Although the SDC department of the German Federal statistics office is armed with a range of tools, their application is rarely straightforward. The prevailing method involves suppressing disclosive table cells. To avoid disclosure-by-differencing attacks, secondary suppression or coarsening is applied additionally. Another common method is the Cell Key Method (CKM), which adds controlled noise to table cells instead of complete suppression. Occasionally however, we are faced with statistics which do not align well with either approach. The road traffic accidents dataset fall into this category. Whenever a traffic accident happens on a public road, the police officer on duty files a report describing the accident. These files are then transferred to the statistical offices of the states where they undergo certain plausibility checks before being forwarded to the Federal Statistical Office.

These statistics consist of a huge dataset with more than 500 mixed-type variables. The overwhelming majority of variables are deemed non-sensitive.

Some sensitive variables are nonetheless included in very coarse aggregations and published before the main body of tables, which do require a SDC procedure. Additionally, the statistical office maintains a geocoded open data file of non-sensitive variables called the *German Accident Atlas*. The combination of these two data puts heavy restrictions on the effectiveness of post-tabular SDC methods. Due to the published table margins, many internal table cells need to be targeted in secondary suppression, leading to drastic utility reduction. Similarly, CKM noise is restricted to internal cells. Due to many published table margins, an attacker can compute accurate value ranges for protected cells.

In this work, we tackle the SDC problem via synthetic data. Our approach is to synthesize the sensitive variables, without modifying the variables included in the open data file and the pre-released tables. We use the R package *synthpop* [12] to generate synthetic microdata. The domain experts can then use the generated synthetic microdata as input when producing output tables for publication. Ideally, one would evaluate the quality and risk of the synthetic tables beforehand. This is however not possible, since many of the tabular structures are not known in advance. Therefore, it would be beneficial to estimate the quality of potential aggregates directly from the microdata. In order to verify this, we aim to investigate the validity of certain microdata level risk and utility measures as a proxy for the quality of arbitrary tabulations.

Throughout the paper, we refer to aggregated data as a “table(s)” or “tabular data”. Whenever we talk about non-aggregated microdata, we will use the term microdata.

## 2 Synthetic Data

The idea of using synthetic data for limiting disclosure risk stems from a paper by Rubin [17]. There he maps the task of disclosure control to an imputation problem, treating unsampled records as missing values. By doing so, one can then use an imputation model to create synthetic samples which replace the missing data. This approach is what we would call a fully synthetic dataset nowadays. A less perturbative approach, called partially synthetic data, was published by Little [10]. He proposed to limit the synthesis to sensitive variables only, thus leaving the dataset partially unchanged. By changing only a subset of variables, the resulting data quality is improved, albeit the risk of disclosure potentially increases. This approach is especially useful for controlling data quality of aggregates, since these tend to exhibit a lower disclosure risk.

Both systems of data synthesis rely on either a joint or sequential modelling framework. The joint modelling approaches intend to approximate the fully joint distribution. In the recent year they have been studied increasingly by the deep learning community through the use of models based on Generative Adversarial Networks (GAN) [6]. The sequential approach on the other hand, factorizes the full distribution into many conditional distributions. This way, sensitive variables are synthesized sequentially and potentially conditioned on each other. One can freely choose an appropriate model that describes the conditional distributions

for each variable. Model choices range from simple linear regression to more complex machine learning models [1, 3, 15].

## 2.1 Application to the Road Traffic Accidents Dataset

We started by reducing the dataset to 85 variables by removing redundant variables. The next step was a feature selection step with respect to our 9 sensitive variables (7 categorical, 2 numerical) in order to speed up data synthesis. To do this, we used the R Package *Boruta* [9], which offers a feature selection procedure based on random forests (among others). Reducing the number of variables is a necessity when synthesizing big datasets based on the CART algorithm. According to [7], categorical predictors pose a computational challenge for CART because the number of possible partitions grows exponentially with the number of levels. To further cut down the run time, we applied varying degrees of binning to categorical variables when they were used as predictors and determined the number of predictors based on Boruta. We note that binning was only done to predictors, never to the variable which is being synthesized. Categorical predictors with high variable importance and cardinality were *target encoded* [11] in specific situations. We then used *synthpop* to synthesize the 9 sensitive variables according to the order recommended by [13].

While *synthpop* offers a rule functionality to preserve plausibility in an *if <condition> then <x = “new value”>* fashion, we had to deal with a case where the condition limits the plausible values to a subset rather than a single value. We accomplished this by moving these variables to the end of the synthesis sequence and performing a rejection sampling with externally built *rpart* [21] models.

## 3 Evaluation Methods

The synthetic data community has developed a large body of measures for determining utility and risk of synthetic data [2]. However, the vast majority of evaluation methods tend to operate on microdata level. In this work, we consider the scenario where statistical agencies only publish aggregated data. Depending on the statistics and use cases, the coarseness and dimensionality of such tables may vary greatly. In order to investigate the relationship between measures for these two distinct data types, we conduct our analysis on both microdata and tables.

For tabular data evaluation, we chose three non synthesized and three synthesized variables as well as the geo code for cross tabulation. Next we specified the range of probable table dimensions to be between 2 and 5. Following these specifications we built all possible tables crossing the 6 selected variables, leading to a large set of tables with 2, 3, 4 and 5 table dimensions for different aggregation levels. All measures addressing tabular data in the following chapters are computed on this set of tables.

**Table 1.** N-way tables with 6 variables. Only tables with at least one synthetic variable are considered

table dimension	2	3	4	5	total
number of tables $n(d)$	12	19	15	6	52

### 3.1 Utility Measures

Commonly, microdata measures can be divided into global and analysis-specific measures. While the former strives to evaluate the distributional similarities between original and protected data, the latter tends to focus on measuring the output similarity of a given analysis task like the estimated coefficients from a linear regression.

Information loss in tabular data [18] is roughly split into distance based measures, variance analysis and the investigation of correlations. Among these, distance based measures are the most straightforward and can be easily applied to high dimensional frequency tables. Typical examples are absolute cell deviations (measuring the deviation of cell counts), relative cell deviations or Hellinger distance.

**Microdata Evaluation.** We limited our analysis to the propensity score mean squared error ( $pMSE$ ) [22] because it can handle mixed-type variables. To compute the propensity scores, one first stacks the original and synthetic data and adds a boolean variable indicating whether a record  $i$  with  $i = 1, \dots, N$ , where  $N$  is the number of observations in the stacked dataset, is synthetic or not. Next, the probability  $p_i$  that the record stems from the synthetic dataset is computed by a classification model (CART) fitted on the stacked dataset. The  $pMSE$  is defined as  $\frac{1}{N} \sum_{i=1}^N (p_i - c)^2$ , with  $c = \frac{n_{syn}}{N}$  being the fraction of synthetic records ( $c = 0.5$  in our case). This quantity compares the distributions of the propensity scores from original and synthetic dataset. The smaller the measures the higher the general utility of the synthetic dataset. One downside of the  $pMSE$  is its tendency to increase as the complexity of the model used to calculate the propensity scores (propensity model) grows, even in cases where the model is correctly specified. To address this issue, we also used the  $pMSE$  ratio [19], which is described more detailed in Appendix B. Microdata measures were computed on the whole dataset with 85 variables.

**Tabular Data Evaluation.** We measure tabular data quality using the Hellinger distance, repeating the computation on each of the 3 aggregation levels, state, district and municipality. Cell counts were normalized beforehand so that the sum of all normalized cell counts is equal to 1 and therefore the Hellinger distance is between 0 and 1.

### 3.2 Risk Measures

Measuring risk is generally more difficult than measuring utility. To develop a risk measure, one must first design the attack scenario, which depends on the data product, publication process, agency requirements, and synthesis approach.

Fully synthetic data is generally speaking less susceptible to attacks than partially synthetic data, since the degree of changes to the data is much different. In a partially synthetic setting, some variables are left unchanged and allow for matching with records of the original dataset. When synthesizing key identifiers, the emphasis lies on preventing re-identification. Synthesizing the sensitive variable reduces the probability of attribute disclosure [5]. The former has been thoroughly investigated in [16]. Hornby and Hu created the R package *IdentificationRiskCalculation* [8] for measuring identification risk based on matching categorical or continuous key variables. While these studies focus on re-identification, our present work solely deals with *attribute disclosure*.

We specifically considered one rather recent measure called *Targeted Correct Attribution Probability* (TCAP) [20]. The TCAP computation starts by selecting synthetic microdata records with  $l$ -diversity = 1 with regards to a target variable (see Table 2(a)). [20] describes these records as having *Within Equivalence Class Attribution Probability* (WEAP) = 1. We then identify records with **key** combinations identical to those of the previously selected records, now among the original data ((4, 12) and (5, 12)). Among these original records, we calculate the proportion of records that have the same **key and target** (e.g. keys (4, 12) and target (4) match for original records 3 and 4, but not for 5). This record-level quantity, called TCAP, describes the probability that an intruder who knows the key variables, can guess the true target value when focusing on records with WEAP = 1.

**Tabular Data Evaluation.** Tables of the road traffic accidents dataset are mostly endangered by cases of group disclosure (GD): *Person A and neighbor N live in a small town. A knows that his neighbour N drove his car into a ditch next to the main road recently. He looks up a corresponding table for his municipality and discovers, that in 2022, all accidents on the main road were related to alcohol consumption. A has found out that N has been driving drunk.* For the risk measure, we extracted the cells in the synthetic and original tables that represent a GD, called disclosive cells. We compute the set overlap between these two sets of cells and use:

$$\text{risk} = \frac{\#\text{common disclosive cells}}{\#\text{synthetic disclosive cells}}. \quad (1)$$

Note that one does not have to include all of the key variables to obtain a group disclosure. A subset of the key variables may be sufficient to obtain a group disclosure.

The interpretation of Eq. (1) is as follows: Given a case of GD in the synthetic table, what is the probability this case reveals real information. We do not count

GD cases, when the disclosed value appears with a frequency of  $\nu \geq 90\%$  in the original data. We argue that this equates to the result of a naive guess.

**Microdata Evaluation.** We see from the above, that GD in a synthetic dataset is identical to the statement  $l$ -diversity = 1 or WEAP = 1. With this, we may understand TCAP as a generalization of the tabular data risk measure. In Table 2(a), synthetic GDs are given by the two key combinations (4,12) and (5,12), while only (5,12) corresponds to a GD in the original data (Table 2(b)) with  $TCAP = 1$ .

When computing the proportion of unique key combinations with  $TCAP = 1$  among all unique key combinations that contribute to a synthetic GD (or WEAP = 1), namely the combinations (4, 12) and (5, 12), we get a measure similar to the tabular data measure in Eq. (1). Both give a risk value of 50%. We define this minor modification of TCAP as

$$\text{TCAP ratio} = \frac{\#\text{unique combinations with } TCAP = 1}{\#\text{unique combinations with } WEAP = 1}. \quad (2)$$

This measure gives us the proportion of GDs in the synthetic dataset that are also GDs in the original dataset on microdata level.

**Table 2.** Example of synthetic and original microdata. GD are bold.

(a) synthetic						(b) original			
record	Key 1	Key 2	target	WEAP	TCAP	record	Key 1	Key 2	target
1.	1	2	3	0.5	*	1.	1	2	<b>3</b>
2.	1	2	5	0.5	*	2.	1	2	<b>3</b>
3.	4	12	<b>4</b>	1	2/3	3.	4	12	4
4.	4	12	<b>4</b>	1	2/3	4.	4	12	4
5.	4	12	<b>4</b>	1	2/3	5.	4	12	9
6.	5	12	<b>10</b>	1	1	6.	5	12	<b>10</b>

\* irrelevant, since  $WEAP < 1$

A different approach for risk estimation is the SUDA algorithm [4], which was originally designed to measure the risk of a sample disclosing population uniques. The key concept is a *Minimal Sample Unique* (MSU). A MSU is a unique combination of  $n$  (user defined) variables, such that no unique combination within a subset of those  $n$  variables can be found. SUDA does not differentiate between key and target variables.

The record corresponding to (Key 1 = 5) in Table 2 (right) is a MSU and a GD case at the same time. More MSU can be found in the Target column, namely 9 and 10, of which only Target = 10 corresponds to a GD. In this example, the correspondence between MSU and GD seems minor. We note that in practice,

most GD cases actually correspond to unique key combinations, which in turn also represent MSUs.

Usually the number of MSUs is used to produce a score [4]. Instead we compare the MSU ( $n = 6$ ) of the synthetic and original microdata and compute:

$$\text{risk} = \frac{\#\text{common MSU records}}{\#\text{synthetic MSU records}}. \quad (3)$$

A MSU record is any record, that contains a MSU. We refer to *Common MSU records* when the MSU appears in both datasets, synthetic and original.

## 4 Results

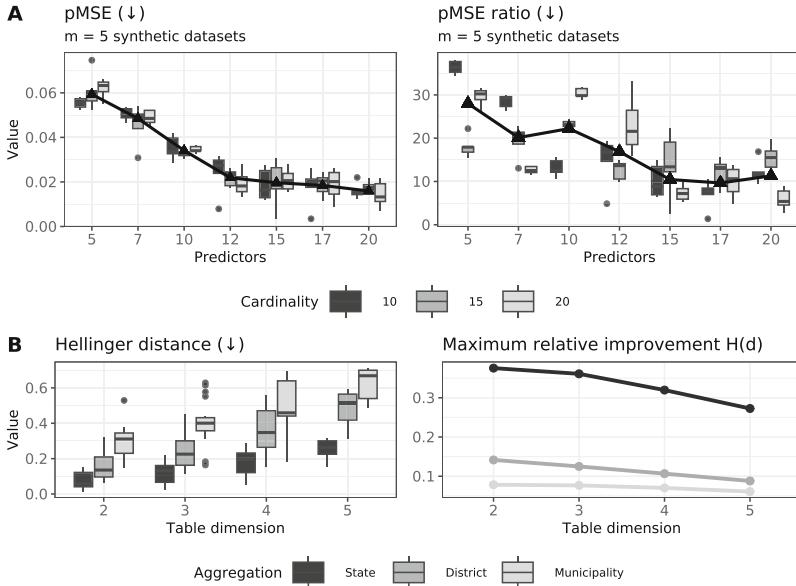
Test results were obtained by running *synthpop* with varying number of predictors and predictor variable cardinality. We ran 21 configurations and produced 5 synthetic datasets for each. Every synthetic microdataset was then used to produce tables with 2, 3, 4 and 5 dimensions using the 6 variables we selected (according to Table 1).

We show the evaluation results with respect to utility and risk in the following section.

### 4.1 Utility

Section A of Fig. 1 shows the microdata metrics *pMSE* and the *pMSE ratio* computed on the **whole dataset** consisting of 85 variables. We varied the number of predictors from 5 to 20 (represented by the x-axis), and binned categorical variables to a maximum cardinality of 10, 15 and 20 (represented by the box color). The improvement of global data quality is clearly visible for the *pMSE* plot. According to the line representing the averaged *pMSE* grouped by predictors, we observe a seemingly linear improvement that slows down significantly as the number of predictors increases to 12. The reason for this saturation behaviour lies in the way we performed the selection of the predictors. As mentioned in Sect. 2.1, the feature selection algorithm ranks potential predictors in a decreasing order of variable importance. When we sequentially increased the number of predictors, we started with the most important variables and continued with less important ones. The Fig. 1 (left) reflects the fact, that most of the predictive power has been captured in around 12 variables. While variable cardinality does play a role, its effects are significantly weaker for the range we tested.

The *pMSE ratio* on the other hand seems to depend much more on predictor variable cardinality without any clear tendencies. Furthermore, it seems that a distinct linear downtrend of the *pMSE ratio* with an increasing number of predictors may not be apparent, particularly with few predictors. We attribute this distinct behaviour to the sensitivity of the null expectation with regards to cardinality. The changes in predictor cardinality can lead to varying distributional differences between the  $m$  synthetic datasets, which in turn directly affect



**Fig. 1.** Comparing original and synthetic microdata and tables. Subplots A show the global utility *pMSE* and *pMSE ratio* in relation to the number of predictors and variable cardinality. Subplot B (left) shows the Hellinger distances for one synthesizer configuration. B (right) shows the maximum improvement over all synthesis setting for the Hellinger distance averaged over all tables of a specific dimensionality.  $\downarrow$  lower the better,  $\uparrow$  higher the better.

the null expectation (see Appendix B). In [14], a *pMSE ratio* below 10 defines acceptable utility for the synthesis.

We can see in Fig. 1B that high aggregation and low table dimensionality lead to smaller Hellinger distances. Both factors directly control the sparseness of table cells. High-dimensional tables at the municipality level have many sparsely populated cells, especially in regions with low population density. This, in turn, means, that the local distribution of the data will be represented by a very small sample size. We recall that CART does the synthesis on a per-record level. When we study the synthetic distribution at a low level of aggregation, we measure the distributional similarities based on a small sample of the data generator. Measuring differences based on small sample sizes will most likely lead to larger distributional differences due to sample variance. By gradually aggregating the synthetic microdata, we increase the number of (partially) synthetic records contributing to the synthetic distribution, making comparisons to the original distribution more reasonable. This property of data synthesizers can in some ways pose a fundamental problem for statistical agencies that want to publish at very low granularity without sacrificing data quality. Requiring

strict distributional similarity at ultra-fine granularity may essentially equate to demanding a close replica of the original data.

In section B (right) of Fig. 1, we displayed the improvement of tabular data quality. We plot the following quantity:

$$\mathcal{H}(d) = \max_{i,j} \left( \frac{|\hat{H}_{s_i,d} - \hat{H}_{s_j,d}|}{\hat{H}_{s_i,d}} \right) \quad (4)$$

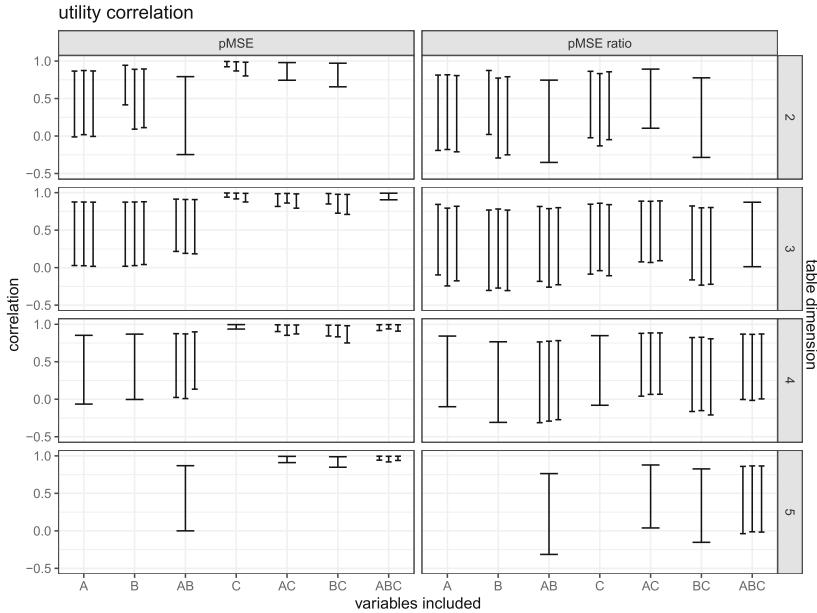
$$\text{with } \hat{H}_{s_i,d} = \frac{1}{m n(d)} \sum_{i,m_i} H_{s_i,d,i,m_i} \quad , \quad (5)$$

where  $i$  denotes one particular table among the  $n(d)$  tables for dimensionality  $d$ ,  $s_i$  denotes one configuration of the synthesizer,  $m_i$  the  $i$ th synthetic dataset and  $H$  denotes the Hellinger distance between original and synthetic table. By averaging over  $m$  and  $n(d)$ , we focus only on the size of the table. The quantity  $\mathcal{H}(d)$  describes the maximum relative improvement in Hellinger distance for one spatial aggregation level. The aforementioned effect of aggregation extends to the maximally achievable utility improvement in section B (right). It shows that advantageous synthesizer configurations affect the aggregation levels very differently, making general tabular data quality improvement non trivial.

## 4.2 Correlation Analysis

For the purpose of relating microdata scores to tabular data metrics, we did a correlation analysis. For each synthesis configuration, we averaged the *pMSE* (*pMSE ratio*) over  $m = 5$  synthetic datasets. Correspondingly, we averaged the Hellinger distances for all possible tables over the five synthetic datasets. This leaves us with one *pMSE* (*pMSE ratio*) and 52 (total number of tables, see Table 1) Hellinger distances per synthesis configuration.

We report the correlations in Fig. 2. The x-axis acts as a grouping variable, indicating which of the three synthetic variables (A, B, C) are included in a table. Empty positions arise because certain table sizes cannot accommodate certain cross tabulations. Groups of multiple intervals are due to tabulations where the non-synthesized variables vary. Confidence intervals were computed with Fisher z-transformation. The left plot indicates that one variable has a significant impact on the correlation between *pMSE* and the Hellinger distance. The synthesized variable C alone drives the correlation, seemingly independent of other participating variables. Tables that do not include variable C show no statistically relevant correlation. It seems that there are no significant differences between the confidence intervals that belong to the same group, which are tables with the same synthesized but different fixed variable set. This finding is largely independent from the table size. Since the microdata utility is by definition the same for every table, absent correlation must be due to the varying trajectory of tabular data utility over the different synthesizer configurations. This in returns implies that data utility has increased heterogeneously for the synthesized variables. Due to the *pMSE* relying on a classifier, emphasis on specific variables is



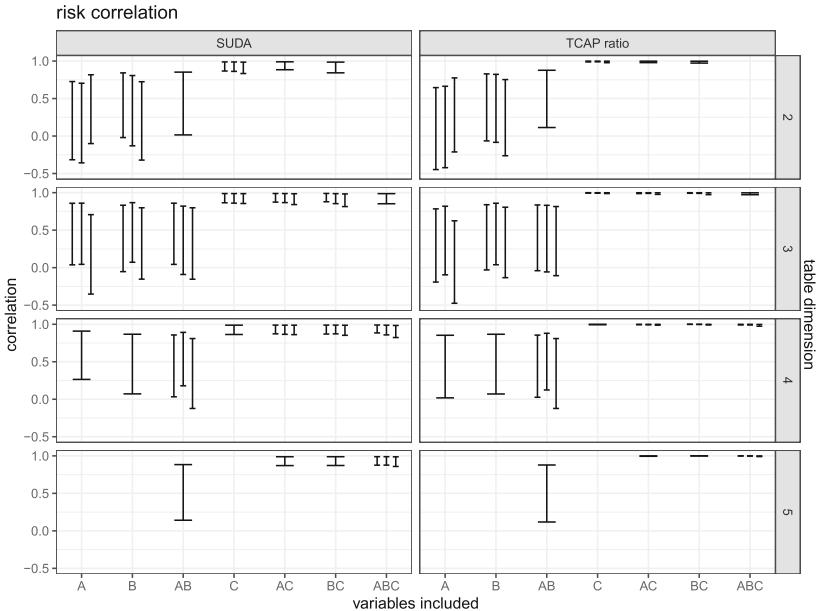
**Fig. 2.** Correlation between microdata utility and Hellinger distances for all possible tables. The x-axis describes the table type, denoting which synthesized variables are included in the table. We show the results for the highest aggregation. A value of 1 implies maximal correlation between tabular and microdata metrics.

to be expected, since the underlying CART algorithm will search for the most expressive variable for prediction. The categorical variable C differs from the other variables in terms of higher cardinality ( $> 60$  levels) and strong correlation with many other variables. A figure showing the improvement of Cramer's V is included in Fig. 5 in Appendix A. We note also, that the other two synthetic variables are numerical and underwent binning before tabulation. Distributional improvements within the variable range of single bins will be hidden in this kind of analysis.

The *pMSE ratio*, on the other hand, shows no correlation at all. The normalization of the *pMSE* seems to reduce the score dependence on specific variables. We note, that even the largest table used here (5 dimensions), is comparatively small with respect to the total variable number of the micro dataset. We expect stronger correlations, when computing the propensity score based metrics on the 6 selected variables alone.

### 4.3 Risk

In accordance with our correlation analysis for utility, we repeated the procedure by correlating SUDA and TCAP with our tabular data risk from Eq. (1). We show the correlation results in Fig. 3 for the lowest aggregation level (municipal-



**Fig. 3.** Correlation between microdata risk and tabular data risk for all possible tables. The x-axis describes the table type, by denoting the synthesized variable that is part of the table. We show the results for the lowest aggregation. A value of 1 implies maximal correlation between tabular and microdata metrics.

ity), since these are the most endangered tables when published. Our findings coincide strongly with the utility analysis. Again, we observe dependence on the variables spanning the table. The variable C drives utility as well as risk. Knowing this, the large confidence intervals found for tables without variable C indicate again that different synthesis configurations did not lead to noticeable changes in tables that include the binned numerical variables A and B. Although the SUDA risk definition does not correspond directly to our definition of risk at tabular level, we still see strong correlation for the tables including the categorical variable C. This may be explained by the fact, that GDs appear most prominently in the form of unique key combinations, since SUDA only focuses on the quantity MSU (Minimal sample unique), which is tied to uniqueness. Not surprisingly, the TCAP ratio correlation is close to 1 in Fig. 3 (right), since we have shown that our application of TCAP finds all cases of GD irrespective of the tabulation. Due to this measure operating on microdata level, no direct conclusion can be made about the way GDs are distributed among all the possible tabulation since TCAP ratio only gives us the ratio of true disclosures of all possible cross tabulations together. This becomes less useful, if the dataset has a huge number of variables, since it is unlikely that a statistical agency will publish all possible cross tabulations, hence overestimating the risk. This holds

as long as we have no knowledge about the potential subset of variables, which will be used in practise for building the tabular data.

## 5 Discussion

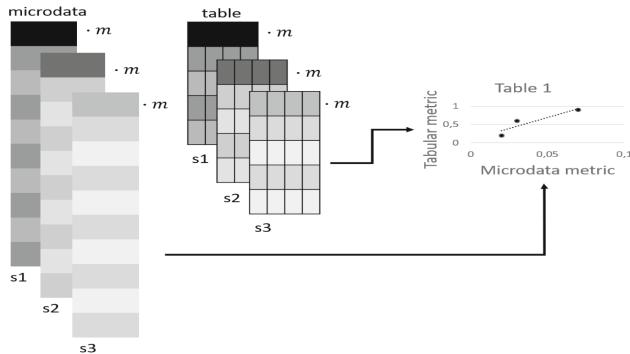
In this study, we focused on a selection of utility and risk metrics, both on microdata and tabular data level. We conducted a correlation analysis to determine whether microdata measures could be used to estimate the quality of tabular data, which is the main data product of statistical agencies. To keep the study as general as possible, we compared the microdata metrics to tabular data metrics computed on all possible cross tabulations given a subset of variables from the road traffic accidents data set.

Our findings suggest that pre-tabular measurements of *global data utility* may not effectively address tabular data quality. In our study, the ordinary *pMSE* is driven by the quality of a single variable while the *pMSE ratio* shows no correlation to tabular data utility at all. The disproportionate influence of single variable to this global utility score makes predicting the tabular data quality difficult. We add that binning numerical variables can affect accuracy, although being a natural step in the tabulation process. This limitation of global utility measures for tabular data products highlights the need for traditional, in-depth quality assessments at both tabular and microdata levels. This brings us back to the question of whether the *global utility* metrics themselves are sufficient to make a concrete statement about the overall utility of a synthetic dataset. On the other hand, we note that the utility measures on tabular data do not attempt to evaluate the utility of the entire dataset, but rather determine the utility for a specific case, which is more similar to the rubric of analysis-specific utility.

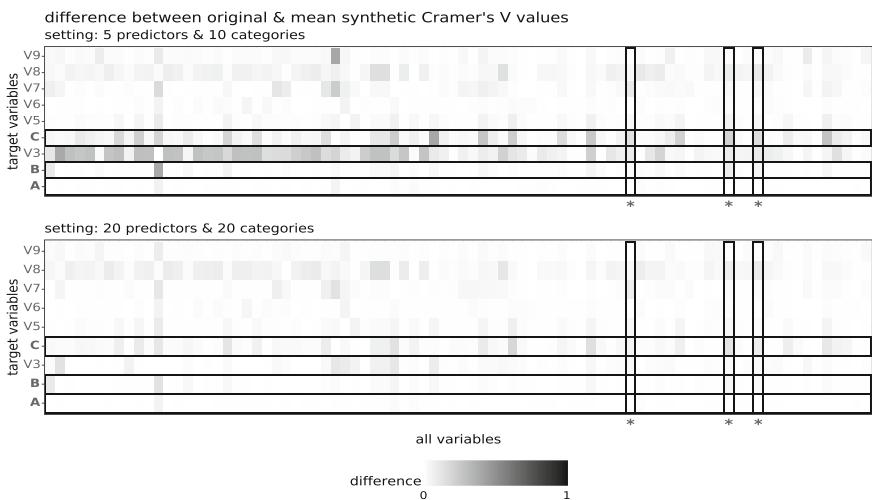
While the correlation analysis for risk metrics shows the same strong dependency on specific variables, microdata and tabular data risk measures are much more similar in nature. This can arguably be attributed to the historical development of data quality measurements preceding concerns for data privacy, which arose firstly in the context of statistical agencies, and hence shaped the scientific community early on. For datasets with few variables, the risk of GD can be confidently analyzed on microdata level in terms of the disclosure risk that the tables exhibit jointly.

## Appendix

### A Additional Figures



**Fig. 4.** Schema of the correlation analysis for a single table type. Every point in the plot corresponds to the metrics computed for one synthesizer configuration  $s_1$  to  $s_3$ . Before computing correlations, we averaged the metrics over the  $m$  synthetic datasets of each configuration.



**Fig. 5.** Visualization of the difference between the original and a synthetic association matrix based on Cramer's V, where the numerical variables have been discretized. The upper plot compares the original matrix with the worst synthesis configuration based on the *pmSE ratio* (i.e. with the highest mean *pmSE ratio* value) and the lower plot with the best configuration. On the y-axis are the target variables, where A, B, and C were used in the correlation analysis (Table 1). The asterisks on the x-axis mark the non synthesized key variables used in the correlation analysis. The darker the tiles, the greater the difference in the bivariate relation.

## B *pMSE Ratio*

[19] introduces the *pMSE ratio* to address the issue that the *pMSE* increases as the model complexity of the propensity model grows, even when the model is correctly specified. It is an extension of the *pMSE* and allows a clear interpretation for synthetic data. A *pMSE* of 0 indicates that the original and synthetic data are identical, which is highly unlikely and not the goal of synthetic data. Instead of aiming for an exact copy of the original, the focus is on achieving distributional similarity between observed and synthetic data. This goal, called *correct synthesis* (CS) by [19], assumes that synthetic and original data are samples from the same distribution and is reflected in a *pMSE ratio* of 1.

To obtain the *pMSE ratio* for  $m$  synthetic datasets, we compute the ratio between the empirical *pMSE* and the expected *pMSE*,

$$\text{pMSE ratio}_i = \frac{\widehat{pMSE}_i}{\mathbb{E}(pMSE)}, \quad \text{with } i = 1, \dots, m. \quad (6)$$

This expectation value, also called *null expectation*, is the expected *pMSE* of synthetic data generated under the assumption of CS. Hence the *pMSE ratio* is the synthetic data utility relative to a *correct synthesis*.

Since we are using CART, a non-parametric model, as the propensity model for our partial synthesis, we need to use pairwise resampling to approximate the null expectation, which is constructed as follows:

1. **Requirements** To obtain a sufficient number of pair combinations, an adequate number of synthetic datasets are required
2. **Pairing** Build all possible 2-way combinations of the  $m$  synthetic datasets (e.g. for  $m = 3$ , that would be  $(1, 2), (1, 3), (2, 3)$ )
3. **Stacking** For each pair combination, stack the two datasets and add a binary indicator variable (analogous to Sect. 3.1 calculating the *pMSE*)
4. **Prediction** Fit a classifier (here: CART) to the stacked dataset and predict the added indicator using the remaining variables of the stacked dataset as predictors
5. **Interim result** Compute the *pMSE*
6. **Iterating** Repeat step 3 to 5 for each pair combinations from step 2
7. **Output** Compute  $\mathbb{E}(pMSE)$  by averaging all calculated *pMSE*.

## References

1. Burgette, L.F., Reiter, J.P.: Multiple imputation for missing data via sequential regression trees. *Am. J. Epidemiol.* **172**(9), 1070–1076 (2010)
2. Drechsler, J., Haensch, A.C.: 30 years of synthetic data. *Stat. Sci.* **39**(2), 221–242 (2024)
3. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. *Comput. Stat. Data Anal.* **55**(12), 3232–3243 (2011)

4. Elliot, M.J., Manning, A.M., Ford, R.W.: A computational algorithm for handling the special uniques problem. *Internat. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**(05), 493–509 (2002)
5. Giomi, M., Boenisch, F., Wehmeyer, C., Tasnádi, B.: A unified framework for quantifying privacy risk in synthetic data (2022)
6. Goodfellow, I., et al.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, vol. 27 (2014)
7. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer, Heidelberg (2009). <https://doi.org/10.1007/978-0-387-84858-7>
8. Hornby, R., Hu, J.: Identification risks evaluation of partially synthetic data with the *IdentificationRiskCalculation* R package. arXiv preprint [arXiv:2006.01298](https://arxiv.org/abs/2006.01298) (2020)
9. Kursa, M.B., Rudnicki, W.R.: Feature selection with the Boruta package. *J. Stat. Softw.* **36**(11), 1–13 (2010)
10. Little, R.J.A.: Statistical analysis of masked data. *J. Off. Stat.* **9**(2), 407–426 (1993)
11. Micci-Barreca, D.: A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explor. Newsl.* **3**(1), 27–32 (2001)
12. Nowok, B., Raab, G.M., Dibben, C.: synthpop: bespoke creation of synthetic data in R. *J. Stat. Softw.* **74**, 1–26 (2016)
13. Raab, G.M., Nowok, B., Dibben, C.: Guidelines for producing useful synthetic data. arXiv preprint [arXiv:1712.04078](https://arxiv.org/abs/1712.04078) (2017)
14. Raab, G.M., Nowok, B., Dibben, C.: Assessing, visualizing and improving the utility of synthetic data. arXiv preprint [arXiv:2109.12717](https://arxiv.org/abs/2109.12717) (2021)
15. Reiter, J.P.: Using CART to generate partially synthetic public use microdata. *J. Off. Stat.* **21**(3), 441–462 (2005)
16. Reiter, J.P., Mitra, R.: Estimating risks of identification disclosure in partially synthetic data. *J. Priv. Confidentiality* **1**(1) (2009)
17. Rubin, D.B.: Statistical disclosure limitation. *J. Off. Stat.* **9**(2), 461–468 (1993)
18. Shlomo, N.: Statistical disclosure control methods for census frequency tables. *Int. Stat. Rev./Revue Internationale de Statistique* **75**(2), 199–217 (2007)
19. Snöke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. *J. R. Stat. Soc. Ser. A Stat. Soc.* **181**(3), 663–688 (2018)
20. Taub, J., Elliot, M.: The synthetic data challenge. In: *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality* (2019)
21. Therneau, T., Atkinson, B.: RPART: recursive partitioning and regression trees (2022)
22. Woo, M.J., Reiter, J.P., Oganian, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confidentiality* **1**(1) (2009)



# Synthetic Data Outliers: Navigating Identity Disclosure

Carolina Trindade<sup>1</sup>(✉) , Luís Antunes<sup>1,2</sup> , Tânia Carvalho<sup>1</sup> , and Nuno Moniz<sup>3</sup>

<sup>1</sup> Faculdade de Ciências da Universidade do Porto, Porto, Portugal

[carolina.trindade@fc.up.pt](mailto:carolina.trindade@fc.up.pt)

<sup>2</sup> TekPrivacy, Porto, Portugal

<sup>3</sup> Lucy Family Institute for Data & Society, Notre Dame, IN, USA

**Abstract.** Multiple synthetic data generation models have emerged, among which deep learning models have become the vanguard due to their ability to capture the underlying characteristics of the original data. However, the resemblance of the synthetic to the original data raises important questions on the protection of individuals' privacy. As synthetic data is perceived as a means to fully protect personal information, most current related work disregards the impact of re-identification risk. In particular, limited attention has been given to exploring outliers, despite their privacy relevance. In this work, we analyze the privacy of synthetic data w.r.t the outliers. Our main findings suggest that outliers re-identification via linkage attack is feasible and easily achieved. Furthermore, additional safeguards such as differential privacy can prevent re-identification, albeit at the expense of the data utility.

**Keywords:** Synthetic Data · Outliers · Deep Learning · Differential Privacy · Data Privacy · Data Utility

## 1 Introduction

Synthetic data [28] refers to data generated by methods designed to capture the distribution and statistical properties of the original dataset, producing data with similar characteristics. In the scope of data protection, synthetic data surged as an alternative to traditional methods such as generalization [13]. Legal frameworks like the General Data Protection Regulation (GDPR) highlight the need to adopt privacy measures to protect private information. As a result, synthetic data has become prevalent across different industries as a proxy for original data in many tasks, like software testing, where access to high-quality data is challenging due to privacy restrictions [11]. Numerous synthetic data generation approaches have been proposed in the literature [13], for instance, models based on deep learning, such as Generative Adversarial Networks [15] or Variational Autoencoders [19]. However, the greater the similarity between the synthetic data and the original data, the greater the risk of disclosure.

Although synthetic data generation is commonly seen as a means to fully protect the privacy of individuals, it has also raised numerous questions about

the memorization of deep learning models [23]. In particular, outliers are potentially more susceptible to attack due to their deviation from the general data. These extreme data points, often correspond to sensitive information. Existing literature focuses on assessing privacy risks through the application of diverse attack methodologies to quantify different privacy vulnerabilities, such as Membership Inference Attacks (MIA) [10, 14, 18, 31]. Stadler et al. [31] also evaluate privacy gain in synthetic data sets using outliers, but only focus on a small portion of rare cases, namely five data points. Although it is theoretically known that extreme data points are particularly more susceptible to intruders' attacks, empirical studies confirming this are scarce. To the best of our knowledge, no studies have yet focused exclusively on investigating the re-identification risk through linkage attacks of synthetic datasets concerning outliers.

In this paper, we conduct a linkage attack in a group of outliers to demonstrate the effectiveness of re-identification and how easily an attacker can link back supposedly protected personal information. We fine-tuned deep learning and differential privacy-based models to generate multiple synthetic data variants to evaluate their capacity to protect outliers.

Our main findings are summarized as follows.

- The effectiveness of protecting outliers during the synthesis process is ultimately model-dependent;
- Differential privacy-based models are more efficient in protecting the privacy of individuals than deep learning-based models, however with higher data utility degradation; and,
- Deep learning-based models provide high data quality with increased *epochs*, but at the expense of individuals' privacy.

Experiments were run using an Inter Core i7 Processor ( $4 \times 1.80$  GHz) and 8 GB RAM in a Ubuntu 18 partition with 100GB. Our experimental evaluation focuses on one original dataset due to computational limitations. However, it is noteworthy that **the re-identification of a single instance renders the dataset subject to the provisions outlined in the GDPR.**

The remainder of this paper is organized as follows. Section 2 provides an overview of data privacy and current related work. Section 3 presents the methodology and materials used in our experiments. The results are presented in Sect. 4 and discussed in Sect. 5. Section 6 concludes the paper.

## 2 Literature Review

In this section, we focus on the main strategies for de-identification, namely, traditional techniques and deep learning-based solutions, including privacy and utility measures. We also discuss our contributions to the related state-of-the-art.

### 2.1 Notions

The de-identification process was designed for secure data publication, aiming to modify data until an acceptable level of disclosure risk and data utility is

achieved [5]. In this process, quasi-identifiers (QIs) are selected to measure the risk of disclosure. Such attributes, when combined, can provide information that could lead to re-identification of individuals (e.g. gender and date of birth). Privacy-Preserving Techniques (PPTs) are then applied according to the level of privacy and utility. If the desired level is not met, the parameters of the PPTs are adjusted or another PPT is selected. Otherwise, the data is ready for release.

Multiple PPTs were proposed in the literature to reduce disclosure risk while maintaining data utility useful for tasks such as decision-making. Currently, synthetic data is at the forefront of data de-identification methods, given its reputation for providing potentially secure data while maintaining utility. Deep learning-based models have been proven to be more versatile and to provide better results in comparison to other approaches, such as interpolation methods [13]. Despite their popularity, the generation of synthetic data may overlook the protection of outliers, which makes them more vulnerable to exploitation [5].

Outliers are data points that deviate from the expected patterns or central tendencies of a dataset [16]. They can be classified as univariate, which are extreme values in the distribution of a particular attribute, or multivariate, which are unique combinations of values in the observations. The most common approach for outlier detection is based on the concept of standard deviation [21]. Let  $X = \{x_1, x_2, \dots, x_n\}$  be a dataset with  $n$  observations, in which  $\mu$  represents the mean and  $\sigma$  corresponds to the standard deviation of the dataset. Also, consider a threshold value  $k$ , that is used to determine the boundary beyond which data points are considered outliers. The outlier detection is then performed by

**Definition 1 (Outlier detection).** *Any data point  $x_i$  that lies outside the range  $[\mu - k\sigma, \mu + k\sigma]$  is an outlier.*

Researchers have developed a plethora of methods for identifying and handling outliers in datasets. Examples include Tukey's fences [33], Peirce's criterion [27] and graphical approaches such as box plots or normal probability plots. The choice of method generally depends on factors such as the distribution of the data and the desired level of sensitivity to outliers.

Concerning the privacy evaluation, we focus on identity disclosure, also known as re-identification, which occurs when an attacker associates a record of the released dataset as belonging to an individual w.r.t QI values. This type of disclosure can occur under a linkage threat [25]. Common approaches for record linkage are based on the similarity functions between pairs of records [12, 22, 24].

Many privacy measures have been proposed in the literature. The most known is  $k$ -anonymity [29], used to evaluate identity disclosure by indicating whether a dataset respects the desired level of  $k$ . Each individual cannot be distinguished from at least  $k-1$  other individuals. A more sophisticated measure is Differential Privacy (DP) [9] which provides a rigorous mathematical definition of privacy guarantees. With DP, the disclosure risk is assessed avoiding assumptions on previous background knowledge that an attacker may have or may learn about individuals. This is achieved by bounding the sensitivity of released data to the presence of any individual within the dataset. Generally, the Laplace distribution is used to add noise to the output of  $K$ , which is called Laplace Mechanism.

**Definition 2 (Differential Privacy [9]).** A randomized function  $K$  is  $\epsilon$ -differential privacy if for all data sets  $D_1$  and  $D_2$  differing on at most one element, and all  $S \subseteq \text{Range}(K)$ ,  $P_r[K(D_1) \in S] \leq \exp(\epsilon) \times P_r[K(D_2) \in S]$ .

A crucial aspect of de-identification is also the evaluation of utility since the transformations applied can potentially reduce the utility of the data. Thus, utility measures are used to quantify the similarity between the distributions of the original and the de-identified data [5]. Information loss measures aim to compare records between the original and the de-identified datasets as well as statistics computed from both to assess if the transformed data is still analytically valid (e.g. discernibility [2]). Measures such as mean, variance, and correlation are used to quantify the changes in statistics. Predictive performance metrics evaluate the ability to make predictions (e.g. Accuracy, Precision and Recall [1]).

## 2.2 Current Research Directions

Tao et al. [32] benchmark differentially private synthetic data generation algorithms and classify them as GAN-, Marginal- and Workload-based. Conclusions indicate that Marginal-based methods outperform the remaining approaches and GAN-based methods could not preserve the 1-dimensional statistics of data. Hotz et al. [17] discuss the major disadvantages of the use of synthetic data with differential privacy. The authors suggest that, the more accurate the synthesis process in representing the original data, which may contain outliers, the greater the risk of disclosure, thus synthetic data could be used by an attacker to make confident guesses on the identity or attributes of a real person.

El Emam et al. [10] propose a full risk model to evaluate re-identification and an attacker's ability to learn new information about an individual from matches between the original and synthetic records. Their findings indicate a low identity disclosure risk in both cases and concluded that the use of synthetic data reduces re-identification significantly. Houssiau et al. [18] propose TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. This threat modeling framework defines privacy attacks considering different attacker background knowledge and their learning objectives to evaluate synthetic data, additionally providing analysis reports. Giomi et al. [14] present Anonymeter, a statistical framework aiming to perform attack-based evaluations of different types of privacy risks in synthetic tabular datasets. In their experiments, synthetic data has the lowest vulnerability against linkability, suggesting that one-to-one relationships between real and synthetic data records are not preserved. Stadler et al. [31] propose a framework that allows the privacy evaluation of differentially private synthetic data and compares it to standard techniques. The authors conclude that synthetic data generated with generative models without any layer of protection do not protect outliers from linkage attacks and differentially private synthetic data protect against membership inference with a large loss of utility.

Previous studies have explored the privacy of synthetic data by conducting diverse attacks to measure different types of privacy risk [10, 14, 18, 31]. Among these, the work by Stadler et al. [31] is most closely aligned with ours; the authors

assess the privacy gain of synthetic data records, including outliers, with and without differential privacy regarding linkage attacks. While the authors perform a complex game to verify privacy gain, our approach involves a simpler linkage attack to demonstrate how easy is to perform a re-identification on outlier records. Furthermore, in contrast to their work, which focuses on only five extreme data points, our analysis includes a dataset with a substantially higher portion of outliers. Additionally, we generate multiple synthetic data variants using three deep learning models and three differential privacy models, each adjusted with an extensive range of hyperparameters, to analyze the influence of specific hyperparameters in synthetic data quality and privacy protection.

### 3 Experimental Evaluation

In this section, we provide our experimental methodology by describing the data and methods used. Given an original dataset, we generate multiple synthetic data variants using deep learning and differential privacy models. Then, we evaluate the generated synthetic variants in terms of privacy and utility. Concerning data utility, we use different metrics to analyze how synthetic data retains the original statistical properties. For privacy, we perform a linkage attack by comparing each data variant with the original data w.r.t a set of QIs in the outliers subset.

#### 3.1 Data

We use the Credit Risk dataset [8] as the original dataset for our study. This dataset has 22.910 records and presents a critical portion of outliers, essential for our experimental evaluation. Table 1 contains a summary of the attributes.

**Table 1.** Summary of the attributes of the Credit Risk dataset.

Attribute	Type	Description
<i>person_age</i>	Numerical	The person's age
<i>person_income</i>	Numerical	The person's annual income
<i>person_home_ownership</i>	Categorical	The person's home ownership
<i>person_emp_length</i>	Numerical	The person's employment length in years
<i>loan_intent</i>	Categorical	The loan intent
<i>loan_grade</i>	Categorical	The loan grade
<i>loan_amnt</i>	Numerical	The loan amount
<i>loan_int_rate</i>	Numerical	The loan interest rate
<i>loan_status</i>	Numerical	The loan status
<i>loan_percent_income</i>	Numerical	The loan percent income
<i>cb_person_default_on_file</i>	Categorical	The historical default
<i>cb_person_cred_hist_length</i>	Numerical	The credit history length

This dataset has approximately 12% missing records. To preserve the integrity of future analysis, all instances of missing values have been removed.

Considering the data properties, we select *person\_age*, *person\_income*, *person\_home\_ownership* and *loan\_intent* as QIs. We hypothesize that their accessibility could facilitate easier re-identification by an attacker.

### 3.2 Methods

**Synthetic Data Generation.** We use models available in the SDV [6, 26] and DPART [20, 30] tools to obtain respectively both deep learning- and differential privacy-based synthetic dataset variants. Regarding SDV, we use the TVAE, CTGAN and CopulaGAN models. From DPART, we use the Independent, PrivBayes and DPsynthpop models. The combination of the hyperparameters resulted in 27 synthetic variants for each SDV model and 7 synthetic variants for each DPART model, all with the same size as the original dataset. Thus, in total, we generated 102 synthetic dataset variants. Table 2 contains a summary of the hyperparameters tested in this experiment.

**Table 2.** Synthetic data generation tools with models and hyperparameters used.

Tool	Models	Parameters
SDV	TVAE	$epochs \in \{150, 300, 500\}$
	CTGAN	$batch\_size \in \{20, 50, 100\}$
	CopulaGAN	$embedding\_dim \in \{12, 64, 128\}$
DPART	Independent	$epsilon \in \{0.01, 0.1, 0.2, 0.5, 1.0, 5.0, 10.0\}$
	PrivBayes	
	DPSynthpop	

**Data Utility.** In terms of data utility, we use the SDMetrics [7] tool which is integrated with SDV and provides various metrics to evaluate different aspects of synthetic data. We need both original and synthetic dataset variants for such an evaluation. Table 3 contains a summary of the used metrics.

**Table 3.** Utility metrics used from SDMetrics [7].

Metric	Description
BoundaryAdherence	Verifies whether a synthetic column respects the minimum and maximum values of the real column
CategoryCoverage	Verifies whether a synthetic column covers all the possible categories that are present in a real column
RangeCoverage	Verifies whether a synthetic column covers the full range of values that are present in a real column
StatisticSimilarity	Verifies the similarity between a real and a synthetic column by comparing a summary statistic (median)

CategoryCoverage and RangeCoverage are equivalent metrics for categorical and numerical attributes, so we combined them as AttributeCoverage.

**Outliers Attack.** As we conduct linkage attacks on outliers, we use the  $z-score$  method to select the outliers subset. This method is described as follows.

$$Z = \frac{x_i - \mu}{\sigma} \quad (1)$$

$Z - score$  is applied to numerical QIs, where  $x_i$  represents the  $i^{th}$  the attribute value,  $\mu$  is the sample mean and  $\sigma$  is the standard deviation. We set a threshold  $k = 3$ ; thus,  $x_i$  is an outlier if the  $z - score$  is greater than 3 or less than -3. As we have multiple numerical QIs, we perform the multivariate outliers selection.

To perform a linkage attack on outliers we use Record Linkage [4] toolkit. The parameters in Record Linkage depend on QIs type. For the numerical QIs we used the Gauss method with *offset* and *scale* of 5 for the *person\_age* attribute and 1.000 for the *person\_income* attribute. These parameters allow us to have a wider range of values to consider as potential matches. For the categorical attributes, we used the Levenshtein method without defining a range, because the package looks for similar words and synthetic data is generated according to distribution, so we are only interested in exact matches here. The output of the tool contains the index of the record in the original dataset and the index of the record in the synthetic dataset variant that are considered possible matches, as well as a score for each QI. These scores range from 0 to 1, where 1, indicates a possible match and 0, indicates not a possible match.

We filter the score results by applying a threshold of 0.5 for numeric Quasi-Identifiers and 1 for categorical attributes. In scenarios where potential matches exist across all QIs (worst-case scenario), we aggregate these results to determine the number of instances where each record in the original dataset corresponds to only one possible match in the synthetic variant.

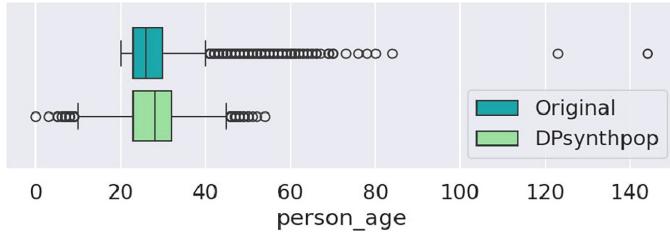
## 4 Results

In this section, we present the results obtained from the development focusing on data utility and linkage attack.

### 4.1 Data Utility

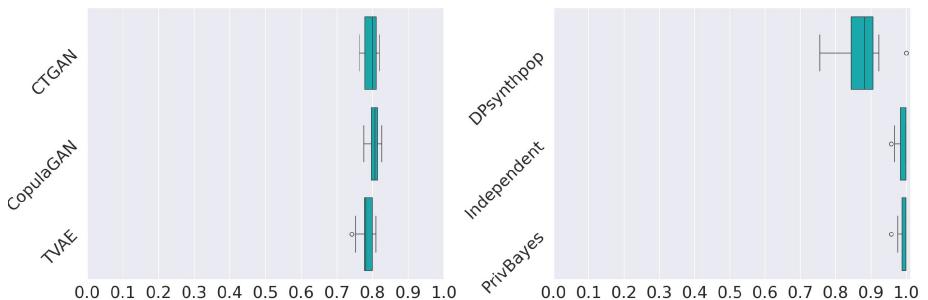
We report the utility results for each synthetic data variant, as outlined in Table 3. The metrics, calculated for each attribute, are averaged across each variant.

Regarding BoundaryAdherence, all models scored a 1, indicating that data points were generated within established bounds. However, the DPsynthpop model performed poorly, with a median BoundaryAdherence of approximately 0.45, as it failed to adhere to the minimum and maximum values of the original dataset's attributes. Figure 1 illustrates this issue by comparing the distributions of the original dataset and a specific DPsynthpop variant.



**Fig. 1.** Example of the distribution of the attribute *person\_age* in the original dataset compared to a DPsynthpop variant.

Figure 2 presents the AttributeCoverage for each synthetic data variant. Results show that differential privacy-based models outperform deep learning-based models, likely because the latter focuses on replicating the most frequent values from the original dataset. In contrast, differential privacy-based models, generate records to cover all possible values of each attribute, resulting in an overrepresentation of outliers compared to the original dataset (617 outliers). Table 4 provides the average number of outliers generated by each model.

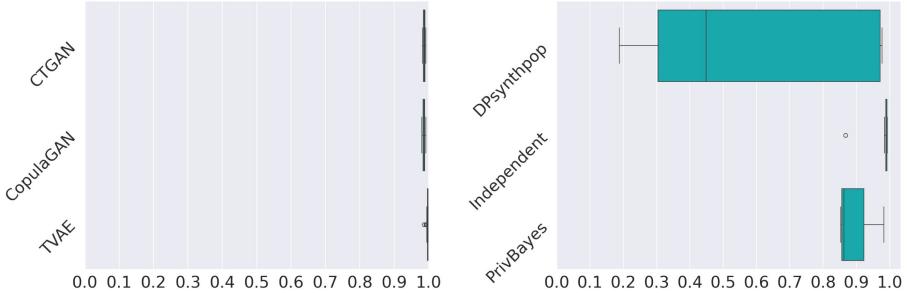


**Fig. 2.** AttributeCoverage of the synthetic dataset variants generated with deep learning-based models (left) and differential privacy-based models (right).

**Table 4.** Average number of outliers for each synthetic data generation model.

Tool	Models	Outliers
SDV	TVAE	720
	CTGAN	840
	CopulaGAN	794
DPART	Independent	1449
	PrivBayes	1483
	DPsynthpop	114

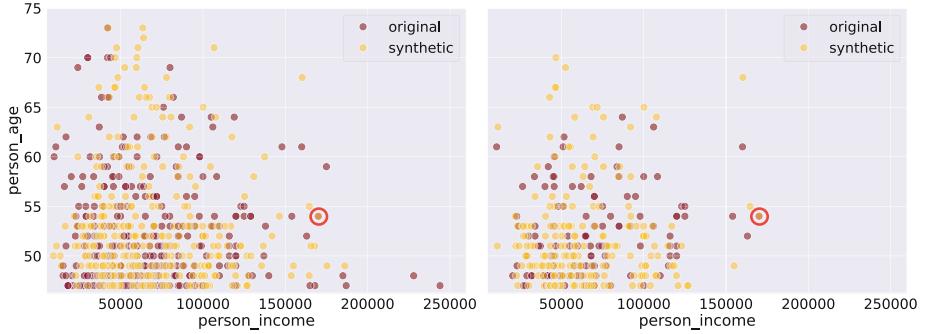
Note that DPsynthpop is an exception. As previously mentioned, this model fails to respect attribute value boundaries (Figs. 1 and 2), resulting in incomplete coverage of all possible attribute values. This model also demonstrates weak performance in StatisticSimilarity, with a median lower than 0.5 as shown in Fig. 3. This indicates a poor replication of the original data distributions. All the remaining models show statistical similarity with the original data.



**Fig. 3.** StatisticSimilarity of the synthetic dataset variants generated with deep learning-based models (left) and differential privacy-based models (right).

## 4.2 Linkage Attack

We now demonstrate the feasibility of re-identification using synthetic data through linkage attacks, particularly by leveraging outliers. We select a certain synthetic data variant, for demonstrative purposes. Specifically, we use a TVAE variant ( $epochs = 150$ ,  $batch\_size = 20$ ,  $embedding\_dim = 12$ ). Our objective is to show the effectiveness of this attack by expanding the supposed background knowledge of an attacker. Figure 4 presents on the left the potential matches based on the numerical QIs by comparing this variant against the original (4279 possible matches). In comparison, the right image shows the possible matches for all QIs (490 possible matches). For visual clarity, we only illustrate the potential matches for numerical QIs for individuals aged over 45.



**Fig. 4.** Distribution of records that are possible matches regarding the *person\_age* and *person\_income* attributes (left) and also including *person.home\_ownership* and *loan\_intent* attributes (right) with one possible match highlighted.

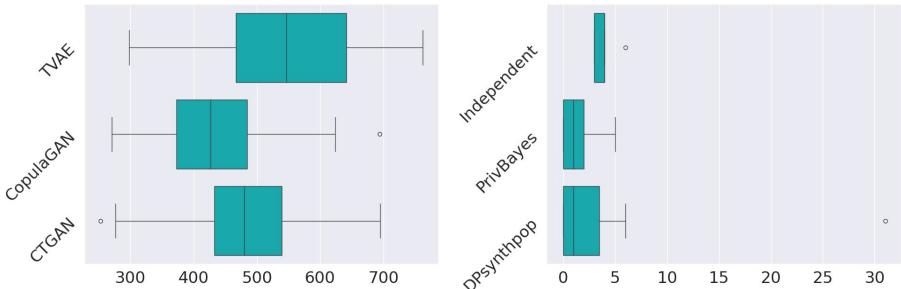
**Table 5.** Example of possible matched records.

Attribute	Original	Variant
<i>person_age</i>	<b>54</b>	<b>54</b>
<i>person_income</i>	<b>170000</b>	<b>170262</b>
<i>person.home_ownership</i>	<b>MORTGAGE</b>	<b>MORTGAGE</b>
person.emp.length	12.0	33.0
<i>loan_intent</i>	<b>PERSONAL</b>	<b>PERSONAL</b>
<i>loan_grade</i>	A	A
<i>loan_amnt</i>	11000	8053
<i>loan_int_rate</i>	6.62	7.27
<i>loan_status</i>	0	0
<i>loan_percent_income</i>	0.06	0.06
<i>cb_person_default_on_file</i>	N	N
<i>cb_person_cred_hist_length</i>	20	21

In both cases, we observe multiple potential matches where the original and synthetic variant data points overlap. However, as more QIs are incorporated, the number of these potential matches decreases. This is due to the refined specificity of background knowledge, which narrows the pool of individuals who match these characteristics. Focusing on the highlighted example, Table 5 verifies these overlapping points. This particular match was obtained by filtering results for *person\_age* = 54 and  $160000 \leq person\_income \leq 180000$ .

Despite some attribute values being distant, all QIs matched. Given that these cases are outliers, and the models typically do not generate outliers considering non-outlier records, it is very unlikely that another data point could be a possible match. Therefore, we can infer that this variant record is a synthetic version of that original record, illustrating the feasibility of re-identifying

synthetic data. Figure 5 shows the number of possible matches for each model’s synthetic dataset variants under the worst-case scenario – using all QIs.



**Fig. 5.** Possible matches for the synthetic dataset variants generated with deep learning-based models (left) and differential privacy-based models (right).

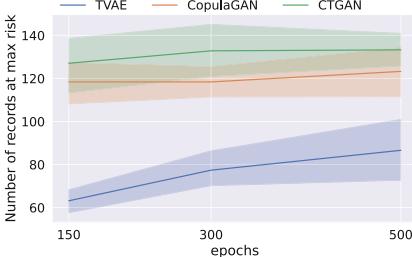
Deep learning-based models yield a higher number of possible matches compared to differential privacy-based models, indicating a heightened risk of re-identification. Although some of these values may be false positives due to the adjustments of scale and offset to widen our search of possible matches, the potential for privacy breaches remains substantial. Nevertheless, we considered the scores obtained on the attributes of the possible matches to assess success rate.

## 5 Discussion

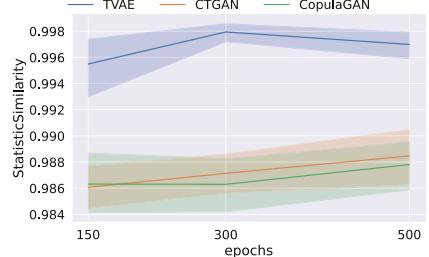
Our experiments demonstrate that the protection of outliers ultimately depends on the synthetic data generation model. Deep learning-based models tend to adhere to the general distribution of the original dataset, focusing on generating more frequent values. On the other hand, differential privacy-based models deliberately generate records that cover all the values of each attribute, which results in a disproportionately high number of outliers (Table 4).

We also observe that adding an extra layer of protection, namely differential privacy, compromises data quality. Notably, the DPSynthpop model generated inferior quality synthetic data because it failed to accurately learn the distribution of the original dataset (Figs. 2 and 3). Despite this, the extra layer of protection resulted in fewer instances where only one possible match existed between each record of the original dataset and the synthetic variants. However, this enhancement in privacy protection might not outweigh the substantial loss in data quality and utility. To further understand this relationship, we present in Fig. 6 and Fig. 7 the number of cases that are at maximum risk (1 possible match) for each synthetic data variant concerning the *epochs* parameter, and how similar each synthetic data variant is concerning this parameter. We do not

include differential privacy-based models in this analysis as the parameters are not comparable and they perform very poorly in terms of data utility. This leads us to the question: *how does the number of epochs influence utility and privacy?*



**Fig. 6.** Original data records with one possible match in each synthetic data variant generated from deep learning-based models w.r.t *epochs* parameter.



**Fig. 7.** StatisticSimilarity of each synthetic data variant generated from deep learning-based models w.r.t *epochs* parameter.

Results indicate that as the parameter *epochs*, which represents the number of iterations the models use to optimize their parameters, increases, the number of unique matches between the original dataset and the synthetic variants also rises. For variants generated using CTGAN and CopulaGAN, increased *epochs* enhance the similarity between the original and synthetic data, whereas for those generated using TVAE, the similarity decreases. In general, a higher *epoch* value allows models to better capture the characteristics of data, resulting in synthetic data that more closely resemble the original. Consequently, this similarity also elevates the re-identification risk. It is therefore important to tune the hyperparameters of synthetic data generation models according to the data's characteristics to achieve an optimal balance between data quality and privacy.

For future work, we plan to include records that are not outliers, a larger number of datasets, different QI sets, and other types of attacks to corroborate these results. We considered the risk of MIA using the DOMIAS [3] tool which, to the best of our knowledge, is the only tool performing MIA on synthetic data. DOMIAS is a density-based MIA model that allows MIA against synthetic data by targeting local overfitting of the generative model. Unfortunately, a big limitation of this tool is that it only operates on numerical data. Furthermore, we aim to investigate the impact of outlier treatment strategies at different stages (before, during, and after synthesis) on both utility and privacy. Also, evaluate the necessity of excluding outliers from the synthetic data generation process.

Although synthetic data does not fully protect the privacy of individuals, it serves as an effective proxy for the original data in many tasks, highlighting its value. We stress the importance of developing and updating synthetic generation approaches for a secure and robust data environment that balances utility with privacy considerations specially focused on extreme and rare data points.

## 6 Conclusion

In this paper, we conducted an analysis focused on the effectiveness of the re-identification associated with synthetic data generation models, specifically examining their efficacy in protecting extreme data points, i.e. outliers.

Our results showed that outlier protection is model-dependent. The deep learning-based models tested focused on generating more frequent values, while the differential privacy-based models generally generated a higher number of outliers. However, the differential privacy-based models also resulted in poorer data quality. Most importantly, we conducted a linkage attack to demonstrate how outliers can be exploited to re-identify personal information, highlighting the vulnerability associated with synthetic data.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Allen, K., Berry, M.M., Luehrs Jr., F.U., Perry, J.W.: Machine literature searching viii. operational criteria for designing information retrieval systems. *Am. Doc.* (pre-1986) **6**(2), 93 (1955)
2. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: 21st International Conference on Data Engineering (ICDE 2005), pp. 217–228. IEEE (2005)
3. van Breugel, B., Sun, H., Qian, Z., van der Schaar, M.: Membership inference attacks against synthetic data through overfitting detection. arXiv preprint [arXiv:2302.12580](https://arxiv.org/abs/2302.12580) (2023)
4. de Bruin, J.: Recordlinkage. Online (2016). <https://pypi.org/project/recordlinkage/>. Accessed March 2023
5. Carvalho, T., Moniz, N., Faria, P., Antunes, L.: Survey on privacy-preserving techniques for microdata publication. *ACM Comput. Surv.* (2023)
6. DataCebo, I.: SDV. Online (2018). <https://github.com/sdv-dev/SDV>. Accessed January 2023
7. DataCebo, I.: Sdmetrics. Online (2020). <https://github.com/sdv-dev/SDMetrics>. Accessed January 2023
8. Dua, D., Graff, C.: Credit risk dataset. Online (2020). <https://www.kaggle.com/datasets/laotse/credit-risk-dataset>. Accessed April 2023
9. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). [https://doi.org/10.1007/11787006\\_1](https://doi.org/10.1007/11787006_1)
10. El Emam, K., Mosquera, L., Bass, J.: Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *J. Med. Internet Res.* **22**(11), e23139 (2020)
11. El Emam, K., Mosquera, L., Hopetroff, R.: Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. O'Reilly Media (2020)
12. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. *J. Am. Stat. Assoc.* **64**(328), 1183–1210 (1969)

13. Figueira, A., Vaz, B.: Survey on synthetic data generation, evaluation methods and GANs. *Mathematics* **10**(15), 2733 (2022)
14. Giomi, M., Boenisch, F., Wehmeyer, C., Tasnádi, B.: A unified framework for quantifying privacy risk in synthetic data. arXiv preprint [arXiv:2211.10459](https://arxiv.org/abs/2211.10459) (2022)
15. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
16. Grubbs, F.E.: Procedures for detecting outlying observations in samples. *Technometrics* **11**(1), 1–21 (1969)
17. Hotz, V.J., et al.: Balancing data privacy and usability in the federal statistical system. *Proc. Nat. Acad. Sci.* **119**(31), e2104906119 (2022)
18. Houssiau, F., et al.: TAPAS: a toolbox for adversarial privacy auditing of synthetic data. arXiv preprint [arXiv:2211.06550](https://arxiv.org/abs/2211.06550) (2022)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114) (2013)
20. Mahiou, S., Xu, K., Ganev, G.: Dpart: differentially private autoregressive tabular, a general framework for synthetic data generation. arXiv preprint [arXiv:2207.05810](https://arxiv.org/abs/2207.05810) (2022)
21. Mateo-Sanz, J.M., Sebé, F., Domingo-Ferrer, J.: Outlier protection in continuous microdata masking. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 201–215. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-25955-8\\_16](https://doi.org/10.1007/978-3-540-25955-8_16)
22. Muralidhar, K., Domingo-Ferrer, J.: Rank-based record linkage for re-identification risk assessment. In: Domingo-Ferrer, J., Pejić-Bach, M. (eds.) PSD 2016. LNCS, vol. 9867, pp. 225–236. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-45381-1\\_17](https://doi.org/10.1007/978-3-319-45381-1_17)
23. Nikolenko, S.I.: Synthetic data for deep learning. arXiv preprint [arXiv:1909.11512](https://arxiv.org/abs/1909.11512) (2019)
24. Pagliuca, D., Seri, G.: Some results of individual ranking method on the system of enterprise accounts annual survey. Esprit SDC Proj. Deliverable MI-3 D **2**, 1999 (1999)
25. Party, A.D.P.W.: Opinion 05/2014 on anonymisation techniques. European Commission (2014)
26. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410. IEEE (2016)
27. Peirce, B.: Criterion for the rejection of doubtful observations. *Astron. J.* **2**(45), 161–163 (1852)
28. Rubin, D.B.: Statistical disclosure limitation. *J. Off. Stat.* **9**(2), 461–468 (1993)
29. Samarati, P.: Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001)
30. Mahiou, S., Xu, K., G.G.: Dpart. Online (2022). <https://github.com/hazy/dpart>. Accessed May 2023
31. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data–anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 2022), pp. 1451–1468 (2022)
32. Tao, Y., McKenna, R., Hay, M., Machanavajjhala, A., Miklau, G.: Benchmarking differentially private synthetic data generation algorithms. arXiv preprint [arXiv:2112.09238](https://arxiv.org/abs/2112.09238) (2021)
33. Tukey, J.W., et al.: Exploratory Data Analysis, vol. 2. Reading (1977)



# Privacy Risk from Synthetic Data: Practical Proposals

Gillian M. Raab<sup>(✉)</sup>

Scottish Centre for Administrative Data Research, University of Edinburgh,  
Edinburgh, Scotland, UK  
[gillian.raab@ed.ac.uk](mailto:gillian.raab@ed.ac.uk)

**Abstract.** This paper proposes and compares measures of identity and attribute disclosure risk for synthetic data. Data custodians can use the methods proposed here to inform the decision as to whether to release synthetic versions of confidential data. Different measures are evaluated on two data sets. Insight into the measures is obtained by examining the details of the records identified as posing a disclosure risk. This leads to methods to identify, and possibly exclude, apparently risky records where the identification or attribution would be expected by someone with background knowledge of the data. The methods described are available as part of the **synthpop** package for **R**.

## 1 Introduction

It is now thirty years since the first proposals were made to use synthetic data (SD) as a privacy enhancing technology (PET) [17, 27]. These were concerned with using data synthesis to make confidential versions of microdata available to analysts, while reducing the risk of breaching the privacy of individuals contributing records to the original “Ground Truth” (GT) data used to create the SD. In recent years data synthesis has been used in many other contexts, such as anonymizing images that might identify people, or concealing geographic locations from mobile phone transaction data; see [34] for examples. This paper is restricted to the use of SD to create confidential microdata, often by national statistics agencies (NSAs), as described in a recent UNECE report [32]. Two recent reviews of the field [5, 26] have highlighted the need for practical privacy metrics to evaluate the disclosure risk (DR) from the release of SD. This is in contrast to the variety of measures of utility available to compare SD with GT [1, 14, 15, 23, 29, 33, 35].

For the past ten years we<sup>1</sup> have been developing the **synthpop** package for creating synthetic microdata. We have mainly implemented methods proposed by others to make them available for use on real problems. The package provides a variety of methods for creating SD and includes tools to allow the person creating the SD to check that analyses based on the SD will mirror what would

<sup>1</sup> The team was led by Beata Nowok and also included Chris Dibben and the author.

be found from the GT. These can be broadly defined as measures of utility. They include measures that compare results from particular analyses (specific utility) and overall measures that compare the whole distributions (general utility) [29].

It has been argued that SD, based only on models, has no records that can be associated with identifiable individuals and is thus not personal data as defined by GDPR<sup>2</sup>. Most people would now dispute this because it may be possible to infer, from the SD alone, the characteristics of individuals in the GT used in its creation. This view has been expressed recently by the European Data Protection Supervisor<sup>3</sup>. Ideally the SD should reproduce the relationships between variables that are features of the population that the GT data represent, but obfuscate the random variation about this model; i.e. reproduce the signal but not the noise. There is no agreed definition of what methods can be used to create SD and some models may give too good a fit to the GT so that the noise can be identified in the SD. Data synthesis carried out via iterative techniques such as GANs [9] may be especially prone to this problem, as can popular tree-based models [4], such as the default model in **synthpop**, if the trees are allowed to grow too large.

This paper is concerned with developing tools that data custodians can use to evaluate the DR of SD. In Sect. 2 we discuss what information needs to be assumed to define the risks. Then in Sect. 3 we provide details of the scenario we are proposing to define these risks. Measures of identity and attribute DR are defined and recommendations made as to which should be used. The measure “replicated uniques”, *repU*, is proposed for identity disclosure and “Disclosive in Synthetic Correct in Original” (*DiSCO*) for attribute disclosure. Results from one small example are discussed, and include methods to identify cases where an apparent disclosure might not really be an unexpected risk. Section 4 then presents methods of excluding such records using a larger data set. Section 5 discusses the differences and commonalities between disclosure risk and utility of synthetic data and discusses other work on quantifying DR that has developed a different approach to this [8]. The paper concludes with practical recommendations and suggestions for future steps.

## 2 Practical Considerations

Synthetic data has the potential to widen access to administrative data to those outside the organisation where it was created. But the lack of any way of assessing DR is a major barrier to allowing custodians to sanction its release. Those with the responsibility of securing public data cannot be expected to be reassured as to the safety of releasing SD without a practical demonstration of how the privacy of data subjects is being protected. When NSAs do release data to the

---

<sup>2</sup> See [https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu\\_en](https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en), accessed 19/5/2024 In the UK this is now incorporated within the Data Protection Act 2018.

<sup>3</sup> See [https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data\\_en](https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en) Accessed 20/5/2024.

public as either tables or microdata (e.g. the Sample of Anonymised Records provided from the UK Censuses<sup>4</sup>) it will have gone through extensive testing and been subject to statistical disclosure control (SDC) to ensure its safety. Section 5 outlines how these procedures could be adapted to synthetic data.

During 2023 the author was part-funded by Research Data Scotland<sup>5</sup> to develop practical measures that could be used to evaluate the DRs of synthetic data that were being considered for release to researchers. The latest version of the **synthpop** package for **R** now includes routines to measure disclosure risk<sup>6</sup>. Details of the disclosure routines are described in Raab et al. (2024) [24]. While evaluating the utility of synthetic data is relatively straightforward, the assessment of DR and its integration into a protocol for producing synthetic data is more difficult because it depends on other factors, as discussed in the anonymization decision-making framework [7]. These include:

- (a) the context of the data release: when, how, to whom, under what regulation
- (b) whether the person with access to the SD believes it to be real
- (c) what other sources of data are available for records in the GT
- (d) what the person with access to the SD knows about the GT in general and specific ways
- (e) what information is released about how the SD has been created.

Item (a) is a decision that must be taken by, or on behalf of, the data holder and DR must be interpreted in this context. Item (b) is a major concern to data holders. Someone with SD might inadvertently allow access to others who might believe it to be real leading to a loss of reputation for the data custodian. Another situation where this could arise would be if a laptop with SD were lost or stolen<sup>7</sup>. The disclosure measures presented here assume that the SD are approached it as if it were real. While this may seem a worst-case scenario it may be a useful one, since it leads to DR measures that compare risks from the SD to those from the GT, as we illustrate in the examples below. Items (c) and (d) are clearly important. In the next section we explain how they affect the disclosure measures. Item (e) is specific to the means of creating SD. It may include general information about the type of method used in the creation of the SD, as well as specific methods. The latter is much more likely to lead to a privacy violation, especially if details of the model are released to a sophisticated intruder [28]. This

---

<sup>4</sup> see <https://www.ons.gov.uk/census/aboutcensus/censusproducts/microdataspaces>.

<sup>5</sup> An organisation with the mission “...to work with researchers, analysts and policy-makers to unlock the potential of public sector data for the benefit of public good”; see <https://www.researchdata.scot/>.

<sup>6</sup> This is version 1.8.1 that can be installed from Github at <https://github.com/gillian-raab/synthpop>.

<sup>7</sup> The author’s involvement with SD nearly came to an end when her laptop, with SD created for a training course, was stolen. Fortunately, she was able to reassure the security staff from the data holders that the SD was fully encrypted as well as being clearly labelled as “Fake data”.

is an area that would benefit from further investigation. Meanwhile, it would be prudent of data holders not to release details of their synthesis models.

### 3 Scenario and Definitions

#### 3.1 Setting the Scene

These disclosure measures are intended to assess what a person who only has access to the SD can infer about known individuals who are present in the GT. We use the term “intruder” for such a person, though no malicious intent is implied. The intruder is assumed to have information for one or more individuals about the value of certain key variables (quasi-identifiers) that are present in the GT. The identification of quasi-identifiers is an important aspect of DR assessment and a data holder may have to update decisions about this if new data sources are accessed that allow people to discover information about known individuals, e.g. scraping information from the web. The intruder first attempts to see if the individual with these quasi-identifiers is present (identity disclosure), and then to determine the value of other, potentially sensitive, items in the data file that we refer to as targets (attribute disclosure). We are assuming a worst-case scenario where the intruder believes they are querying the original data.<sup>8</sup> Disclosure measures from the GT are each compared to similar measures for someone with access only to the SD. The difference between these two measures (e.g. *Dorig - DISCO*, for the preferred measures described below) is a measure of the disclosure protection afforded by the synthesis.

Here we introduce the measures with an example. Formal definitions with notation and formulae are in Appendix 1. The first step in evaluating DR, as described here, is to identify a set of keys that might be expected to be known to an intruder. These keys are then combined to form a combination of quasi-identifiers that we designate as  $q$ . For example, if we have hospital records we might define age, sex date and hospital as keys and this would give a  $q$  with levels such as “78 | M | 1/1/2024 | WG” for a 78 year old man admitted to hospital WG on 1/1/2024.

#### 3.2 Identity Disclosure Measures

The concept of k-anonymity is central to identity disclosure for microdata. First proposed in 1998 [3], it is discussed fully in [7]. A table is k-anonymous with respect to a set of keys if a record cannot be distinguished from at least  $k - 1$  others. Based on this idea, the percentage of records for which the keys identify just one individual (i.e. that breach 2-anonymity or, more simply, are unique in the data) give identity disclosure measures. Tables of  $q$  values are produced from the GT and the SD. *UiO* and *UiS* (Unique in Original and Unique in Synthetic)

---

<sup>8</sup> This may not be too unrealistic if the data are made available inadvertently, or if the intruder thinks that efforts to label the SD as e.g. “Fake Data” are thought to be just a cover up.

are the percentages of records in the GT and in the SD with unique values of  $q$ . An intruder checking out a record for their known set of keys will look for it in the SD. Some records in  $UiO$  will not be in the SD and  $UiOoS$  (Unique in Original in Synthetic) gives the percentage that would be found. These records are then checked for uniqueness in the SD, giving  $repU$  (replicated uniques) as the percentage of unique GT records that are also unique in the SD.

The percentage  $repU$  has been used as a disclosure measure to evaluate SD by [12] and by [25]<sup>9</sup>. Replicated uniques are used in **synthpop** as part of the statistical disclosure control function, **sdc**, that includes the option of reducing DR by removing them from the SD. Nowok et al. [20] have evaluated this and give an example where this process has very little effect on utility.

### 3.3 Attribute Disclosure Measures

Attribute disclosure uses the same composite identifier  $q$ , from the keys, and uses it to identify the level of a target  $t$ . The attribute disclosure measure,  $DCAP$ , was proposed by Elliot [6] and has been used in other evaluations of SD [21, 31].  $DCAP$  is calculated as the average percentage of records with  $t$  correctly predicted from  $q$  in the SD. Its use was proposed for the Synthetic Data Challenge<sup>10</sup> where teams created SD sets based on the 2001 Census of Scotland. In response to criticisms by team members that this measure might be a measure of utility rather than privacy, the measure was adapted to  $TCAP$  by restricting disclosure to cases when  $q$  will predict a unique value of  $t$  in the SD - i.e. records with an l-diversity of 1 [18].  $TCAP$  was used in the final report of the data challenge [31] and in other evaluations of SD [2, 16]. It is close to the  $DiSCO$  measure that we propose below. Appendix 1 gives formal definitions of these measures.

We approach DR from the point of view of an intruder with access to the SD and to keys forming  $q$  for one or more individuals in the original data.

Modelling what an intruder might do, we calculate the following measures, each of which is a proportion of the original records:

- For a given  $q$  in GT search for the  $q$  value in the SD. The proportion found becomes  $iS$  (in Synthetic)
- Check if all records with the same  $q$  in SD have the same level of the target  $t$ . The proportion passing this further test becomes  $DiS$  (Disclosive in Synthetic).
- Then check if these apparent disclosures corresponds to the value of  $t$  in the GT. The proportion of GT records for which this is true becomes  $DiSCO$  Disclosive in Synthetic Correct in Original.

---

<sup>9</sup> Jackson et al. in [12] argue that the denominator for  $repU$  should be the number of records in SD, rather than those in GT. This is inappropriate because our scenario is to consider the risk to the GT data.

<sup>10</sup> This took place at the Newton Institute programme on Data Linkage and Anonymisation, 2016 see <https://www.newton.ac.uk/event/dla/>.

Note that records contributing to  $DiSCO$  may not be disclosive in the original data; this information would not be available to the intruder. A further measure  $DiSDiO$  (Disclosive in Synthetic Disclosive in Original) restricts the score to those also disclosive in the GT. These measures are defined formally in Appendix 1.

### 3.4 A Simple Example

To illustrate our proposed measures we use a data set about quality of life in Poland (SD2011), available as part of the **synthpop** package. The target is a score for depression (`depress`) and  $q$  is created from variables “sex”, “age”, “region” and “placesize”. The depression score has 23 levels and the  $q$  formed from these keys has 3,459 levels in the GT but this increases to 4,475 when GT and SD are combined. Almost 50% of the SD records have unique levels of  $q$  and this reduces to 15% for  $repU$ .

Five synthetic data sets are created and Table 1 illustrates the attribute disclosure measures for the target “`depress`” for each of these. From the GT data 53.3% ( $Dorig$ ) of records could identify the level of “`depress`” with certainty. Turning to the synthetic data we start by seeking the  $q$  values from the GT in the SD and 64% ( $iS$ ) are found. The requirement for a record to be disclosive in the SD reduces this to 33% ( $DiS$ ) and again requiring the disclosure to be correct reduces this to around 9% ( $DiSCO$ ). This rate reduces further to around 6%, by restricting to records that were also disclosive in the GT ( $DiSDiO$ ).

The final column in Table 1 gives the values for  $DCAP$ , the average percentage of records correctly predicted, but not restricted to those predicted with certainty. As expected the rates are higher than for  $DiSCO$ . The corresponding rate from the GT ( $CAPd$ ) was 74%.

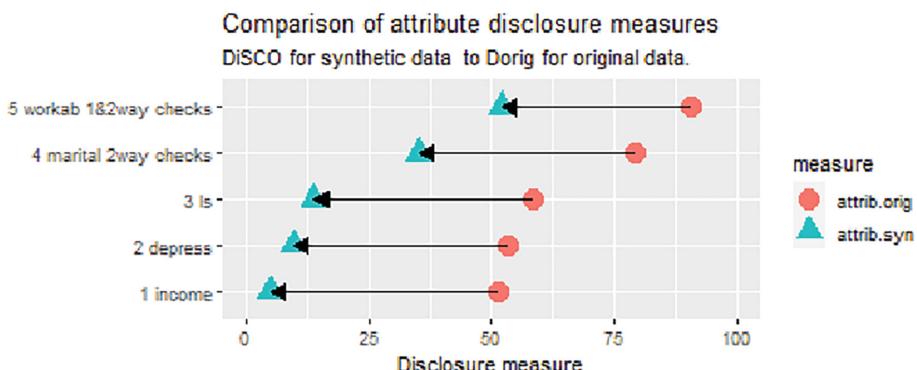
**Table 1.** Attribute disclosure measures for 5 syntheses of the SD2011 data with target “`depress`” identified from keys “sex”, “age”, “region” and “placesize”

	$iS$	$DiS$	$DiSCO$	$DiSDiO$	max_denom	mean_denom	$DCAP$
1	64.90	34.18	9.54	6.14	3.00	1.16	16.20
2	64.00	32.50	10.26	6.78	4.00	1.19	17.45
3	64.02	32.14	9.10	5.92	4.00	1.19	15.92
4	63.88	33.38	9.20	5.52	4.00	1.21	16.12
5	63.44	31.46	9.34	5.80	4.00	1.23	16.38

The  $Dorig$  and  $DiSCO$  measures are not restricted to disclosures that are identified from unique records for a  $q$  value and a level of the target in either the GT or the SD. For each disclosive record, the number of SD records with the same level of the target and the same  $q$  in the SD is the denominator that applies to that record. The columns `max denom` and `mean denom` refer to the denominators

in the SD that contribute to the *DiSCO* measures. We can see from the mean that here the majority of disclosive records had unique key combinations in the SD, and the maxima was 3 for the first synthesis and 4 for the others. The disclosure functions in **synthpop** use information on large denominators to check for a key in  $q$  that is highly predictive of one level of the target. When such a relationship is identified it causes a two-way check to be triggered and details printed. A target where one level accounts for a very high proportion of the disclosures can also lead to high levels of *DiSCO* for similar reasons and this may be flagged as a one-way check.

Results for *DiSCO* as an attribute disclosure measure for the SD and *Dorig* for the GT are shown in Fig. 1, for five targets from the example discussed above. As well as **depress**, the other variables are **ls**, a score for life satisfaction, **workab**, the intention to work abroad, **marital**, marital status and **income** income with 8, 3, 9 and 407 distinct values respectively<sup>11</sup>. The plot is ordered by the disclosure measure for the SD.

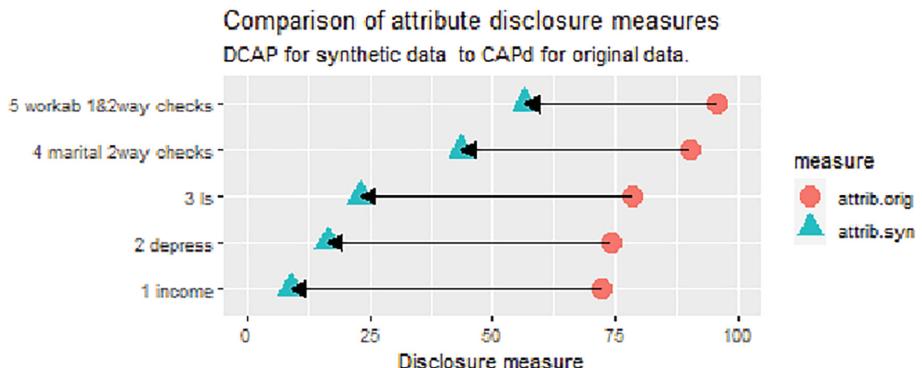


**Fig. 1.** Attribute disclosure measures *Dorig* and *DiSCO* for 5 targets from synthesis of the SD2011 data identified from keys “sex”, “age”, “region” and “placesize”

The top two variables **workab** and **marital** are flagged as requiring checking; **workab** for both one-way and two-way relationships and **marital** for two-way. Detailed output from the disclosure functions give information on the target levels for one-way checks and the target-key pairs that contribute most to two-way checks. In this example it was the answer “NO” for “workab” that was flagged by the one-way check; most survey respondents (89%) had never worked abroad and 93% of all disclosive records had this level of the target. For “marital” several target-key pairs contributed to the two-way check, such as the marital status “single” for the youngest ages.

Figure 2 shows the same analysis as Fig. 1, but using *DCAP* as a disclosure measure. It shows a very similar pattern to Fig. 1, although all measures are

<sup>11</sup> including missing value categories.



**Fig. 2.** Attribute disclosure measures  $CAPd$  and  $DCAP$  for 5 targets from synthesis of the SD2011 data identified from keys “sex”, “age”, “region” and “placesize”

higher. Most importantly the two variables flagged for one-way or two-way checks show similar disclosure patterns as we saw for the  $DiSCO$  measure. These checks, and what to do about them, are discussed further in Sect. 4.

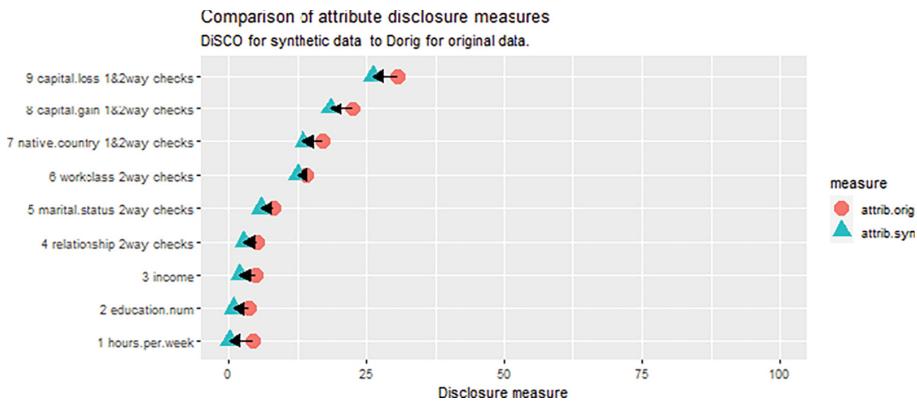
Note that, by default, our disclosure methods treat all variables as if they were categories. Although the `income` variable may have been an exact value in the original survey, it is rounded in the data that are released so that there are only 406 distinct values in the 5000 records used here. This is typical of all data made available to researchers by NSAs. If SD contains more finely-grained continuous data, there are two possible approaches. The first would be to count as disclosure records where the distance between the GT and SD is lower than a threshold. The second is to group the continuous variables into categories that are small enough to make knowledge of them be considered disclose. The second, simpler, approach is implemented in the `synthpop` package by allowing the user to specify the number of groups into which each numerical key or target will be categorised. For the target `income` with 406 distinct values in Fig. 1 gave a  $DiSCO$  of 4.90%. Grouping it into 20 categories increased  $DiSCO$  to 6.14%.

## 4 Excluding Disclosive Records that are Not Risky

Having identified records where the apparent disclosure is something that would generally be known, one option is to exclude these key-target combinations explicitly from the measures. For example:

- All records with levels of the target identified as contributing a high proportion of disclosive records can be excluded from all disclosure measures. These can be identified from the one-way checks described in the previous section.
- Selected key-target pairs can be excluded from all measures. These can be identified from the two-way checks described in the previous section.
- Missing values of some or all of the keys or the targets can be excluded from leading to a disclosure

- Disclosure measures can be restricted to those disclosive records where the denominators of the disclosive cells for the level of  $q$  and  $t$  are less than or equal to a limit defined by the user. A denominator limit of 1 will restrict disclosure in the SD to records that are unique for the target  $t$  and the combined identifiers  $q$ .



**Fig. 3.** Attribute disclosure measures *Dorig* and *DisCO* for 9 targets from synthesis of the Adult data identified from keys “sex”, “age”, “occupation” and “race”

To illustrate exclusions, we use the Adult data set from the UCI machine learning repository [19] with almost 50 thousand records from the US Census income study. The data set was synthesised with the default method (CART) in **synthpop**. Figure 3 summarises the attribute DR from the keys **age**, **occupation**, **race** and **sex** for the other 9 variables in the data<sup>12</sup>. The disclosiveness of the original data is relatively low, compared to the previous example, because of the large sample size and the absence of any geographic identifiers. The three variables with the highest attribute disclosure **capital.gain**, **capital.loss** and **native.country** are flagged to check one-way relationships. Details show that the levels contributing to disclosure are zero for the first two and **United-States** for **native.country**. These levels make up 95%, 92% and 89% of all disclosive records, respectively.

Three other targets are flagged as having disclosive two-way relationships **workclass**, **marital**, **relationship**. Detailed output from the disclosure functions showed that the largest contribution to two-way disclosures for **workclass** was due to this being missing when occupation was missing, although other relationships between these two variables also contributed. For **marital**, the largest contributions came from the key **age** with patterns similar to **marital** in our first example.

<sup>12</sup> Although the data set lists 15 variables, two are identical except that one (education) appears as both a factor and a numeric variable and We have excluded the weight, as it is not an analysis variable.

**Table 2.** Attribute disclosure measured by *DiSCO* from Adult data from the keys age, occupation, race and sex and different exclusion criteria.

target	Column 0		Column 1		Column 2		Column 3		Column 4	
	No exclusions		One target level excluded		missing values excluded		denominator limit 1 and a) and b)		denominator limit 1 only	
	orig	syn	orig	syn	orig	syn	orig	syn	orig	syn
capital.gain	22.55	18.83	0.21	0.00	0.21	0.00	0.21	0.00	2.68	1.18
capital.loss	30.61	26.35	0.08	0.00	0.08	0.00	0.08	0.00	2.68	1.23
education.	3.71	1.00	3.71	1.00	3.71	1.00	2.68	0.44	2.68	0.44
hours.per.week	4.36	0.19	4.36	0.19	4.36	0.19	2.68	0.14	2.68	0.14
income	4.97	2.10	4.97	2.10	3.51	1.74	1.74	0.58	2.68	0.79
marital.status	8.23	5.18	8.23	5.18	8.23	5.18	2.68	0.77	2.68	0.77
native.country	17.09	13.85	0.94	0.09	0.83	0.08	0.73	0.07	2.68	0.86
relationship	5.17	2.66	5.17	2.66	5.17	2.66	2.68	0.72	2.68	0.72
workclass	14.27	11.71	14.27	11.71	9.14	6.91	2.45	0.81	2.68	0.93

Table 2 gives the results of excluding different entries in the tables of  $q$  and  $t$  from the attribute disclosure measures. Excluding the levels of the target flagged by one-way checks (capital gain, capital loss and native country) reduces the *DiSCO* to almost zero for these 3 variables. Adding the exclusion of missing values reduced the disclosure for `workclass` and for some other variables a little. Adding the restriction to denominators of 1 reduced the disclosure for variables, identified as requiring two-way checks, to low levels. In the final columns we see that restricting to denominators of 1, by itself, gives levels of disclosure below 1% for all but the first two variables. Note that the attribute disclosure of the original data with denominators of 1 is the same for all targets at 2.68%. This is the same as the % unique values of  $q$  in the GT, because all records with a unique  $q$  are disclosive in the GT for all targets.

Which of these columns in Table 2 represents an approach that an NSA should adopt to exclude these one-way and two-way contributions from the attribute disclosure measures? In the case of the Adult data this is a rhetorical question because the data have already been made available to the public. Both column 3 or column 4 give low risks. Column 2 gives low risks except for the three targets identified as having contributions from two-way  $tq$  relationships. This could be remedied by excluding the specific  $tq$  combinations, but at the cost of more examination of the data. Column 4 offers a quick and easy solution, but perhaps at the price of failing to count some really disclosive records in the DRs.

The choice of using the easy measure (Column 4) or the more onerous one (modified column 2) may depend on the context of the release, as discussed in Sect. 2, factor (a). For SD that will be released to trusted researchers, with guarantees that it will not be made public, the easy measure given by the method in Column 4 might be enough. But for SD to be openly shared, the data holder would be expected to carry out more extensive DR assessment.

This paper presents only a very limited evaluation of the DR measures, using two examples. It is hoped that the availability of the new `synthpop` tools for DR assessment will increase our understanding of DR, the factors that influence it and how best it should be assessed.

## 5 Can We or Should We Exclude Utility from Attribute Disclosure Risk and Does it Matter?

Inevitably, measures of utility and attribute DR have much in common. Attribution risk involves the ability to predict one variable from others. This is clearly related to utility. In a comprehensive review of privacy metrics Wagner and Eckhoff [34] note that authors have suggested that “privacy metrics are orthogonal to cost and utility metrics”. Yet, several of the metrics they propose are have also been cited as general utility metrics [23].

When disclosure is an aspect of utility, it may still be a real disclosure for the individual concerned, if it correctly predicted a sensitive attribute. The exception would be if it would be the result of a relationships in the data that was already known. Ideally we would have a full specification of the intruder’s prior belief about what the relationships in the data might be. With that information we could assess the additional information that the SD provides compared to the prior [10]. Specification of a prior would be an onerous task, so the simplification we propose in Sect. 4 is an attempt to identify some of the prior information that might be available for low-dimensional relationships in the data<sup>13</sup>. It is specific to the knowledge of the intruder. For example, a table of UK occupations by income might identify that all university vice-chancellors had incomes in the highest group. Whether this would be a new DR would depend on what was known about UK academic salaries.

A recent proposal of a framework for quantifying privacy risk, entitled the Anonymeter, has been proposed by Giomi et al. [8] who provide open source code at <https://www.anonos.com/products/anonymeter> where they also post a quote from a letter to them from the French data protection authority, the Commission Nationale de l’Informatique et des Libertés (CNIL) as:

“The results produced by the tool Anonymeter should be used by the data controller to decide whether the residual risks of re-identification are acceptable or not, and whether the dataset could be considered anonymous”.

They propose methods to assess privacy risk as *singling out*, *linkability* and *inference*. Their measures of *singling out* and *inference* appear to be identical to the measures *RepU* and *DCAP* described here. However they are computed by a simulation rather than from the exact formulae given in Appendix 1. They claim that the Anonymeter provides a method that makes it possible to “measure how much of an attacker’s success is simply due to the *utility* of the SD and how much instead is an indication of *privacy violation*”. They propose is the following:

- divide the GT data into two sections as **Training** and **Control**. **Training** is used to create the SD and the disclosiveness is assessed from **Control**. They suggest that **Control** might be a smaller fraction (e.g. 20%) so as to maintain the quality of the SD

---

<sup>13</sup> While restricting to one-way or two-way relationships seems limited, in practice stratification by segments of the GT such as area can make this more effective.

- create SD based on the Training data
- measure the DR for the Control data from this SD.

The authors argue that the success rate of the Control is a consequence of general inference (i.e. utility) and the extent to which the Training risk exceeds the Control is an indication that information has been leaked from the SD. They propose a measure  $R$  which the term a “specific privacy risk” as

$$R = \frac{(r_{train} - r_{control})}{1 - r_{control}}, \quad (1)$$

where  $r_{train}$  and  $r_{control}$  are the proportions of disclosive records identified in the control and training data. It is intended to measure the proportion of the DR that excludes disclosure due to utility ( $r_{control}$ ). A limited comparison of this method has been carried out with the analyses of attribute disclosure from the example in Sect. 3.4. For the five targets illustrated in Figs. 1 and 2, attribute disclosure was calculated using the Anonymeter methodology for a Control with 1000 records (20% of the GT data) and a Training data set of the remaining 80%. Results are given in Table 3.

**Table 3.** Attribute disclosure measures  $DiSCO$  and  $DCAP$  calculated from the training and control segments of the GT data with SD created by the training data. The columns “All” are each measure from the complete data as illustrated in Figs. 1 and 2.

	$DiSCO$				$DCAP$			
	All	Train	Ctrl	R	All	Train	Ctrl	R
income	4.90	5.53	1.90	0.04	8.91	9.06	2.92	0.06
depress	9.54	9.45	4.20	0.05	16.39	15.98	5.80	0.11
ls	13.78	11.88	8.60	0.04	23.03	19.79	12.53	0.08
marital	35.18	34.50	21.30	0.17	43.66	41.50	24.94	0.22
workab	52.10	49.28	31.30	0.26	56.56	52.92	33.31	0.29

There is some agreement here with the results in Figs. 1 and 2. The two targets (marital and workab) that were identified as having large contributions from one-way or two-way relationships are also those identified here as having the largest Control measures that [8] identify as being the contribution to attribute disclosure from utility. The attribute disclosure measures for “workab” after excluding records answering “NOc” as disclosive became 1.1% for  $DiSCO$  and 1.6% for  $DCAP$ , are much lower than the Ctrl values in Table 3 of 12.5% and 13.3%. This suggests that the anonymeter procedure may not be identifying all of the utility. The same was true for “marital” when the identified  $tq$  combinations were excluded from the disclosure measures.

Further investigations of the comparison between the anonymeter methods and what is proposed in this paper may help the understanding of both methods.

It is interesting to note the qualification by the CNIL after the quote from their letter above:

“The anonymity of a synthetic dataset can only be determined on a case-by-case basis, i.e. for each generated dataset, and should therefore not be assumed from analyses performed on other datasets coming from the same provider or data synthesis tool. However, as a solution provider, the company Static should also provide practical elements describing precisely how to use the tool and interpret the results obtained (such as examples, tutorials, thresholds, etc.)”.

The metrics introduced in Sect. 3 have been developed for practical use by data custodians who are deciding on the release of each individual SD. This is not so clear in the anonymeter paper [8]. There is no attempt there to identify sensitive variables, or those variables whose value would be expected to be known for individuals in the data. Results are presented that are averaged over all variables in the GT as targets using random selections from the remaining variables as keys. It may be that their metrics have some value for evaluating the DR of methods to create SD rather than evaluating individual synthetic data sets.

## 6 Integrating Utility and DR Measures Into Synthetic Data Creation

Before a data holder decides to release SD it needs to be evaluated for Utility and DR. The person creating the SD has the means that allow each of these to be altered. They include the synthesis methods that are used, and also SDC procedures that can be carried out on either the GT or the SD. This will need to be an iterative process where the synthesis methods and the SDC are adapted until the SD is satisfactory with respect to utility and DR.

These modifications of the SD occur at three stages:

- (a) Pre-processing the GT before synthesis.
- (b) Changing the methods used to create the SD.
- (c) Post-processing the SD before it is released.

Techniques that can be used at (a) and (c) include recoding variables with small categories, grouping or smoothing numeric variables and the removal of records that are unique in the GT data or are replicated uniques in SD, as well as other techniques from Statistical disclosure control [11].

Under item (b) many different methods can be used to create SD and each may have parameters that can be tuned to influence utility or DR. The default method in **CART** can be adjusted to improve utility via parameters like the sequence of the conditional distributions and the restriction of the predictors used for each conditional model; see [22] for examples. An adjustment of the **CART** method to limit the number of records in the terminal nodes of the tree can improve DR.

Using synthesis methods that satisfy differential privacy(DP) have been advocated as a method of reducing the DR of SD. Our methods of assessing DR are, in one sense, the opposite of the formulation of DP. Our proposals are very specific as to what is known about individuals in the data and to knowing which items may be sensitive to being disclosed. Differential privacy, on the contrary, attempts to protect against arbitrary outside knowledge on the part of an intruder. There have been several important criticisms of DP SD: see for example [30] in terms of its lack of utility, and even its poor DR. These underline the importance of using independent DR assessment even for DP synthesis.

## 7 Discussion and Future Work

This paper has shown how using routines to calculate DR measures on real data sets can lead to understanding and improving the DR metrics. It is hoped that more feedback will improve the metrics even more. The author would welcome feedback on the new **synthpop** disclosure measures either to suggest improvements or (more likely at this stage) to report bugs or other problems.

There is much more work that could be done on factors that influence the DR of synthetic data. One important aspect would be to investigate whether over-fitting the synthesis model will increase DR. An example of a potentially over-fitted model is the use of a saturated model for categorical data from a cross-tabulation of all the variables. This has been proposed by Jackson et al. [13] and is implemented as the method **cata11** in **synthpop**. The disclosure of SD created by **cata11** could be compared to that created using the **ipf** which uses iterative proportional fitting to a set of defined margins to create the SD.

Further work comparing the metrics presented here with those recommended by Anonymeter may help us to clarify the differences between utility and attribute DR for SD.

**Acknowledgement.** Research Data Scotland (<https://www.researchdata.scot/>) has funded some of Gillian Raab's time to carry out the research reported here and to expand the capabilities of the **synthpop** package (from version 1.8-0 available on CRAN at <https://CRAN.R-project.org/package=synthpop>) to include measures of disclosure control. We also thank the Scottish Centre for Administrative Data Research for continue to support the development of the **synthpop** package since its creation was supported by the ESRC funded SYLLS project in 2012–14. We would also like to thank two anonymous referees for their helpful comments.

## Appendix 1: Notation and Formal Definitions

### Computational Approach and Notation

Before defining the measures of identity and attribute DRs we need to introduce the notation that will be used to calculate them. The first step is to create the quasi-identifiers from the keys for the GT and SD. For the keys used in the

example given in Sect. 3.4 the quasi-identifier that we will designate as  $q$  for the first record in the original data is:

"FEMALE | 57 | Lubuskie | URBAN 100,000–200,000"

and that for the first record in the synthetic data:

"FEMALE | 39 | Zachodnio-pomorskie | URBAN 100,000–200,000".

In order to calculate identity disclosure measures, we need to compare the tables of  $q$  from the GT and SD. For attribute disclosure measures we need to cross-tabulate  $q$  with each target variable  $t$  and compare findings from the SD with what would have been found from the GT. In general, the levels of  $q$  and sometimes  $t$  in the GT and SD will not be the same. Before creating any tables, we need to define sets of  $q$  and  $t$  values that give the union of both sets of levels and align the tables so that their indices correspond.

For the GT data  $d_{.q}$  is the count of records with the keys corresponding to the levels of  $q$  and  $d_{tq}$  the count of records with this  $q$  and level  $t = 1, \dots, T$  of the target. The equivalent counts from the synthesised data are designated by  $s_{.q}$  and  $s_{tq}$ . When a member of  $q$  is in the GT data but not in the SD,  $s_{.q}$  and  $s_{tq}$  are all zero. Similarly when a member of  $q$  is in the SD but not in the GT,  $d_{.q}$  and  $d_{tq}$  are all zero. The two tables can be written as shown in Table 4, where the total records in the GT data is  $N_d$ , made up of  $N_d$  *only* and  $N_d$  *both*. The equivalent totals for the SD are  $N_s$ ,  $N_s$  *only* and  $N_s$  *both*.

**Table 4.** Notation for tables from quasi-identifier ( $q$ ) and target ( $t$ ) from GT (upper table) and SD (lower table).

	only in original			in both			only in synthetic			Total
1	...	$d_{1q}$	...	...	$d_{1q}$	...	...	0	...	$d_{1.}$
...	...	...	...	...	...	...	...	...	...	...
t	...	$d_{tq}$	...	...	$d_{tq}$	...	...	0	...	$d_{t.}$
...	...	...	...	...	...	...	...	„,	...	...
T	...	$d_{Tq}$	...	...	$d_{Tq}$	...	...	0	...	$d_{T.}$
Column sums	$d_{.q}$		...	$d_{.q}$		...	...	0	...	$N_d$
Totals	$N_d$ <i>only</i>			$N_d$ <i>both</i>			0			$N_d$
	only in original			in both			only in synthetic			Total
1	...	0	...	...	$s_{1q}$	...	...	$s_{1q}$	...	$s_{1.}$
...	...	...	...	...	...	...	...	...	...	...
t	...	0	...	...	$s_{tq}$	...	...	$s_{tq}$	...	$s_{t.}$
...	...	...	...	...	...	...	...	...	...	...
T	...	0	...	...	$s_{Tq}$	...	...	$s_{Tq}$	...	$s_{T.}$
Column sums	...	0	...	...	$s_{.q}$	...	...	$s_{.q}$	...	$N_s$
Totals	0			$N_s$ <i>both</i>			$N_s$ <i>only</i>			$N_s$

## Identity Disclosure

To calculate the % of unique records in the GT and synthetic data we need:

$$\% \text{ Unique in Original} = \text{UiO} = 100 \sum (d_{.q}|d_{.q} = 1)/N_d. \quad (2)$$

$$\% \text{ Unique in Synthetic} = \text{UiS} = 100 \sum (s_{.q}|d_{.q} = 1)/N_d. \quad (3)$$

The intruder has information about the keys for an individual in the GT that they attempt to identify in the SD. They first attempt to find them in the SD, and the % found is:

$$\% \text{ Unique in Original in Synthetic} = \text{UiOoS} = 100 \sum (d_{.q} = 1|s_{.q} = 1 \wedge d_{.q} > 0)/N_d. \quad (4)$$

Some of these records would not be unique in the SD, restricting to such records gives:

$$\% \text{ replicated Uniques} = \text{repU} = 100 \sum (s_{.q}|d_{.q} = 1 \wedge s_{.q} = 1)/N_d. \quad (5)$$

## Attribute Disclosure

To find an attribute from a set of keys, it is necessary to examine the distribution of  $s_{tq}$  for groups defined by  $q$ . We define column proportions for the GT and SD as  $pd_{tq} = d_{tq}/d_{.q}$  and for the synthetic as  $ps_{tq} = s_{tq}/s_{.q}$ .

The measure  $DCAP$  is the percentage of  $t$  correctly predicted in the SD: This gives

$$DCAP = 100 \sum_{tq} (ps_{tq}d_{tq})/N_d \quad (6)$$

As a comparator for  $DCAP$  we need the % that would be predicted with someone with access to the GT, giving  $CAPd$ :

$$CAPd = 100 \sum_{tq} (pd_{tq}d_{tq})/N_d, \quad (7)$$

Returning to the scenario described in Sect. 3.1, we first calculate a measure of attribute disclosure for the GT data that requires that each set of records defined by  $q$  has a unique value of the target. This is an attribute disclosure measure for the GT data:  $\% \text{ Disclosive in Original}$ :

$$Dorig = 100 \sum^q \sum^t (d_{tq}|pd_{tq} = 1)/N_d. \quad (8)$$

An intruder with access only to the SD, but with knowledge of  $q$  from one or more individuals in the GT, would look them up in the SD. Some of their  $q$  levels be key combinations that do not appear in the SD leaving the proportion that do appear as  $iS$  (in Synthetic)

$$iS = 100 \sum^q \sum^t (d_{tq}|s_{tq} > 0)/N_d. \quad (9)$$

A level of  $q$  from a GT record with level  $i$  of  $t$  may identify any level  $j$  of the target as disclosive in the SD giving

$$DiS = 100 \sum^q \sum^{i=1,..T} \sum^{j=1,..T} (d_{iq}|ps_{jq} = 1)/N_d. \quad (10)$$

Some of these will identify the wrong target. To exclude these we restrict to records where  $i = j$  giving Disclosive in Synthetic Correct Original:

$$DiSCO = 100 \sum^q \sum^t (d_{tq}|ps_{tq} = 1)/N_d. \quad (11)$$

Note that  $DiSCO$  can include records that are not disclosive in the GT data giving a further measure Disclosive in Synthetic and Disclosive in the Original:

$$DiSDiO = 100 \sum^q \sum^t (d_{tq}|ps_{tq} = 1 \wedge pd_{tq} = 1)/N_d. \quad (12)$$

As we comment above the intruder would not be able to tell if records were identified as  $DiSDiO$  rather than  $DiSCO$ , so we prefer the latter measure. However, the intruder can identify when the apparently disclosive record is not unique in the SD. This restriction can be imposed by requiring the denominator in the SD not to exceed a 1, as described and discussed in Sect. 4.

The measures  $DCAP$  and  $DiSCO$  computed by **synthpop** use the total number of records in the GT as a denominator. This differs from the measures used by [6] and [31]. Their  $DCAP$  is the same as Eq. 6, except that it uses as denominator  $N_{d\ both}$ , the count of records in the GT that have  $q$  values represented in the SD. This denominator is also used to calculate  $TCAP$  in [16], thus differentiating it from  $DiSCO$ , giving

$$TCAP = 100 \sum^q \sum^t (d_{tq}|ps_{tq} = 1)/N_{d\ both}. \quad (13)$$

Both  $DCAP$  and  $TCAP$  can be scaled with respect to the disclosure that would be found for someone who had access to the marginal distribution of the target. The intruder guesses the level for each target according to the relative frequencies  $pd_t$ . in the GT. Averaging this over all observations gives

$$baseCAPd = \sum (pd_t)^2/N_d. \quad (14)$$

The  $DCAP$  or  $TCAP$  score can be expressed by scaling from 1 to  $baseCAPd$  [16, 21], although this can result in some negative values.

## Appendix 2: R Code for Examples Analysed in Sects. 3 and 4

```
#####
# R code #####
devtools::install_github("Gillian-Raab/synthpop", build_vignettes = TRUE)
library(synthpop)
rm(list=ls())
ods <- SD2011[, c("sex", "age", "region", "placesize", "depress",
  "income", "ls", "marital", "workab")]
#####
#####----- Table 1 -----
s1 <- syn(ods, seed = 8564, cont.na = list(income = -8))
s5 <- syn(ods, seed = 8564, m = 5, cont.na = list(income = -8))
t5 <- disclosure(s5, ods, print.flag = TRUE, target = "depress",
  keys = c("sex", "age", "region", "placesize"))
t5
#####
#####----- Figures 1 and 2 -----
tti <- disclosure.summary(s1, ods, keys = c("sex", "age", "region", "placesize"))
tti
ttib <- disclosure.summary(s1, ods, attrib.meas ="DCAP", keys = c("sex", "age", "region", "placesize"))
ttib
## grouping
disclosure(s1, ods, target = "income", keys = c("sex", "age", "region", "placesize"))
disclosure(s1, ods, target = "income", ngroups_target = 20, keys = c("sex", "age", "region", "placesize"))
#####
#####----- ADult data set analis -----
library(arules)
data(AdultUCI)
##syn.AdultUCI <- syn(AdultUCI) gives warning about "-" in variable names
names(AdultUCI) <- gsub("-", ".", names(AdultUCI))
myAdult <- AdultUCI
for (i in c(5,15)) myAdult[,i] <- factor(as.character(myAdult[,i])) ## logical and ordered factor changed to factor
myAdult <- myAdult[,-c(3:4)] ## drop fnlwgt and one of education
# order changed to put education.num (now a factor) at the end
system.time(
  syn1.myAdult<- syn(myAdult, cont.na = list(capital.gain = 99999, hours.per.week =99,
    method = c("sample",rep("ctree", 12)), visit.sequence = c(1,2,4:13,3) ,models = TRUE)
)
summary.disc1 <- disclosure.summary(syn1.myAdult, myAdult,
  key = c("age","sex","occupation","race"))
summary.disc1

#####
#####----- table 2 -----
tab1 <- summary.disc1$attrib.table[,1:2]
dimnames(tab1)[[1]] <- substring(dimnames(tab1)[[1]],3)
dimnames(tab1)[[1]]
tab1 <- tab1[order(dimnames(tab1)[[1]]),]; tab1
targs <-names(myAdult)[!names(myAdult)] summary.disc1_nottarg <- disclosure.summary(syn1.myAdult, myAdult,
  not.targetslev= c("", "", "", "0", "0", "", "United-States", ""),
  key = c("age","sex","occupation","race"))
tab2 <- summary.disc1_nottarg$attrib.table[,1:2]
dimnames(tab2)[[1]] <- substring(dimnames(tab2)[[1]],3)
tab2 <- tab2[order(dimnames(tab2)[[1]]),]; tab2
summary.disc1NA <- disclosure.summary(syn1.myAdult, myAdult,usetargetsNA = FALSE,
  not.targetslev= c("", "", "", "0", "0", "", "United-States", ""),
  key = c("age","sex","occupation","race"))
tab3 <- summary.disc1NA$attrib.table[,1:2]
dimnames(tab3)[[1]] <- substring(dimnames(tab3)[[1]],3)
tab3 <- tab3[order(dimnames(tab3)[[1]]),]; tab3
summary.disc1NAdi <- disclosure.summary(syn1.myAdult, myAdult,usetargetsNA = FALSE,
  denom_lim =1, exclude_ov_denom_lim = TRUE, not.targetslev= c("", "", "", "0", "0", "",
  "United-States", ""),
  key = c("age","sex","occupation","race"))
tab4 <- summary.disc1NAdi$attrib.table[,1:2]
dimnames(tab4)[[1]] <- substring(dimnames(tab4)[[1]],3)
tab4 <- tab4[order(dimnames(tab4)[[1]]),]; tab4
summary.d1 <- disclosure.summary(syn1.myAdult, myAdult, denom_lim =1,
  exclude_ov_denom_lim=TRUE,
  key = c("age","sex","occupation","race"))
tab5 <- summary.d1$attrib.table[,1:2]
dimnames(tab5)[[1]] <- substring(dimnames(tab5)[[1]],3)
tab5 <- tab5[order(dimnames(tab5)[[1]]),]; tab5
alltab <- cbind(tab1,tab2,tab3, tab4, tab5)
round(alltab,2)
#####
```

## References

1. Bowen, C.M., Snoke, J.: Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. *J. Priv. Confidentiality* **11**(1) (2021)
2. Chen, Y., Taub, J., Ellieot, M.: Trade-off between information utility and disclosure risk in GA synthetic data generator. UNECE Work Session on Statistical Data Confidentiality, Skopje, North Macedonia (2019). [https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019\\_S3\\_UK\\_Chen\\_Taub-Elliot\\_AD.pdf](https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Chen_Taub-Elliot_AD.pdf). Accessed 06 Jan 2024
3. Dalenius, T.: Finding a needle in a haystack. *J. Off. Stat.* **2**, 329–336 (1986)
4. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. *Comput. Stat. Data Anal.* **55**(12), 3232–3243 (2011). <https://doi.org/10.1016/j.csda.2011.06.006>
5. Drechsler, J., Haensch, A.C.: 30 years of synthetic data. *Stat. Sci.* **39**(2), 221–242 (2024). <https://doi.org/10.1214/24-STS927>
6. Elliot, M.: Final report on the disclosure risk associated with the synthetic data, produced by the SYLLS team (2014). <https://tinyurl.com/syllsDR>. Accessed 23 Feb 2022
7. Elliot, M., Mackey, E., O’Hara, K.: The anonymisation decision-making framework: European practitioners (2020). <https://ukanon.net/framework/>. Accessed 23 Feb 2022
8. Giomi, M., Boenisch, F., Wehmeyer, C., Tasnádi, B.: A unified framework for quantifying privacy risk in synthetic data (2022). <https://arxiv.org/abs/2211.10459>. Accessed 30 June 2024
9. Goodfellow, I.J., et al.: Generative adversarial networks (2014)
10. Hu, J., Reiter, J.P., Wang, Q.: Disclosure risk evaluation for fully synthetic categorical data. In: Domingo-Ferrer, J. (ed.) PSD 2014. LNCS, vol. 8744, pp. 185–199. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-11257-2\\_15](https://doi.org/10.1007/978-3-319-11257-2_15)
11. Hundepool, A.: Statistical Disclosure Control, 1st edn. Wiley, Chichester (2012)
12. Jackson, J., Mitra, R., Francis, B., Dove, I.: Using saturated count models for user-friendly synthesis of large confidential administrative databases. *J. R. Stat. Soc. A. Stat. Soc.* **185**, 1613–1643 (2022)
13. Jackson, J., Mitra, R., Francis, B., Dove, I.: Using saturated count models for user-friendly synthesis of categorical data. *J. Roy. Statist. Soc. Series A* (2022, accepted). <https://arxiv.org/abs/2107.08062v2>
14. Kaloskampis, I., Joshi, C., Cheung, C., Pugh, D., Nolan, L.: Synthetic data in the civil service. *Significance* **17**, 18–23 (2021)
15. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A framework for evaluating the utility of data altered to protect confidentiality. *Am. Stat.* **60**(3), 224–232 (2006)
16. Little, C., Elliot, M., Allmendinger, R.: Comparing the utility and disclosure risk of synthetic data with samples of microdata. In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 234–249. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_17](https://doi.org/10.1007/978-3-031-13945-1_17)
17. Little, R.J.A.: Statistical analysis of masked data. *J. Off. Stat.* **9**(2), 407–26 (1993)
18. Machanavajjhala, A., Gehrke, J., Kifer, K., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. In: 22nd International Conference on Data Engineering (ICDE 2006). IEEE (2006)

19. Newman, C.B.D., Merz, C.: UCI repository of machine learning databases (1998). <http://www.ics.uci.edu>
20. Nowok, B., Raab, G., Dibben, C.: Recognising real people in synthetic microdata: risk mitigation and impact on utility. UNECE Work Session on Statistical Data Confidentiality, Skopje, North Macedonia (2017). <https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/3.risk.mitigation.pdf>. Accessed 23 Feb 2022
21. Pater, L., Smid, S.: Making attribute information of synthetic data interpretable with the aggregation equivalence level. UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS Expert Meeting on Statistical Data Confidentiality, 26–28 September 2023, Wiesbaden (2023). [https://unece.org/sites/default/files/2023-08/SDC2023\\_S4\\_3-Netherlands\\_Pater\\_D.pdf](https://unece.org/sites/default/files/2023-08/SDC2023_S4_3-Netherlands_Pater_D.pdf). Accessed 23 Feb 2024
22. Raab, G.M., Nowok, B., Dibben, C.: Guidelines for producing useful synthetic data (2017). <http://arxiv.org/abs/1712.04078>
23. Raab, G.M., Nowok, B., Dibben, C.: Assessing, visualizing and improving the utility of synthetic data (2021). <https://arxiv.org/pdf/2109.12717.pdf>. Accessed 24 June 2024
24. Raab, G.M., Nowok, B., Dibben, C.: Practical privacy metrics for synthetic data (2024). <https://arxiv.org/abs/2406.16826>. Accessed 27 June 2024
25. Raab, G.: Utility and disclosure risk for differentially private synthetic categorical data. In: Muralidhar, K., Domingo-Ferrer, J. (eds.) PSD 2022. LNCS, vol. 13463, pp. 250–265. Springer, Berlin (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_18](https://doi.org/10.1007/978-3-031-13945-1_18)
26. Reiter, J.: Synthetic data: a look back and a look forward. Trans. Data Priv. **16**, 15–24 (2023)
27. Rubin, D.B.: Discussion: statistical disclosure limitation. J. Off. Stat. **9**(2), 461–8 (1993)
28. Shokri, R., Strobel, M., Zick, Y.: On the privacy risks of model explanations. [arXiv.org](https://arxiv.org) (2021)
29. Snöke, J., Raab, G., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. J. Roy. Statist. Soc. Serues A **181**(3), 663–688 (2018)
30. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data – anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22), Boston, MA, pp. 1451–1468 (2022). <https://www.usenix.org/conference/usenixsecurity22/presentation/stadler>
31. Taub, J., Elliot, M., Pampaka, M., Smith, D.: Differential correct attribution probability for synthetic data: an exploration. In: Domingo-Ferrer, J., Montes, F. (eds.) PSD 2018. LNCS, vol. 11126, pp. 122–137. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-99771-1\\_9](https://doi.org/10.1007/978-3-319-99771-1_9)
32. UNECE: Synthetic data for official statistics; a starter guide (2022). <https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf>. Accessed 01 Jan 2024
33. Voas, D., Williamson, P.: Evaluating goodness-of-fit measures for synthetic microdata. Geogr. Environ. Model. **5**, 177–200 (2001)
34. Wagner, I., Eckhoff, D.: Technical privacy metrics: a systematic survey. ACM Comput. Surv. **51**(3), 1–38 (2018)
35. Woo, M.J., Reiter, J.P., Oganian, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. J. Priv. Confidentiality **1**, 111–124 (2009)



# Attribute Disclosure Risk in Smart Meter Data

Guillermo Navarro-Arribas<sup>1</sup> and Vicenç Torra<sup>2</sup>

<sup>1</sup> Department of Information and Communications Engineering,  
Universitat Autònoma de Barcelona, Barcelona, Spain  
[guillermo.navarro@uab.cat](mailto:guillermo.navarro@uab.cat)

<sup>2</sup> Department of Computing Science, Umeå University, Umeå, Sweden  
[vtorra@cs.umu.se](mailto:vtorra@cs.umu.se)

**Abstract.** This paper studies attribute disclosure risk in aggregated smart meter data. Smart meter data is commonly aggregated to preserve the privacy of individual contributions. The published data shows aggregated consumption, preventing the revelation of individual consumption patterns. There is, however, a potential risk associated to aggregated data. We analyze some datasets of smart meter data consumption to show the potential risk of attribute disclosure. We observe that, even if data is aggregated with the most favorable aggregation approach, it presents this attribute disclosure risk.

**Keywords:** Smart meter data ·  $k$ -anonymity · Attribute disclosure

## 1 Introduction

Smart meter data provides a detailed insight on home energy consumption and can be used for different purposes: prediction of future consumption, redesign of power grids, or simply for marketing and public knowledge. Several approaches exist to anonymized smart meter data, and most of them are based on providing some sort of aggregation [1, 4, 11]. That is, data is aggregated to preserve the privacy associated to each smart meter, and consequently to each home power consumption.

There are different approaches to aggregate smart meter data with privacy purposes. One approach is to globally aggregate all the data from all available sources. In other schemes the data is aggregated geographically in distribution points. In any case, aggregated data is usually considered safe. Several smart meters (homes) readings are aggregated together forming an anonymity set, this is commonly extended to ensure  $k$ -anonymity [17] in the released dataset. There are however some potential risks associated to aggregated smart meter data. In this paper we investigate a potential attribute disclosure attack, where a user, whose data is included in the dataset, can estimate some partial information from the anonymized dataset.

The idea is to check if the aggregation of data could lead to situations where the aggregated value can be used to approximate the original value of a respondent. To that end, we investigate smart meter data aggregated using microaggregation, which can be seen as the most favorable aggregation to avoid this kind of attacks.

The goal of the paper is to outline the problem and make aware the community about it. This paper complements our previous results [2,3] discussing some specific attacks to smart grid data. More particularly, we used Non-Intrusive Load Monitoring (NILM) with the goal of detecting individual appliances in aggregated data.

The paper is organized as follows. Section 2 discusses smart meter data and its aggregation, Sect. 3 describes the attribute disclosure problem in aggregated data, and Section 4 analyzes the attribute disclosure in smart meter data. Section 5 discusses some aspects of e attribute disclosure attack, and Sect. 6 concludes the paper.

## 2 Aggregation of Smart Meter Data

We consider a generic case of smart meter data, where the data provided by a smart meter can be seen as time series. A time series is a sequence of values taken at some time interval. In our case, we will consider regular time periods, and for each time period we will have the specific power consumption reading provided by the smart meter. We denote the time series of smart meter  $i$  as  $x_i = (x_i^t : t \in T)$ , where  $T$  is the time based index.

In our case, we will consider  $T = t_0, \dots, t_m$ , for all time series in the dataset. Each time series corresponds to a different smart meter, and we consider readings for a single day, having time periods of 30 minutes, thus  $m = 48$ . In some sense the dataset could be seen as a microdata file where each record is a time series and the attributes are the consumption for each time period.

As previously stated, it is quite common to aggregate smart meter data in order to provide some degree of privacy regarding individual consumption patterns. Moreover, this aggregation could be done to provide  $k$ -anonymity guarantees in the protected data.

We can observe two different approaches to perform this aggregation:

- Global aggregation: the data publisher has the data from all smart meters and can then aggregate them at once. This allows to perform, for example, microaggregation of the time series. The advantage of this approach is that the aggregation can better preserve information loss.
- Local aggregation: in this case, the data is aggregated locally, usually by some power grid distribution center. Records are aggregated together based on e.g. geographical distribution and not the actual consumption pattern.

In this study, we will consider the global aggregation case since it is the one expected to have a better behavior regarding attribute disclosure risk. If attribute disclosure is possible in globally aggregated data, it will surely be

possible in locally aggregated data. This claim is based on the intuition that global aggregation will provide more homogeneous groups, where consumption patterns will be more similar and thus, more difficult to differentiate.

## 2.1 Datasets

We have used 3 datasets from two different sources. From each dataset we took only one day, that is, 48 readings for each smart meter. We chose the day as the one that had more readings from the dataset. The three datasets are:

- *ch1banes*: *BANES Energy Data Electricity* [5] is a dataset with electricity energy usage data in Council buildings from Bath and Nord East Somerset. The dataset shows consumption in 30 minutes slots. We have taken the consumption for 2019-11-13, which consists of 79 different buildings.
- *cer*: which contains electricity and gas consumption data from the Commission of Energy Regulation, as provided by the *Irish Social Science Data Archive* [10], also for 30 minutes slots. Here we consider:
  - *cer-elec*: electricity consumption for 2009-08-20, with consumption from 983 houses.
  - *cer-gas*: gas consumption for 2009-12-03, with consumption from 1493 houses.

We also assume that the values of each reading will be positive, since we are only considering energy consumption.

## 2.2 Smart Meter Data Microaggregation

We have considered a global microaggregation of the smart data. Microaggregation [7,8,13] is a well known method for data privacy that is commonly used to provide  $k$ -anonymity. It builds small clusters and then replaces each or the records in the cluster by the cluster center. As each cluster has at least  $k$  records, when all the records are replaced by the same cluster center,  $k$ -anonymity is satisfied.

Microaggregation is formulated as an optimization problem with specific constraints. The objective function resembles that of  $k$ -means, where cluster centers are considered, and records are assigned to the nearest cluster center. Constraints ensure that each record is assigned to exactly one cluster, with each cluster containing at least  $k$  records (and at most  $2k$ ).

When considering more than one variable, that is, multivariate microaggregation, the problem becomes NP-hard [16] so heuristic methods have to be used. MDAV [7], ch18ch1refspsTemplsps2008spsTDP2008v01n02 is one of such methods and has been extensively used in the literature. In this work we use MDAV for microaggregation. We have used values of  $k = 2, \dots, 41$ .

Moreover, to microaggregate time series we need to define a distance function to form the clusters, and an aggregator operator to compute the cluster representative. Given that the time series are aligned, we use the Euclidean distance and the average. The average is commonly used in smart meter data aggregation.

### 3 Attribute Disclosure in Aggregated Smart Data

Regarding  $k$ -anonymity, attribute disclosure is commonly associated to categorical confidential attributes. Common attacks such a homogeneity, similarity, or skewness attacks [18] are considered on the distribution of confidential attributes. To prevent such attacks there are well known proposals such as  $p$ -sensitivity [21],  $l$ -diversity [15], or  $t$ -closeness [14].

Less known are attacks on the masked numeric attributes. In a smart meter dataset, all attributes are masked, and no confidential attributes are considered. In this case, we show that some types of attacks are possible, usually considering an internal attacker.

In [19] some metrics are proposed to estimate the likelihood of this kind of attacks. These metrics are based on sensitivity rules commonly used in tabular data (see e.g. [6, 9, 12] for details). We adapt them here to measure the sensitivity of aggregated smart meter data.

We consider a given smart meter dataset  $X$  with  $n$  time series or records, with  $m$  consumption readings. We assume the data has been protected resulting in a protected dataset  $X' = \rho(X)$ , where  $\rho$  is the protection method, in our case, microaggregation.

Such dataset  $X'$  consists of clusters of record, each cluster with equal time series. In each one we can observe different cells, one for each time reading. Let us consider that for a given cell we have  $t$  contributors which provide the original values  $c_1, \dots, c_t$ . If  $\bar{c}$  is the average,  $\bar{c} = \frac{1}{t} \sum_i c_i$ , and the protected cell will have this value for all the records.

Considering the contribution of each record to the cell average, we can check the following sensitivity rules to denote if a cell is sensitive. In this context, a sensitive cell is a cell where attribute disclosure can take part.

**( $n, r$ )-dominance:** The rule  $(n, r)$ -dominance determines that the cell is sensitive when  $n$  contributors represent more than the  $r$  fraction of the total. If we consider the values  $c_i$  ordered in decreasing order,  $c_{\sigma(1)} \geq c_{\sigma(2)} \geq \dots \geq c_{\sigma(t)}$ , this rule will detect a cell as sensitive when

$$\frac{\sum_{i=1}^{nr} c_{\sigma(i)}}{\sum_{i=1}^t c_i} > r. \quad (1)$$

**$p\%$  rule:** The rule  $p\%$  is stated as follows. A cell is sensitive when an intruder can estimate the contributor within  $p$  percent, taking into account the released table. It can be proven that the best estimation is the one of the second-largest contributor (i.e., the one which contributes with  $c_{\sigma(2)}$ ) on the largest one. Then a cell is sensitive when

$$\sum_{i=3}^t c_{\sigma(i)} < p c_{\sigma(1)}. \quad (2)$$

In this expression we use  $p$  as a value in  $[0,1]$  instead of a percentage.

Hundepool et al. [12] recommend the use of  $p' = (1 - r_r)/r_r$  (and  $p\% = 100p'$ ) as providing a risk assessment similar to the  $(2, r_r)$  rule. E.g., for  $n_r = 2$  and  $r_r =$

0.6, we would have  $p = 66\%$ . In general, for the dominance rule, parameterization with  $n = 1$  or 2 and  $r > 0.6$  have been considered in the literature. For the rule  $p\%$ , a parameter larger than 60% has also been considered in the literature. We will use  $n = 1$ ,  $n = 2$ ,  $r = 0.6$ , and  $p = 66\%$  in our experiments.

## 4 Attribute Disclosure Risk in Smart Meter Data

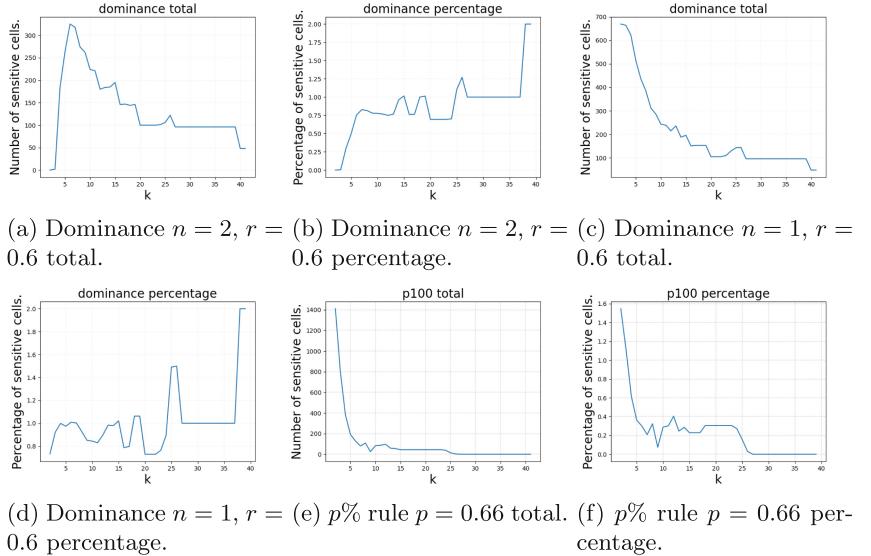
In this section, we analyze the sensitivity of smart meter data according to the sensitivity rules commented in Sect. 3. We have considered the three datasets commented in Sect. 2.1. Each dataset has been protected with microaggregation for  $k = 2, \dots, 41$  using the MDAV algorithm. Table 1 shows a summary of the number of cells for each dataset and some values of  $k$  for each dataset. We include the number of sensitive cells according to the  $(n, r)$ -dominance for  $n = 2$ ,  $n = 1$ ,  $r = 0.6$ , and the  $p\%$  rule for  $p = 66\%$  (See 3).

**Table 1.** Number of cells (*cells*), dominance for  $n = 2$ ,  $r = 0.6$  (*dom2*), and for  $n = 1$ ,  $r = 0.6$  (*dom1*), and  $p\%$  for  $p = 66\%$  (*p100*) for each dataset masked with different values of  $k$ .

<i>k</i>	<i>ch1banes</i>				<i>cer-elec</i>				<i>cer-gas</i>			
	<i>cells</i>	<i>dom2</i>	<i>dom1</i>	<i>p100</i>	<i>cells</i>	<i>dom2</i>	<i>dom1</i>	<i>p100</i>	<i>cells</i>	<i>dom2</i>	<i>dom1</i>	<i>p100</i>
2	1872	0	669	1410	23568	0	8671	23556	35808	0	4728	19306
4	912	184	623	384	11760	1426	10861	5068	17904	980	5582	9650
6	624	325	436	129	7824	5566	7683	723	11904	2283	4787	6425
8	432	274	312	109	5856	5349	5818	153	8928	2313	4187	4838
10	336	224	243	83	4704	4508	4688	62	7152	2218	3873	3789
15	240	195	196	44	3120	3097	3118	6	4752	1954	3103	2234
20	144	100	105	44	2352	2341	2352	2	3552	1722	2565	1454
30	96	96	96	0	1536	1536	1536	0	2352	1403	1920	743

More detailed results are shown in Figs. 1, 2, 3. Both sensitivity rules are show in absolute value and percentage over the total number of cells for each dataset.

In general, the dominance rule highlights more cells as sensitive than the  $p\%$  rule. Specially, the dominance rule for  $n = 1$  gives a very high percentage of sensitive cells. The attacker could estimate the highest consumer with a 60% of confidence. These results show that even masked data can have sensitive cells, yielding attribute disclosure. This means cases in which a user can have a good estimation of the consumption of the rest of the users. It is also important to note that we have employed a global masking approach. All records are protected globally resulting in a more homogeneous microaggregation, a geographical aggregation will lead to a higher number of sensitive cells.



**Fig. 1.** Dominance and  $p\%$  rule for the dataset *banes*.

## 5 Internal Attacks on Aggregated Data

In this section, we discuss the potential attack performed by an internal user attempting to estimate consumption from the masked dataset. As an example scenario, suppose that a user knows its own consumption for the day, and has the protected dataset with the consumption from all users aggregated. This user can attempt to estimate the consumption of users that fall in its own anonymity set (microcluster). To do that, the attacker needs to:

1. Identify the microcluster where its consumption has been aggregated.
2. Estimate the average consumption of the other users.

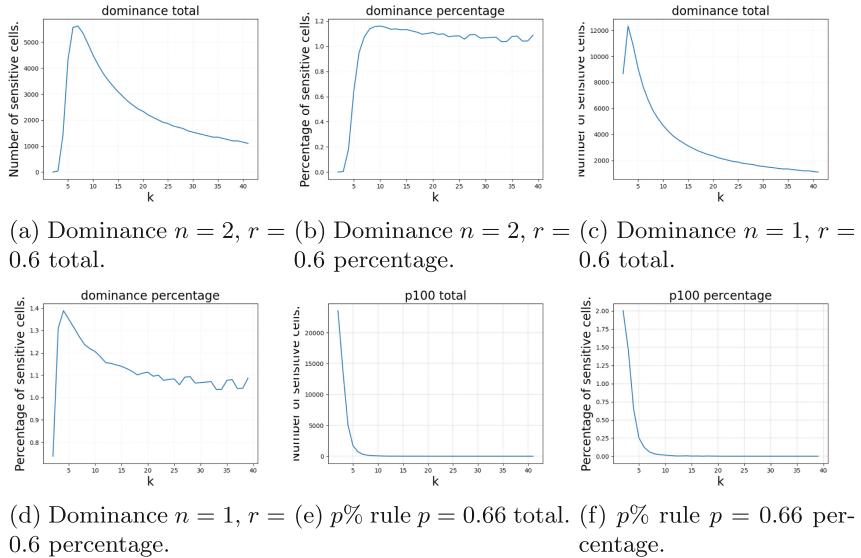
For the first step, the attacker can estimate the closest aggregated time series consumption to its own consumption and assume that it will be its own microcluster or anonymity set.

We denote the aggregate dataset  $X'$  as composed of  $s$  microclusters of time series:  $C_1, C_2, \dots, C_s$ , with their respective cluster centers  $\bar{C}_1, \bar{C}_2, \dots, \bar{C}_s$ . The attacker with a consumption time series  $x_a$  attempts to identify the cluster  $C_a$  where its data have been aggregated as:

$$C_a = \{C_i \mid \arg \min_i d(x_a, \bar{C}_i)\} \quad (3)$$

The attack is successful if  $c_a \in C_a$ . Here,  $d$  is a distance function for time series. We use the Euclidean distance (see Sect. 2.2).

Figure 4 shows the microcluster reidentification success ratio for each dataset. This has been computed by randomly selecting a record from the original dataset



**Fig. 2.** Dominance and  $p\%$  rule for the dataset *cer-elec*.

and then attempting to identify the cluster in the protected dataset according to Eq. (3). The result is the average of 100 executions for each case.

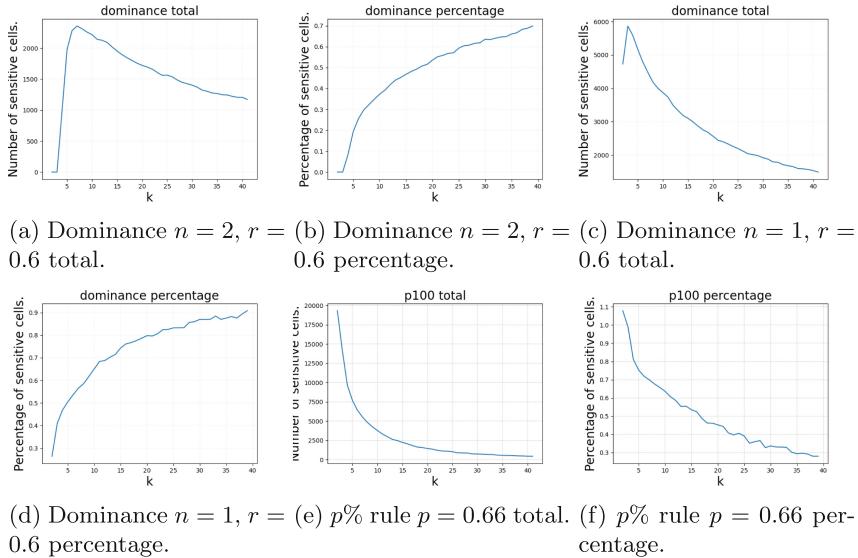
We can see that the success ratio is mostly above 50%, but in general, one could expect a bigger success ratio given the microaggregation was performed on the whole dataset. It is thus expected that a locally aggregated dataset based on e.g. geographical distribution could yield worse results.

The second step, estimating the average consumption of other users, can obviously be estimated from the published cluster representative. Given the cluster where the attacker data is masked,  $C_a$  with its representative  $\bar{C}_a$ , if we consider that the cluster has  $r$  records (time series) we could see a cell of consumption values for a given time period  $t$  as  $C_a^t = \{c_{a1}^t, c_{a2}^t, \dots, c_{ar}^t\}$  with an average of  $\bar{C}_a^t$ . The attacker can estimate the average consumption of the other  $t - 1$  users. Let  $c_{a1}$  be attacker's consumption, with this information can easily set an upper bound of the estimation. The maximum value for any  $c_{ai}$  for  $i = 2, \dots, r$  will be:

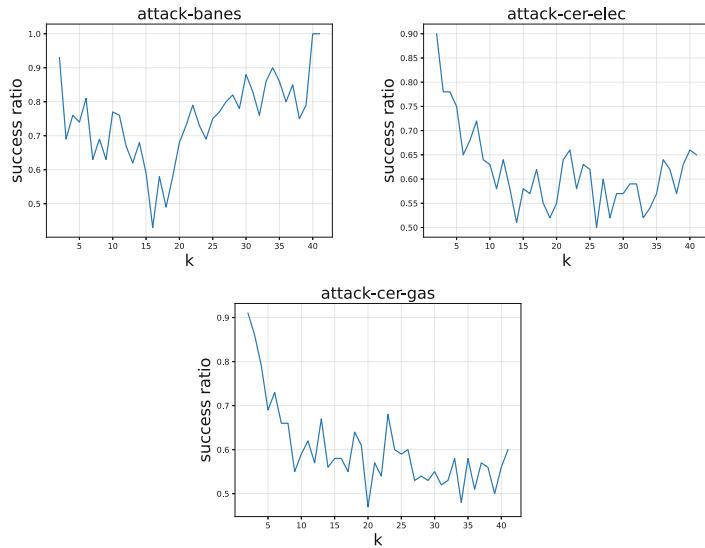
$$\max_{i=2,\dots,r} \{c_{ai}\} \leq (r \cdot \bar{C}_a^t) - c_{a1} \quad (4)$$

The minimum will be 0 for  $r > 2$ .

This estimation is somehow expected in aggregated data, but if the cell is sensitive, it means that the attribute disclosure is more critical. Prior knowledge will allow the attacker to estimate the higher consumer with more precision.



**Fig. 3.** Dominance and  $p\%$  rule for the dataset *cer-gas*.



**Fig. 4.** Microcluster reidentification success ratio for global microaggregation in the *ch1banes*, *cer-elec*, and *cer-gas* dataset.

### 5.1 Practical Implications Discussion

From a practical point of view, an attacker could force the previously described attack if its smart meter is taking part in the anonymized dataset.

We have considered the worst case for the attacker in order to give some insight on the potential problem of attribute disclosure, but the attack is more significant on locally aggregated data. In such cases, the attacker will attempt to gain information about a neighbor (probably known neighbors).

One problem with locally aggregated data is that the attacker might not easily identify its microcluster in the protected dataset. To increase the probability of succeeding in the attack, the attacker can generate a high consumption pick in a specific time period. This will affect the average of this specific time period, making the aggregate easily identifiable. Alternatively, the attacker can induce some particular patterns in the consumption to inject specific patterns (signatures) in the aggregated data. Then, using non-intrusive load monitoring (NILM) techniques, these signatures [3] can be extracted from aggregated data.

We observe that this attack is feasible on smart meter data using real data and should thus be considered for further research.

## 6 Conclusions

In this paper we have analyzed attribute disclosure in aggregated smart meter data. We have shown that there is potential risk for attribute disclosure by checking sensitivity of aggregated values in a global microaggregation. This leads to assume that locally aggregate data will introduce more risk, leading to higher possibilities of attribute disclosure.

**Acknowledgements.** This research has been partially supported by the project DANGER C062/23 within the Plan de Recuperación, Transformación y Resiliencia funded with Next Generation EU funds. Spanish Ministry under Grant PID2021-125962OB-C33 SECURING/NET. Catalan AGAUR under Grant SGR2021-00643. The Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation. Support also received from the Swedish Research Council under the project Privacy for complex data (VR 2022-04645).

## References

1. Adewole, K.S., Torra, V.: DFTMicroagg: a dual-level anonymization algorithm for smart grid data. *Int. J. Inf. Secur.* (2022). <https://doi.org/10.1007/s10207-022-00612-8>
2. Adewole, K. S., Torra, V.: Energy disaggregation risk resilience through microaggregation and discrete Fourier transform. *Inf. Sci.* **662**, 120211 (2024). <https://doi.org/10.1016/j.ins.2024.120211>
3. Adewole, K.S., Torra, V.: Privacy issues in smart grid data: from energy disaggregation to disclosure risk. *Proc. DEXA* 71–84 (2022)
4. Alsaid, M., Slay, T., Bulusu, N., Bass, R.B.: K-anonymity applied to the energy grid of things distributed energy resource management system. In: Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, pp. 581–582. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3498361.3538794>

5. BANES Energy Data Electricity. (2020) Bathhacked, Bath and North East Somerset. <https://data.bathhacked.org/datasets/banes-energy-data-electricity>
6. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. *Eur. J. Oper. Res.* **171**, 39–52 (2006)
7. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.* **14**(1), 189–201 (2002)
8. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous  $k$ -anonymity through microaggregation. *Data Min. Knowl. Disc.* **11**(2), 195–212 (2005)
9. Duncan, G.T., Elliot, M., Salazar, J.J.: Statistical Confidentiality. Springer, New York (2011). <https://doi.org/10.1007/978-1-4419-7802-8>
10. Commission for Energy Regulation (CER). (2012). CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010 [dataset]. 1st Edition. Irish Social Science Data Archive. SN: 0012-00, and 0013-00. <https://www.ucd.ie/issda/data/commissionforenergyregulationcer/>
11. Gerlitz, C., Eriksson, A., Hansson, C.: Anonymisation score for time series consumption data. In: 27th International Conference on Electricity Distribution (CIRED 2023), pp. 428–432. Institution of Engineering and Technology, Rome, Italy (2023). <https://doi.org/10.1049/icp.2023.0338>
12. Hundepool, A., et al.: Statistical Disclosure Control. Wiley, Hoboken (2012)
13. Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Trans. Knowl. Data Eng.* **17**(7), 902–911 (2005)
14. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: privacy beyond  $k$ -anonymity and  $l$ -diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, pp. 106–115 (2007). <https://doi.org/10.1109/ICDE.2007.367856>
15. Machanavajjhala, A., Gehrke, J., Kiefer, D., Venkatasubramanian, M.: L-diversity: privacy beyond  $k$ -anonymity. In: Proceedings of the IEEE ICDE (2006)
16. Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control, statistical. *J. United Nat. Econ. Comm. Eur.* **18**(4), 345–354 (2000)
17. Samarati, P.: Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001)
18. Torra, V.: Guide to Data Privacy: Models, Technologies, Solutions. Springer International Publishing, Cham (2022). <https://doi.org/10.1007/978-3-031-12837-0>
19. Torra, V., Navarro-Arribas, G.: Attribute disclosure risk for  $k$ -anonymity: the case of numerical data. *Int. J. Inf. Secur.* **22**, 2015–2024 (2023). <https://doi.org/10.1007/s10207-023-00730-x>
20. Templ, M.: Statistical disclosure control for microdata using the r-package sdcMicro. *Trans. Data Priv.* **1**(2), 67–85 (2008)
21. Truta, T.M., Vinay, B.: Privacy protection: p-sensitive  $k$ -anonymity property. In: Proceedings of the 2nd International Workshop on Privacy Data management (PDM 2006), p. 94 (2006)



# The statbarn: A New Model for Output Statistical Disclosure Control

Elizabeth Green , Felix Ritchie , and Paul White

University of the West of England Bristol, Coldharbour Lane, Bristol BS16 1QY, UK  
Felix.ritchie@uwe.ac.uk

**Abstract.** A major success for research this century has been the growth of secure facilities allowing research access to detailed sensitive personal data. This has also raised awareness of the problem of output disclosure risk, where statistics may inadvertently breach the confidentiality of data subjects, a risk that grows with the detail in the data.

Managing this risk is a concern for these secure facilities. While there is a well-established literature on the protection of frequency tables and linear aggregates, researchers in secure facilities produce a wide range of statistical outputs. The theory covering non-tabular outputs is small, fractured, and has grown ad hoc. This is also reflected in the guidance available to data service staff, which typically consists of a long list of outputs and some rules to be applied to them.

This paper describes a significant new concept in output statistical disclosure control: the statistical barn or ‘statbarn’. This is a framework to classify all statistical terms by their disclosure characteristics, including risk, exceptions and mitigation measures. This statbarn massively reduces the dimensionality of the disclosure checking problem, as well as providing improved clarity. It also creates a feasible basis for automatic disclosure control checking.

**Keywords:** statbarn · output checking · statistical disclosure control

## 1 Introduction

Analysts have used sensitive record-level data for research since the scientific method became formalised in the nineteenth century. Since the development of the computer, both the amount of data collected and the ability to analyse it has grown exponentially. But alongside this has grown the awareness that the use of this data can infringe rights of the data subjects to confidentiality and privacy. For most of scientific history this was dealt with by a ‘good chaps’ approach to ethics, wherein scientists (and they were almost certainly chaps) were assumed to be moral and trustworthy individuals. In the late twentieth century this paternalism began to be replaced by data protection models, where the risks of sharing confidential data were managed by reducing the detail in the data. More risky data was made accessible by licensing users, but data reduction was still the first defence. However, national statistical institutes (NSIs) became aware that the tabulations that were their core activities could inadvertently reveal confidential information, and developed tabular protection methods from the 1980s onwards.

A major success for research this century has been the growth of secure facilities allowing research access to detailed sensitive personal data. These ‘trusted research environments’ (also called research data centres, safe havens, or secure enclaves or similar terms) allow researchers access to very detailed and sensitive data, but in an environment controlled by the data holders. With remote access increasingly being offered by these facilities, the TRE provides a very secure but very flexible research environment.

One part of the ‘trust’ in a TRE is knowing what has been taken out of the environment, and so all TREs use some form of output checking to ensure inappropriate material is not released. This in turn highlighted the inadequacy of tabular protection methods developed for NSIs, and so from the early 2000s output statistical disclosure control (OSDC) began to be developed.

A problem for TREs is the wide range of statistical outputs generated by researchers. Whilst tabular outputs have a large and well-established literature, the theory covering non-tabular outputs is small, fractured, and has grown ad hoc. This is also reflected in the guidance available to data service staff, which typically consists of a long list of outputs and some rules to be applied to them. There is therefore a large and growing need for a change of thinking on how best to deal with the complexity of research outputs.

This paper describes a significant new concept in OSDC: the statistical barn or ‘statbarn’. This is a framework to classify all statistical terms by their disclosure characteristics, including risk, exceptions and mitigation measures. This statbarn massively reduces the dimensionality of the disclosure checking problem, as well as providing improved clarity. It also creates a feasible basis for automatic disclosure control checking, as in the SACRO tool (Smith et al. 2023) designed to provide this capability to all TREs.

The first draft of the manual describing statbarns [1] was released for comments in October 2023, and the concept has been used in teaching both researchers and output checkers since then, with strongly positive results. In this paper, we review the historical background, describe the conceptual framework and the process of identifying statbarns, illustrate the guidance that this leads to, and provide examples of how the re-examination of the statistical foundations of OSDC has led to some new perspectives. Finally, we consider how this new framing will affect output checking in the future, particularly in the context of automated checking.

## 2 The Historical Development of OSDC Theory

As noted in the introduction, OSDC is a relatively new concept in statistics. Statistical disclosure control mainly began in the 1970s, and focused on protecting confidentiality in the source data – what would now be called ‘input SDC’. This then developed into extensive guidelines for the production of tabular outputs (frequencies, and magnitudes such as mean and total). Research on input SDC and tabular outputs was strongly supported by NSIs, as their core work was to produce public statistics (and sometimes datasets for research) from data which was often confidential at the point of collection. Texts such as [2] in 1996 were comprehensive and non-controversial, and indeed still are in the covered areas – although more is known about risks, the evaluation of risk, and counter-measures, the core theory remains largely unchanged.

However, the increasing use of highly sensitive data for research use prompted data services to consider whether there were gaps in the knowledge of risks in output. A series of papers 2004–2006 [3–7] were the first papers or guides considering non-tabular outputs, all addressing the hitherto-unconsidered issue of linear regression outputs.

A more comprehensive approach was first presented in 2007 [8] which discussed the limitations of traditional rules-based approaches for research outputs. [8] also introduced the idea of the ‘research zoo’: designing disclosure control regimes should be like a zoo manager working out which are the animals that bite, which that fly, which that burrow, and building cages to deal with each group. [9] expanded on this, arguing that mathematical form not statistical use should be the basis for classification. This led to the concept of ‘safe statistics’ and ‘unsafe statistics’: a binary grouping of statistics by mathematical form into high and low risk outputs.

These works prompted Eurostat in 2008 to commission research into OSDC. The resulting report [10] was the first attempt to provide a general-purpose guide to output checkers covering common research outputs and using these systemic models. [10] was incorporated into the Eurostat-funded guide to good SDC practice [11] and these (and an updated version [12]) have been the key references in SDC ever since.

Many TREs now produce guides for researchers. In addition, in the UK the Secure Data Access Professionals (SDAP) expert group produced a significantly more comprehensive guide [13] which built on [10] but expanded both the range of statistics covered and the use of examples to illustrate topics. The SDAP manual is currently the main reference used in UK TREs.

However, while there is more information about how to deal with a wider range of outputs, these guides still follow the pattern set by [10]: identify a statistic, describe the risks and mitigations, move on to the next one. This is not sustainable. First, there are very many statistics. Second, different disciplines use statistics in different ways. Do all the different uses need to be considered? And do output checkers need to be familiar with all the different potential uses?

The answers are no, and no; but they require a fundamental rethink of OSDC. This is the statbarn.

### 3 Theoretical Development

#### 3.1 The Building Block: Safe and Unsafe Statistics

The building block is the concept of ‘safe/unsafe statistics’, first defined in [9]. Certain statistics are ‘unsafe’, in that they present clear risks in their expected uses. For example, consider a frequency table. The risks associated with frequency tables are

- low numbers, with cells relating to one or two data subjects
- class disclosure, with empty or full cells indicating that the members associated with that cell do or do not have some specific characteristic
- differencing across tables, allowing implicit tables with low numbers to be generated

These can all be reasonably expected in the regular course of research. This is of course a general statement; a specific frequency table may be fine for release. Whether it

can be released or not depends on the data, and is specific to that particular output, not the statistic or even the dataset. A frequency table of the demographics of data subjects may be acceptable for release; but a table which describes the subset of graduates may not be suitable for release, either because of inherent problems, or because of differencing concerns when compared with the whole table. So an ‘unsafe’ statistic is data dependent and can be released but only once it has been inspected.

In contrast, a ‘safe statistic’ has no meaningful disclosure risk in normal use. Consider the coefficients from a linear regression. As [14] demonstrates, there is no disclosure risk in regular use. There are theoretical risks, but those are not meaningful in practical research environments: they require the researcher to do something nonsensical for research purposes specifically to hide confidential data, and output checking is based on the assumption that the researcher is not deliberately trying to cheat the system.

So regression coefficients are ‘safe’ and can be published irrespective of the underlying data. This does not mean that there are no rules around them. It is possible (if unlikely) that an inexperienced researcher could inadvertently generate a magnitude table by running a linear regression on dummy variables and including all possible interactive terms (fully saturated model). In practice, this is only likely to occur if a very naïve researcher runs a regression with just one or two binary variables as explanators. The SACRO manual [1] therefore requires that a fully saturated regression should be treated as a magnitude table. Note however that that rule does not require any assessment of the data: it is the functional form that is being checked (is this really a regression, or just a table?).

The safe/unsafe distinction is based on the mathematical, not the statistical, characteristics, and these classifications are inherited by more complex forms. A linear regression is safe. A hazard or panel model is even safer, but the safe/unsafe model does not have this level of granularity. A frequency table is unsafe. Weighted frequencies, rounded values or survival tables are all safer than a plain frequency table, but they are still unsafe and need to be checked.

So the safe/unsafe statistics model simplifies the output checking problem by putting outputs into two large boxes marked “check” or “don’t check”. This in itself has been of substantial value to TREs, but the natural extension is to consider whether a more detailed classification model would be even more effective without losing the benefits of simplicity?

### 3.2 The statbarns

A statistical barn (‘statbarn’) defines a group of statistics in terms of four characteristics:

- Classification as safe/unsafe
- Risk factors
- Rules to be checked
- Mitigation measures

All statistics in the statbarn should have the same characteristics on those four criteria. That means that the output checker only needs to know one set of risks to consider, one set

of rules to be applied, and one set of mitigation measures to recommend to researchers – irrespective of the number of statistics in that group. Consider the statbarn labelled ‘frequencies’. This includes

- frequency tables
- pie chart, waterfall charts, histograms
- reported counts
- decision trees

For all of these statistics, and any others in the statbarn, each of these statements applies:

- they are unsafe and need to be checked before release
- low numbers, class disclosure, class exclusion and differencing are all potential sources of risk
- a minimum threshold is an appropriate rule to be applied
- suppression, rounding, noise addition and table redesign are all appropriate mitigation techniques

The output checker therefore now has a clear model for dealing with a very large range of outputs. Some of these may require repeat application, as in the case of decision trees where each node has to be evaluated, but the principles remain the same.

Consider now the statbarn ‘correlation coefficients’, which includes all linear and non-linear estimation models. The corresponding characteristics of these are:

- Safe, no checks before release except administrative ones (see below)
- No meaningful things to check expect making sure it is a regression
- If fully saturated, evaluate as table; check for residual degrees of freedom to prevent a naïve researcher running a model with  $n$  observations and  $n-1$  variables (i.e. a faithful reproduction of the data in equation format)
- There are no meaningful mitigations as an output only fails if it is not a regression

Again, one set of guides/rules covers a very large set of outputs. At the time of writing, twelve statbarns have been identified, and given risks, rules and mitigations [1]:

**Table 1.** Current defined statbarns (from [1])

‘Unsafe’ statbarns	‘Safe’ statbarns
a) Frequencies	g) Correlation coefficients and models
b) Position (median, percentiles, IQR)	h) Statistical hypothesis tests
c) Maxima and minima	i) Shape (SD, skewness, kurtosis)
d) Means, total, other linear aggregates	j) Mode
e) Calculated odds/risk ratios	k) Non-linear concentration ratios
f) Hazard/survival tables	l) Gini coefficient/Lorenz curves

Two other statbarns (cluster analysis and linked multilevel tables) are still to be explored, but clearly the statbarns listed in Table 1 cover almost everything an output checker is likely to encounter, in just twelve groups.

Some of the statbarns are very large (a, d, g, h) whereas others just have one or two statistics in them (c, e, I, j, l). This reflects important differences in characteristics. For example, maxima and minima could be considered as special cases of ‘position’ (and they were originally) but it was decided that there need to be different ways of handling these two groups. Similarly, ‘frequencies’ could be seen as a subset of ‘linear aggregates’, because all the ‘frequencies’ rules apply to both, but linear aggregates also have dominance rules. So one could remove the ‘frequencies’ statbarn and have an extra rule in ‘linear aggregates’ (“ignore dominance rules if dealing with frequencies”) but this seems more confusing than having two clearly recognisable categories. Finally, hazard and survival tables are usually treated as a form of frequency table (e.g. in the SDAP manual, Greci et al. 2019), but the examination of their form showed that their structure, as far as disclosure risk is concerned, is very different.

Allocation of statistics to groups can be unclear, particularly where different forms of something nominally similar is placed into different statbarns. For example, linear concentration ratios (“what is the market share of the top two companies?”) go in the unsafe linear aggregates statbarn; but non-linear concentration ratios (such a Herfindahl index) are safe and have their own category. As another example, odds or risk ratios calculated directly from incidence numbers are unsafe, but estimated odds ratios from logistic regression are safe (providing they include additional explanatory variables and are not a simple recast of a frequency table) and count as correlation coefficients.

The classification is not entirely based on the characteristics of the statistic. For example, calculated odds ratios pose no disclosure risk in themselves; but it is almost certain that they would be published alongside counts (such as numbers in treatment and control groups) that allow a non-negligible risk to emerge. Hence, the guidance for calculated odds ratios is that they should not be assessed but the underlying 2x2 frequency table should also be provided to the output checkers, who will then assess as usual for frequencies.

The statbarns model addresses the two major concerns about the sustainability of output checking: the number of statistics needing rules, and inter-disciplinary variation in use. The first is directly addressed: while the number of statistics is large, from a disclosure control perspective there are very few different types of statistics. As an exercise, all the statistics available in SPSS were extracted from the SPSS manual, and all could be allocated to one of the twelve statbarns.

In terms of developing inter-disciplinary guidelines, the statbarns model simplifies this by making clear when inter-disciplinary differences matter. For safe statistics, there is no difference in disciplines, because this is based on mathematical form, not use. For unsafe statistics, expertise in different disciplines may come into play, because the data needs to be assessed if the rules are not being followed and an exception is requested. However, the rules themselves are cross-disciplinary: a frequency table should be assessed against a minimum threshold, whether this is for health research or economics. Disciplines may choose to set different thresholds depending on the sensitivity of their data but the principle remains the same. This is particular useful as it allows cross-disciplinary training to be developed.

## 4 Implementation

The current position of statbarns is taken from the SACRO manual [1]. For each of the twelve listed barns, [1] provides

- A description
- Risk factors
- Safe/unsafe classification
- Criteria for rules-based approval, for both manual checking and for machine-led checking
- Remedial actions/acceptable mitigations
- Issues to consider if an exception is requested
- And a discussion of the underlying theory

Note that the rules may differ for humans and computers, based on their abilities: humans cannot do dominance checks effectively; computers don't understand structural zeroes, for example. However, this is the first time a comprehensive set of rules for both humans and computers have been set out. This enables TREs to check their current practices against a clearly stated standard.

The rules in [1] have also been directly implemented in SACRO Python tool [15], which provides semi-automated output checking and is planned for adoption by all UK TREs. Not all of the groups have been implemented yet, but for those which are implemented there is now a direct and transparent link between theory and operating practice.

Since October 2023, the authors have also been using the statbarns concept in training, of both researcher and output checkers. This has received a universally positive response, as it simplified the previous material which was based on the ‘show lots of different types of output’ approach.

## 5 Limitations and Areas for Further Work

[1] provides a theoretical foundation for the statbarns models, but the theoretical basis for this varies. For example

- Frequencies and linear aggregates have been studied for thirty years and have a large evidence base on risks, rules and mitigation
- Linear regression is understood well enough for the rules to also apply to non-linear models, but issues with pairwise correlations have not been studied in detail
- The rationale for the mode as ‘safe’ is published for the first time as an appendix
- New arguments on how survival tables should be assessed, which differ considerably from previous guidelines, form another appendix

Community responses to the initial specification in [1] suggest that it is at least robust enough for practical use. More detailed analysis might suggest that some of the newer statbarns need more theoretical analysis. The number of statbarns may change. For example, it has been argued that, as some parametric statistical hypothesis tests could be reframed as regressions, they should not be a separate category. However, these

developments relate to the specification of initial statbarns, not the concept itself, which has received widespread support.

One gap in the usability of the framework is linking specific statistics to the statbarn. The SACRO team produced an initial mapping as an online spreadsheet, circulated with the draft manual for comments. In the short term this works as most output checkers and researchers tend to work with the same sort of statistics repeatedly, and so can learn about their OSDC properties once. In the longer term, this needs to be more accessible, including search function and synonyms (for example, ‘panel data’ versus ‘repeated measurement’).

The second major gap is graphs and plots. As noted above, these should be able to be placed in one of the statbarns as assessed as such; for example, a scatter plot is a (sparse) frequency table; a kernel density plot is a regression model. Currently the authors have generated an initial list but there are a wider range of uncertainties about which statbarns some plots should go into: is a heat map a type of frequency table, or a linear aggregation? Or is a wholly new statbarn needed? We suspect that the demands for new statbarns may come from graphical outputs.

We have not addressed mitigation solutions, such as differential privacy (DP) or cell-key adjustment (CKA). These are outside the scope of the statbarns project, as the focus is on identifying the minimum grouping rather than considering the whole range of remediation options. [1] merely notes that, for example, ‘noise addition’ (which covers both DP and CKA) is an appropriate solution for frequency tables.

We have also not directly addressed the disclosure risk presented by machine-learning (ML) models. As [16] note, such models cause novel risks because of their potential to store data in the models and to saturate the regression, as well as raising concern over what ‘disclosure’ means in the ML world. At present, the research focus is on understanding what the risks are from particular ML processes [17], and the statbarns model helps to identify different classes of ML models. But this is an ongoing area of research and likely to expand considerably in future.

## 6 Conclusion

The statbarns model changes the way researchers and output checkers think about the disclosure risk in different types of outputs. It is a natural extension of the safe/unsafe statistics model, but in a way which is both more specific and more general. It provides a much more manageable set of guidelines, rules and mitigation measures than previous framings. It can provide the necessary simplification to allow semi-automatic disclosure control tools to be developed, such as the SACRO Python tool which is being adopted across UK TEs. Overall, the statbarns model is one of the most significant developments in OSDC since 2010, and the rapid adoption in the UK is evidence of the need for this streamlined framing.

There is still a substantial amount of work to be done. Some of this is practical, providing the necessary guidance and search facilities to allow researchers and output checkers to identify quickly the statbarns associated with their statistics of interest. TEs may also want to review their training programmes. There are also likely to be theoretical gaps, which might change the number of statbarns or the guidance. Finally, there may

be a need to integrate other developments in OSDC, such as the new noise-addition measures which have become popular in the last decade. Nevertheless, we expect the fundamental concept to be unchanged, and to become increasingly central to OSDC management.

**Acknowledgments.** The authors are grateful for comments on earlier drafts of the SACRO manual, and to participants in workshops. The development of the manual was funded by the UK Medical Research Council as part of the SACRO project, which itself built on the Eurostat ACRO project.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. SACRO: SACRO guide to statistical output checking (2003). <https://uwe-repository.worktribe.com/output/11858423>
2. Willenborg, L., De Waal, T.: Statistical Disclosure Control in Practice, vol. 111. Springer, Cham (1996)
3. Reznek, A.: Disclosure risks in cross-section regression models. Mimeo, Center for Economic Studies, US Bureau of the Census, Washington (2004)
4. Reznek, A., Riggs, T.: Disclosure risks in releasing output based on regression residuals. In: Proceedings of the Section on Government Statistics and Section on Social Statistics, ASA 2004, pp. 1397–1404 (2005)
5. Corscadden, L., Enright, J., Khoo, J., Krsinich, F., McDonald, S., Zeng, I.: Disclosure assessment of analytical outputs. Mimeo, Statistics New Zealand, Wellington (2006)
6. Gomatam, S., Karr, A., Reiter, P., Sanil, A.: Data dissemination and disclosure limitation in a world without microdata: a risk–utility framework for remote access analysis servers. *Stat. Sci.* **20**(2), 163–177 (2005)
7. Ritchie, F.: Disclosure control for regression outputs. Mimeo, Office for National Statistics, Newport (2006)
8. Ritchie, F. Statistical Detection and Disclosure Control in a Research Environment. Mimeo, Office for National Statistics, June. WISERD Data Paper no. 6 (2011). <https://uwe-repository.worktribe.com/output/957311> (2007)
9. Ritchie, F.: Disclosure detection in research environments in practice. Paper presented at UNECE/Eurostat work session on statistical data confidentiality - 2007, Manchester, United Kingdom (2008)
10. Brandt, M., et al.: Guidelines for the checking of output based on microdata research. Final Report of ESSnet Sub-group on Output SDC (2010). <https://uwe-repository.worktribe.com/output/983615>
11. Hundepool, A., et al.: Handbook on Statistical Disclosure Control. ESSNet SDC (2019). [http://neon.vb.cbs.nl/casc/SDC\\_Handbook.pdf](http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf)
12. Bond, S., Brandt, M., de Wolf, P.-P.: Guidelines for Output Checking. Eurostat (2015). [https://ec.europa.eu/eurostat/cros/system/files/dwb\\_standalone-document\\_output-checking-guidelines.pdf](https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf)
13. Griffiths, E., et al.: SDAP Handbook on Disclosure Control for Outputs (2019). <https://doi.org/10.6084/m9.figshare.9958520>

14. Ritchie, F.: Analyzing the disclosure risk of regression coefficients. *Trans. Data Priv.* **12**(2), 145–173 (2019). <http://www.tdp.cat/issues16/abs.a303a18.php>
15. Smith, J., et al.: SACRO: semi-automated checking of research outputs. In: UNECE Expert Meeting on Statistical Data Confidentiality 2023, Wiesbaden, Germany (2023). <https://uwe-repository.worktribe.com/output/11060964>
16. Ritchie, F., et al.: Learning models in trusted research environments - understanding operational risks. *Int. J. Popul. Data Sci.* **8**(1), Article 2165 (2023). <https://doi.org/10.23889/ijpds.v8i1.2165>
17. Mansouri-Bensassi, E., et al.: Disclosure control of machine learning models from trusted research environments (TRE): new challenges and opportunities. *Heliyon* **9**(4), Article e15143 (2023). <https://doi.org/10.1016/j.heliyon.2023.e15143>

# **Spatial and Georeferenced Data**



# Masking Georeferenced Health Data - An Analysis Taking the Example of Partially Synthetic Data on Sleep Disorder

Simon Cremer<sup>1</sup>(✉), Lydia Jehmlich<sup>1</sup>, and Rainer Lenz<sup>1,2</sup>

<sup>1</sup> Institute for Production, Cologne University of Technology, Arts and Sciences,  
50679 Cologne, Germany

{simon.cremer, lydia.jehmlich, rainer.lenz}@th-koeln.de

<sup>2</sup> Department of Statistics, Technical University of Dortmund, 44221 Dortmund, Germany

**Abstract.** Spatial health data is becoming increasingly important in health research. However, the desired information can often not be extracted despite the inherent analytical content. The reason for this is that access to personal georeferenced data sets is severely restricted as they are subject to legal data protection. The method of donut masking attempts to alienate original data by shifting it in such a way that data protection is guaranteed without strongly reducing the analytical validity of the data. In this article donut masking is applied to partially synthetic data on sleep disorders. The degree of anonymity of the masked data set is measured by spatial  $k$ -anonymity reviewing additional knowledge of a potential data attacker. In addition to assessing the spatial similarity of the original and masked data set, an attempt is also made to assess the suitability of such data for analysis purposes.

**Keywords:** Privacy-enhancing technologies · geoprivacy · location privacy · medical informatics · spatial analysis · donut masking · geographic health · spatial  $k$ -anonymity

## 1 Introduction

Whether for the use of targeted advertising measures or the detection of the spatial spread of viruses such as the coronavirus, geodata can - depending on its characteristics - offer enormous added value for society, science and research. Important questions about the future and sustainability of our society can only be answered with high-quality and accessible geodata. For these reasons, the German Council for Social and Economic Data set up a working group on the topic of “Georeferencing of data” back in mid-2010. Since then, great efforts have been made not only in Germany (including the Federal Agency for Cartography and Geodesy and the state surveying offices), but also throughout the European Union to improve the European geodata infrastructure.

The anonymization of health data is a major challenge due to the high degree of individuality of personal and household data. According to the GDPR (Art. 9), this data is considered particularly sensitive and worthy of protection. Most publications on

the anonymization of geodata therefore deal with health data [1]. To identify confidential information, a potential data attacker can, for example, use the geocoordinates of the patient's registration address. However, not only numerical tables, but also characteristics such as image data (X-rays, CT scans), names of the treating physicians or medication administered can contribute to the identification of individuals. The rapid development of data availability and methodology in conjunction with growing computing power is increasingly challenging traditional methods of anonymization and statistical confidentiality. Many anonymization methods are based on the scenario of the potential data attacker who has additional knowledge in the form of quasi-identifiers (i.e., common attributes of additional knowledge and confidential target data) and can thus identify sensitive information about respondents [2]. When using anonymization methods, a bicriteria optimization problem usually must be solved. On the one hand, the analysis potential of the data should be retained, on the other hand, the data should be protected as well as possible. In practice, the problem is usually converted into a single-objective optimization problem with constraints: Minimizing the loss of information while setting a threshold for the protection of the data. Or vice versa: minimizing the risk of re-identification by specifying a catalog of possible analyses. Today, with the digital revolution and new digital data and algorithms that come with it, many familiar questions are being asked anew and many new questions have been added.

The aim of this study is to contribute to a better understanding of the impact of what is called geomasking in practice by evaluating the trade-off between privacy and utility for differently perturbed synthetic microdata. We apply different parameter settings of the donut masking method to georeferenced data in combination with traditional recoding and evaluate the masked data in terms of anonymity and quality.

## 2 The Donut Masking Method

Geomasking is a pool of data processing tools that serve to modify individual data in such a way that as much geoinformation as possible is preserved, while the correct linking of geographical information with confidential target data is made more difficult or even impossible. In this respect, geomasking attempts to optimize the conflicting goals of anonymity and quality.

A prominent and very intuitive concept is the so-called donut method [3]. In addition to a circular space into which the point is randomly shifted, an inner ring is defined around the original location, beyond which the shift must extend. Thus, two radii, which resemble the shape of a donut, determine the minimum and maximum distance between which the new coordinates of the point lie [3].

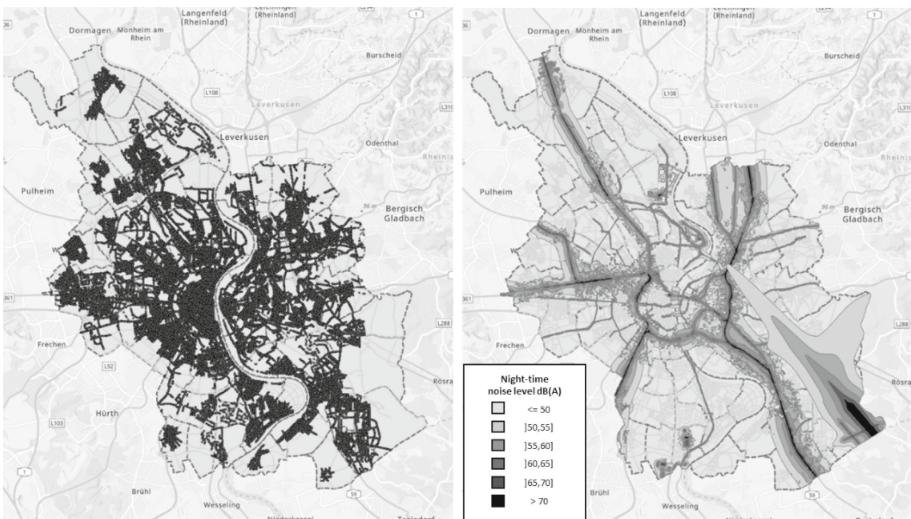
Several other geomasking methods have been developed in addition to donut masking. Nevertheless, donut masking has taken a firm place in scientific literature on geomasking. It is often referenced, further developed, or used as a comparative method. Despite its dominant presence, the method has weaknesses which mean that it has not yet established itself in practice beyond theory.

### 3 Design and Development of the Use Case

To test appropriate variants of donut masking, a partially synthetic georeferenced micro-data set was generated from Cologne administrative data. In this section, we first briefly present the underlying data basis. This is followed by a description of the implementation of methodology and measures chosen to assess both data protection and data utility.

#### 3.1 Cologne Micro Geographical Data Set

The data set used represents a sample of 10% of Cologne's actual population. This sample is based on the data available in [4], describing how many people of which gender and age live in which neighborhood. Based on this information, 109210 data points were generated, to which the features of age, gender and neighborhood were assigned. Whereas the actual distribution within the neighborhoods was considered. The points were then placed on street edges within the respective neighborhoods (the corresponding shapefile can be found in [5]). Using additional cadastre information from [6], it was ensured that no points were placed in green areas, rivers, on bridges or industrial areas to obtain a plausible distribution of the Cologne population (Fig. 1a).



**Fig. 1.** a) Distribution of data points b) Night-time noise level for Cologne.

Representing the night-time noise pollution, measured in dB(A), we used the recently published GeoPackage ‘Raumbezugssystem’ of the City of Cologne containing six what are called night-time noise levels [7]. Including particularly noise pollution from public transport, the DB rail network (Deutsche Bahn), proximity to the flight path of Cologne-Bonn Airport and industrial and harbor areas. An illustration is given in Fig. 1b. The unit dB(A) measures noise level according to the internationally standardized frequency

weighting curve A. This curve takes into account the fact that the human ear perceives sounds with different frequencies to different degrees.

A Python code was then generated to check which night-time noise pollution a data point is exposed to. Depending on the noise pollution, a point was then assigned a probability of suffering from a sleep disorder. The authors are aware that this approach is not suitable for analyzing an actual sleep disorder. However, the test data set should contain a sensitive characteristic (sleep disturbance) that is dependent on spatial conditions (night-time noise level) to examine the extent to which this relationship retained after the application of geomasking methods.

For testing geomasking methodology and analyzing the impact of different characteristics on potential sleep disorder, the five attributes listed below are available, where AC is an abbreviation for ‘Age Category’:

- *Geocoordinates* of residential address [WGS84], according to the World Geodetic System 1984
- *AC1* [0–2, 3–5, 6–14, 15–17, 18–20, 21–34, 35–59, 60–64, 65–74, 75–79, >79]
- *Gender* [male, female]
- *Night-time noise level* ( $\leq 50$ , ]50,55], ]55,60], ]60,65], ]65,70], >70)
- *Sleep disorder* [yes, no]

### 3.2 Methods of Data Anonymization

By recoding the age categories, information is further reduced, and thus additional data protection is achieved in a simple way. The categories of the further recoded variable AC2 are as follows: [0–20, 21–34, 35–59, >59]. Here, attention was paid to a meaningful recoding with focus on the intended analysis (investigation of sleep disorders). This was done in the sense that the classes roughly represent different life stages of people. Anonymization of the gender attribute, for example swapping by defining appropriate transition probabilities, was dispensed with. This means that this attribute is either assumed to be available to a potential data attacker as a quasi-identifier or not.

We now turn to the donut masking method described in Sect. 2 for perturbing the geographical coordinates. Six variants of this method are applied, with the inner radius varying between 100 m and 300 m and the outer between 200 m and 400 m. Specifically, the following settings are selected for the two radii  $\{r_{\min}; r_{\max}\}$ : {100;200}, {100;300}, {100;400}, {200;300}, {200;400} and {300;400}. In total, 18 variants of anonymization are considered, namely:

- 6 parameter settings of *donut masking* (removal of age and gender)
- 6 combinations of *donut masking x AC1 x Gender*
- 6 combinations of *donut masking x AC2 x Gender*

### 3.3 Remarks on Implementation

In this study, Python and ArcGIS were used as primary tools for data processing and analysis. Python served as scripting language for data manipulation, while ArcGIS was primarily used for visualization and analysis of anonymized data on trial as well as for the validation of implemented routines. The integration of Python into the ArcGIS

environment made it possible to effectively realise the required analytical processes, as the Python API of ArcGIS allows direct control of the GIS functionalities.

A specific field of work in this study was the application of donut masking. Although existing open-source implementations already existed for this method [8], the application of these codes revealed several running errors that could compromise data integrity. Given the inconsistencies and errors encountered in the available open-source codes, it was decided for reprogramming of the methods. The newly developed codes were then created with the aim of achieving higher reliability and compatibility with the generated data set and analysis requirements of the study. Quality assurance and debugging of the newly implemented scripts were an integral part of the process to ensure the validity of the results.

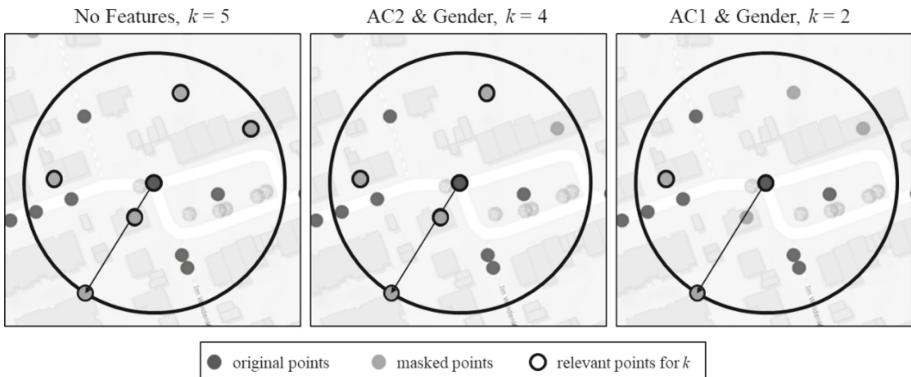
### 3.4 Measurement of Data Anonymity

To re-identify respondents of a survey, a data intruder needs additional knowledge about the units searched for (e.g., in the form of an external database) containing quasi-identifiers which the external and the confidential data have in common. Moreover, he needs response knowledge, that is, knowledge about the participation of the units in the target survey.

The concept of  $k$ -anonymity provides information on the degree to which sensitive information can be revealed by combining those variables which might occur as quasi-identifiers within the external database of a potential data attacker. A data set is called  $k$ -anonymous if each object it contains cannot be distinguished from  $k-1$  other objects, for a definition see e.g. [9]. Specifically,  $k$ -anonymity thus denotes “the number of households from whom a de-identified subject cannot be distinguished” [10] (p.4).

The  $k$ -anonymity can also be transferred to spatial data sets. This is then referred to as spatial  $k$ -anonymity. Spatial  $k$ -anonymity is the most widely used metric for measuring the degree of anonymity of masked sensitive geodata [11]. It describes the number of masked points that are closer or equally close to the original location as the masked original point itself [11]. Thus, given spatial  $k$ -anonymity, at least  $k-1$  masked points are closer to the original address than the associated masked point [12]. Hence, if no further information is available to a potential data attacker, the probability of revealing the original location is  $1/k$ .

As shown in Fig. 2, the value of  $k$  is 5 if no other features are taken into account. In this case all masked data points within the circle are relevant for the calculation of  $k$ . The black-colored rings show which other masked points are considered for calculation of the current  $k$ -value. The more features and the more differentiated these are, the fewer points stay in the neighborhood and can therefore guarantee anonymity. If the data holder provides information on age and gender, the value of  $k$  decreases from 5 to 4 (*AC2* and *Gender*) or to 2 (*AC1* and *Gender*), respectively. In the situation of a data attack, attributes on age and gender can be interpreted as quasi-identifiers, while in the present paper the *Geocoordinates* are interpreted as a sensitive attribute, as it can often identify an individual.



**Fig. 2.** Spatial  $k$ -anonymity

### 3.5 Measurement of Data Quality

Anonymization or ensuring the confidentiality of individual data always goes hand in hand with a reduction of information. Either information is suppressed to such an extent that it can no longer be assigned, or only with disproportionate effort, or protection is achieved by perturbing the data such that the usefulness of the data is (partially) lost. In either case, this means a loss of information. If a specific research objective is already being pursued with the data to be anonymized, the anonymization procedure can be specifically tailored to this aim. In the case of standardized data products, however, criteria must first be formulated to preserve the quality of georeferenced data for the largest possible target group of data users.

Regardless of the actual analysis potential many studies attempt to map the preservation of spatial quality. Spatial quality is to be understood as the topography of the masked point file is as similar as possible to that of the original one. It is exceedingly difficult to capture all topographical aspects. One aspect of approaching that topography or spatial quality is to record the position of the points in relation to each other. The ArcGIS tool “Average Nearest Neighbor” [13] is used to describe this. This function indicates the observed average distance to nearest neighbors as well as the tendency of the distribution of the points to cluster. The assessment of cluster formation is also an aspect to be considered when describing the occurrence of points, since it is useful if hotspots are retained after masking. In this tool, clustering is described by the ‘nearest neighbor ratio’ value. The index considers the relationship between the observed and the expected mean distance. The expected mean distance represents the average distance between neighbors in a hypothetical random distribution. The smallest enclosing rectangle that includes all data points is used as the area for the calculation. If the index is less than 1, the points tend to cluster. If it is greater than 1, the trend is towards dispersion.

In many scientific studies, the spatial similarity of data sets described above is used to assess data quality. Delmelle et al. point out that “a substantive number of studies do not report uncertainty levels of the spatial data and its potential impact on their analytical results” [14] (p.26). This makes analyses based on these data susceptible to bias. Incorrectly drawn conclusions are “followed by wrong decision-making of local

health policies which target specific geographical areas” [14]. Therefore, in addition to examining the spatial similarity of points before and after masking, this paper will also examine the extent to which the analytical validity of the data for is maintained. The aim is to assess whether the researchers would come to comparable conclusions with the masked data as with the original data.

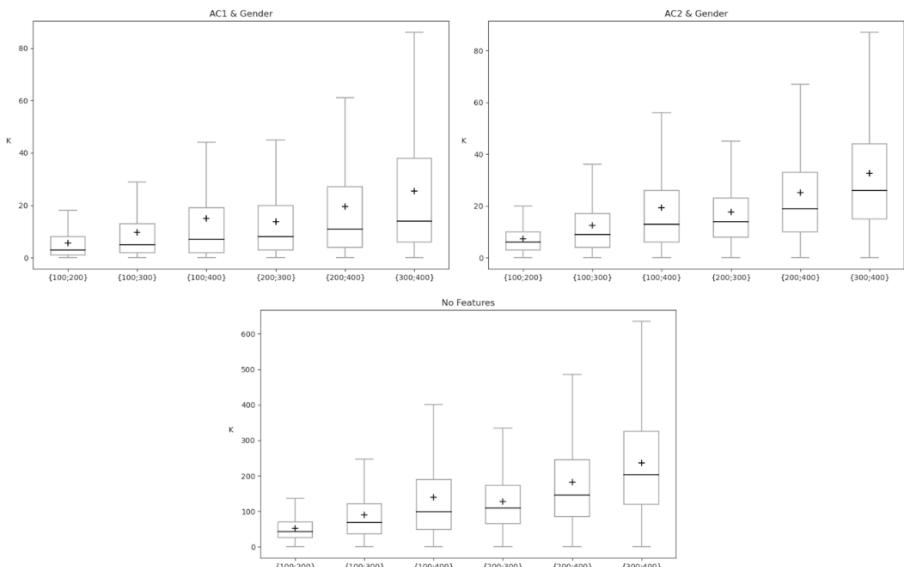
To assess the suitability of the masked data for analysis, a specific feature - sleep disorder - was kept during the masking process. After applying the donut masking to the data, it was analyzed how the sleep disorders were proportionally distributed to the different decibel zones allowing a direct comparison of the data before and after masking. The points containing information on the prevalence of sleep disorders were assigned to different decibel zones to quantify the exposure to environmental noise.

## 4 Application Results

As described in Sect. 3.2, the minimum and maximum radius of the donut method were successively increased in terms of balancing data anonymity and quality, combined with other features (in the form of quasi-identifiers) such as age and gender.

### 4.1 Evaluation of Anonymity

In the following, we consider the effect of anonymizing geocoordinates, age and gender on the criterion of  $k$ -anonymity. Here, the  $k$ -value achieved via anonymization is determined for each observation. Some descriptive parameters of the resulting distributions



**Fig. 3.** Distribution of observed  $k$ -values by donut masking variant  $\{r \text{ min}; r \text{ max}\}$

are shown in Fig. 3 and Table 1. In the boxplots regarding Fig. 3, outliers to the top (very high  $k$ -values) are not shown for reasons of better readability.

To interpret Table 1, we look at the fourth row, i.e. the combination of donut masking {200;300} with  $AC1$  and *Gender*. The mean of the observed  $k$ -values is 13.76, the median is 8, while 22.17% of the observations show a  $k$ -value of less than 3. In literature,  $k$ -values smaller than 3 are usually classified as critical because a potential data attacker himself could have participated in the survey as respondent and hence could contribute to aggregates. The lower quartile has a value of 3, the upper quartile a value of 20, which means that around half of the observations have  $k$ -values between 3 and 20. The increasing  $k$ -anonymity with stronger anonymization - evident in the increasing quantiles, arithmetic means and ranges - was to be expected, as was the right skew of the distributions.

As a result of further recoding the age categories from  $AC1$  to  $AC2$ , the medians approximately double, while the change in the mean values is somewhat smaller. Figure 3 shows that the skewness decreases slightly, and the mean values and medians approach each other after recoding. Within the  $AC2$  variants, the proportion of critical  $k$ -values falls more sharply than expected with successive increases in the radii of the donut masking: for the  $AC1$  variants, this falls from 44.01 to 11.20, for the  $AC2$  variants from 21.61 to 1.06. Regarding the strongest donut masking setting with radii (300;400), the proportion of critical  $k$ -values falls from 11.20 (variant with  $AC1$ ) to 1.06 (variant with  $AC2$ ). If age and gender are completely omitted, i.e. if donut masking is used alone, the critical  $k$ -values even disappear completely when applying stronger donut masking.

From the authors' point of view, a proportion of critical  $k$ -values below 10% appears tolerable when generating research data such as scientific-use-files. This would apply to all anonymization variants beneath the eighth row in Table 1.

**Table 1.** Distribution of spatial  $k$ -anonymity for different donut masking variants

Feature	$r$ min	$r$ max	mean $k$	% $k < 3$	median $k$	lower quartile	upper quartile
AC1 & Gender	100	200	5.61	44.01	3	1	8
	100	300	9.67	33.22	5	2	13
	100	400	15.02	25.72	7	2	19
	200	300	13.76	22.17	8	3	20
	200	400	19.58	16.61	11	4	27
	300	400	25.46	11.20	14	6	38
AC2 & Gender	100	200	7.23	21.61	6	3	10
	100	300	12.42	12.93	9	4	17

(continued)

**Table 1.** (*continued*)

Feature	$r$ min	$r$ max	mean $k$	% $k < 3$	median $k$	lower quartile	upper quartile	
No Features	100	400	19.26	8.91	13	6	26	
	200	300	17.66	4.04	14	8	23	
	200	400	25.10	2.53	19	10	33	
	300	400	32.65	1.06	26	15	44	
	100	200	52.43	0.22	43	26	70	
	100	300	90.22	0.12	68	37	121	
No Features	100	400	139.75	0.08	ara>	98	49	190
	200	300	128.04	0.01	109	65	173	
	200	400	182.07	0.01	146	85	245	
	300	400	236.90	0.00	203	120	326	

## 4.2 Evaluation of Data Quality

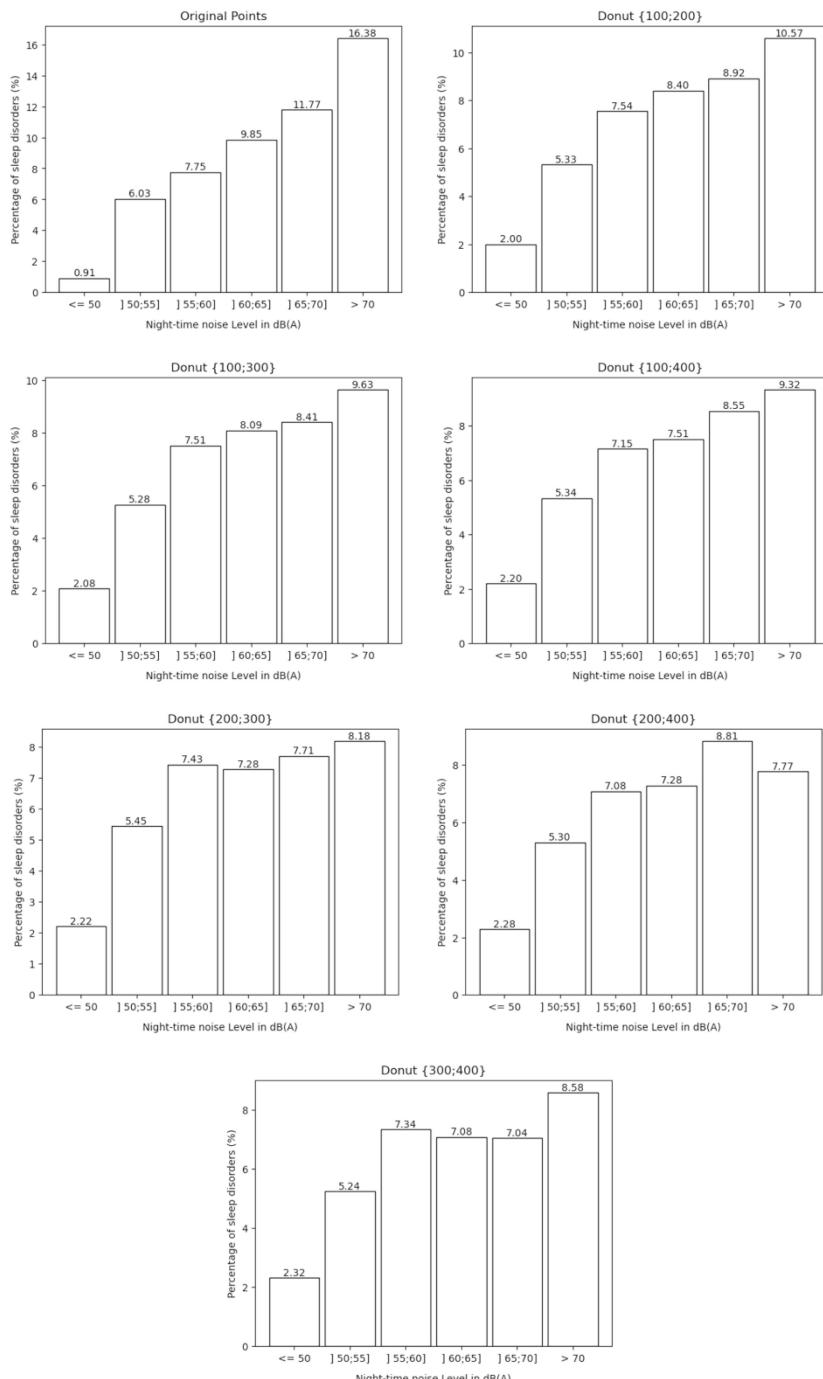
**Spatial Similarity.** The observed mean distance within the original data set is 8.3477 m, the expected mean distance is 36.4899 m and thus we get as nearest neighbor ratio 0.2288. Since the latter is significantly  $< 1$ , the points show a strong trend towards clustering. The associated results of different donut masking settings are shown in Table 2.

**Table 2.** Nearest Neighbor Tool results for the donut masking variants

Masking parameters	{100;200}	{100;300}	{100;400}	{200;300}	{200;400}	{300;400}	{300;500}
Mean distance [m]	23.0442	24.2697	24.8470	24.3554	25.0502	25.1665	25.7502
Expected mean distance [m]	36.9746	37.0750	37.4633	37.2443	37.4761	37.4284	37.7105
Nearest neighbor ratio	0.6232	0.6546	0.6632	0.6539	0.6684	0.6724	0.6828

As the outer radius increases, the average distance between the points also increases. This drifting apart is observed as well in the corresponding nearest neighbor ratio. However, if we look at ratios below 1, we can say that the tendency to form clusters is not lost because of masking. In general, all masking results as depicted in Table 2 show small deviations among each other. From this it can be concluded that although a lot of quality is lost through masking, a gradual modification of the parameters tends to have a small effect.

**Suitability for Analysis.** The results as shown in Fig. 4 indicate that masking with an inner radius of at least 100 m led to a relatively slight shift in the proportions of sleeplessness within the noise zones. However, the general tendency in the sense that the higher the noise exposure, the more frequently sleep disorders are observed, remained largely preserved. The masked data thus reflected similar spatial distributions to those



**Fig. 4.** Percentage of sleep disorders by noise level before and after masking

observed in the original data. The anonymization variants, where the inner radius was greater than or equal to 200 m, showed much stronger deviations from the original distribution.

Maintaining the suitability for analysis after masking is a critical aspect that influences the validity of conclusions in scientific studies. The results suggest that an enlargement of the inner radius has a greater influence on the analysis result than an enlargement of the outer radius. The choice of a smaller inner radius of 100 m thus allows conclusions to be drawn about the spatial distribution of sleep disorders in connection with noise exposure with tolerable restrictions.

### 4.3 Holistic Evaluation and Limitations

The impact of donut masking on the privacy-utility framework in terms of anonymity, quality and suitability for analysis is essentially depending on the existing number of additional features or quasi-identifiers, respectively. Without the knowledge of *Gender* and *Age*, the selection of the smallest radii {100;200} already shows good results. Among the anonymization variants considered in this study, the suitability for spatial analysis is maintained here best possible and the  $k$ -values are sufficiently high.

If the attributes of gender and age are also reviewed in the data set, the anonymization variant using *AC2*, *Gender* and  $\{r_{\min} = 100 \text{ m}; r_{\max} = 400 \text{ m}\}$  as radii for donut masking, according to row 9 in Table 1 is most promising. Regarding the degree of anonymity, the proportion of critical  $k$ -values less than 3 lies below 10%, which appears tolerable in the sense of releasing microdata for research purposes. Although the spatial shifts are quite large for all variants considered here (see Table 2), the analytical results (see Fig. 4) are well preserved for this variant due to the smaller inner radius chosen.

## 5 Conclusion and Future Work

The present study makes it clear that to implement geomasking in practice, it is essential for data holders to review multivariate associations, i.e. the necessity or possibility of anonymizing additional quasi-identifying or sensitive attributes. The practical feasibility on the part of data holders is often neglected in literature. For the everyday use of these methods in research data centers, it is found that a classic recoding of categorial attributes might have a much greater effect on data privacy than the application of donut masking even when using small radii for the latter.

With the aim of granting scientists privileged access to confidential microdata, the authors evaluate a proportion of critical  $k$ -values below the 10% threshold as tolerable when generating standardized research data sets such as what are called scientific-use-files. However, there is also criticism of the concept of  $k$ -anonymity. For example, it does not provide sufficient protection regarding the phenomenon of homogeneity in sensitive attributes. This problem arises when the values within a group are very similar or identical. To counteract related problems, the concept of spatial  $k$ -anonymity could be extended with approaches such as  $l$ -diversity [15] or  $t$ -closeness [16] in future studies. In addition, tests with several data sets could be carried out to ensure the reliability of the results. The authors also recommend setting both radii of the donut method adaptively.

This would counteract the problem that points have so far been alienated globally with the same minimum and maximum parameters, but population density and the occurrence of noise pollution, for example, are local phenomena to be preserved best possible. Our results illustrate the challenges that anonymization faces when aiming to support multiple likely and possibly competing uses, while use case-specific anonymization can provide greater utility. This aspect should be considered when evaluating the associated costs of anonymized data and attempting to maintain sufficiently high levels of privacy for anonymized data.

The research has shown that donut masking can be an effective tool to protect sensitive data while maintaining their analysis potential. It thus provides a basis for further research aimed at balancing data protection and analytical validity.

**Acknowledgments.** This work was carried out within the research cluster “Anonymization of integrated and georeferenced Data” (AnigeD). AnigeD is supported by the Research Network Anonymization for Secure Data Use of the German Ministry of Education and Research supported by the Federal Government’s research framework program on IT security “Digital. Secure. Sovereign.” funded by the European Union – NextGenerationEU.

**Data Availability.** The data sets generated during this study are not publicly available but are available from the corresponding authors on request.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Armstrong, M.P., Rushton, G., Zimmermann, D.L.: Geographically masking health data to preserve confidentiality. *Stat. Med.* **18**, 497–525 (1999). [https://doi.org/10.1002/\(SICI\)1097-0258\(19990315\)18:5%3c497::AID-SIM45%3e3.0.CO;2-%23](https://doi.org/10.1002/(SICI)1097-0258(19990315)18:5%3c497::AID-SIM45%3e3.0.CO;2-%23)
2. Lenz, R.: Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung. *Statistik und Wissenschaft*, 18, Destatis, Germany (2010)
3. Hampton, K.H., et al.: Mapping health data: improved privacy protection with donut method geomasking. *Am. J. Epidemiol.* **172**(9), 1062–1069 (2010)
4. City of Cologne (Stadt Köln), Kölner Stadtteilinformationen - Bevölkerungszahlen 2023. [https://www.stadt-koeln.de/mediaasset/content/pdf15/statistik-einwohner-und-hausalte/koelner\\_stadtteilinformationen\\_zahlen\\_2023\\_einwohner.pdf](https://www.stadt-koeln.de/mediaasset/content/pdf15/statistik-einwohner-und-hausalte/koelner_stadtteilinformationen_zahlen_2023_einwohner.pdf). Accessed 18 May 2024
5. City of Cologne (Stadt Köln), Strassenverzeichnis 2024. [https://www.offenedaten-koeln.de/dataset/strassen-köln](https://www.offenedaten-koeln.de/dataset/strassen-koeln). Accessed 18 May 2024
6. City of Cologne (Stadt Köln), Amt für Landschaftspflege und Grünflächen. <https://offene-daten-koeln.de/dataset/gruenflaechenkataster-koeln-flaechentypen>. Accessed 18 May 2024
7. City of Cologne (Stadt Köln), Umgebungslärm Nacht Köln. <https://offenedaten-koeln.de/dataset/umgebungslaerm-nacht-koeln/resource/775e28c8-1d48-47dc-91b7-270f10b6bef1>. Accessed 18 May 2024
8. Swanlund, D., Schuurman, N., Zandbergen, P., Brussoni, M.: Street masking: a network-based geographic mask for easily protecting geoprivacy. *Int. J. Health Geogr.* **19**(26), 1–11 (2020)

9. Sweeney, L.: K-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowl.-Based Syst.* **10**(5), 557–570 (2002). <https://doi.org/10.1142/S0218488502001648>
10. Allshouse, W.B., et al.: Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto Int. Remote Sens. GIS Hum. Behav. Health Res.* **25**(6), 443–452 (2010). <https://doi.org/10.1080/10106049.2010.496496>
11. Broen, K., Rob, T., Jon, Z.: Measuring the impact of spatial perturbations on the relationship between data privacy and validity of descriptive statistics. *Int. J. Health Geogr.* **20**(3), (2021). <https://doi.org/10.1186/s12942-020-00256-8>
12. Houfaf-Khoufaf, W., Touya, G.: Geographically masking addresses to study COVID-19 clusters. *Univ. Gustave Eiffel* (2021). <https://doi.org/10.1080/15230406.2021.1977709>
13. Esri: ArcGIS Pro. <https://pro.arcgis.com/de/pro-app/latest/tool-reference/spatial-statistics/average-nearest-neighbor.htm>. Accessed 10 May 2024
14. Delmelle, E.M., Desjardins, M.R., Jung, P., Owusu, C., Hohl, A., Dony, C.: Uncertainty in geospatial health: challenges and opportunities ahead. *Ann. Epidemiol.* **65**, 15–30 (2022). <https://doi.org/10.1016/j.annepidem.2021.10.002>
15. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering, pp. 24–36 (2006). <https://doi.org/10.1145/1217299.1217302>
16. Li, N., Li, T., Venkatasubramanian, S. (eds.): t-closeness: Privacy beyond k-anonymity and l-diversity. In: IEEE 23rd International Conference on Data Engineering (2006). <https://doi.org/10.1109/ICDE.2007.367856>



# Privacy and Disclosure Risks in Spatial Dynamic Microsimulations

Hanna Brenzel<sup>1</sup>(✉), Martin Palm<sup>1</sup>(✉) , Jan Weymeirsch<sup>2</sup>(✉) , and Ralf Münnich<sup>2</sup>(✉)

<sup>1</sup> Federal Statistical Office (Destatis), Gustav-Stresemann-Ring 11,  
65189 Wiesbaden, Germany

{hanna.brenzel,martin.palm}@destatis.de

<sup>2</sup> Trier University, Universitätsring 15, 54296 Trier, Germany

{weymersch,muennich}@uni-trier.de

**Abstract.** Microsimulations are used as a tool for evidence-based policy, to better understand the impact of policies on society. However, the quality of the outcome considerably depends on the quality of the data in use. In order to provide a rich set of variables on granular details, synthetic data may be involved. This becomes even more important in dynamic microsimulation where projection into the future is simulated or when providing an open data environment for the research community with a vast amount of variables including geocoded information.

The present paper discusses opportunities and challenges of such synthetic but realistic data generation in a microsimulation data lab, which includes two steps of disclosure control.

First, the base information which is used for the synthetic data generation has to be reviewed in terms of disclosure risks. Second, the data generating process must be ensured to not systematically reproduce rare events for (synthetic) individuals or replicating original input data. Otherwise, due to the large amount of additional information, this may lead to a cumulative effect of the individual re-identification risks. In order to make these output data of dynamic spatial microsimulations available in the sense of open and reproducible research, such statistical disclosure risks must be excluded a priori.

This study examines whether disclosure risks may occur when synthetic data is generated via anonymized data from official statistics sources and how these can be avoided in principle. Furthermore, we discuss methods within the framework of spatial dynamic microsimulation frameworks that automatically ensure the standards of statistical disclosure control as well as official statistics data providers during simulation runs.

**Keywords:** Statistical Disclosure Control · Dynamic Microsimulation · Synthetic Population · Small Area Spatial Simulation · Monte-Carlo Simulation · Confidentiality · Open Reproducible Research

## 1 Introduction

Dynamic microsimulation is a tool to stochastically project microdata throughout simulation periods into the future. In order to simulate regional differences, *spatial* dynamic microsimulations are used that consider geographic differences within the population. This method may not only be used for projection purposes, but also for answering various projection scenarios, giving insights into population dynamics under different political and economic policies as well as demographic developments.

According to Li and O'Donoghue [12], microsimulations consist of two central parts: the actual simulation in the sense of *if-then* processes that are solved simulative, and the data integration and generation. The latter is based on the fact that complete data sets with all the required variables and a sufficiently large number of individuals are rarely available.

One such microsimulation framework is the spatial dynamic microsimulation *Multi-sectoral Regional Microsimulation Model (MikroSim)*. The project is an interdisciplinary endeavour with the goal of creating a concise, close to reality synthetic base population that represents the German population, synthesized from various data sources, including the German census 2011. This digital twin of Germany is then simulated throughout time until the present and projected into the future [13]. For this purpose, synthetic data generation is absolutely necessary in order to obtain a data set of appropriate size and granularity at all, as there is no single original source which contains all features, let alone the spatial resolution required for analyses within small areas [13, 15, p. 9ff].

Thereby, population dynamics should be recreated as they appear in reality both on large scale, for example on national level, and within small areas on almost arbitrary geographic level. While the complex interplay of individuals within this digital twin should be considered, individual real-world units that make up the population should not be re-created for obvious privacy reasons.

To achieve this, the interconnection of various data sources and synthetization techniques are necessary. Besides the data integration aspect, these methods also achieve a certain degree of anonymization within a micro population database [7]. Synthetization as a means of anonymization takes into account, that the individual case is not decisive to meet statistical relationships and frequency distributions. However, the degree to which such synthetic data is *anonymous* is non-trivial to derive. This issue is of particular interest for National Statistical Authorities which provide real-world information for society. The different simulation runs with which results are usually generated and the intended purpose of projecting several years into the future create additional uncertainty in relation to the original data [15, p. 204ff].

In general, the synthetic nature of the simulation base population within MikroSim stems from coefficients that have already been checked for Statistical Confidentiality within the German Federal Statistical Office. Despite this, concerns about exact copies of rare characteristic combinations for real-world individuals may arise in fine-grained spatial simulation setups. Naturally, any common type of synthetization or projection approach will produce units within

the micro-level population which are, by pure chance, very similar to real individuals [21, p. 49ff].

This is not necessarily regarded as a disclosure issue per sé, as long as cases are not reproduced systematically. Whether the replication of individuals is to be considered *systematic*, depends on the number of occurrences as well as the subpopulation this appears in [1, p. 6ff].

Furthermore, within a single simulation run, it depends not only on how many times real-world individuals are reproduced cross-sectionally, but also longitudinally within the projection phase, i.e. whether the life course of individuals are recreated, putting them at risk of actual disclosure.

One goal of the MikroSim project is to make the data available to the research public in the sense of open and reproducible research in the form of a Simulation Data Centre (SimDC), where institutes and researchers may undertake their own analyses. In order to ensure confidentiality even when data with high granularity and detailed spatial references are in use, measures must be taken to control the disclosure risk of individuals present in the input data for the synthetization process. These measures can be implemented both at the data level (data manipulation) and at the data access level (roles and rights). The considerations in this paper are therefore intended to serve as a basis for a role, rights, and confidentiality concept.

The paper is structured as follows: Sect. 2 sheds light on the differentiation between data privacy in academic research and German official statistics as well as the precautions implemented by the according research data centres regarding the usage of real-world information for the research community. Section 3 explains further concepts and methods implemented in MikroSim that prevent disclosure while maintaining an appropriate degree of data utility. Finally, Sect. 4 summarizes our perspective and gives an outlook on future research.

## 2 Data Confidentiality in Research and Official Statistics

Generally, data protection measures are intended to prevent the use of data in bad faith and limit the damage caused by exploitation. The European Code of Practice, which applies to official statistics in the entire EU, explicitly guarantees confidentiality regarding participants' information. Neither their identity nor particularly vulnerable information are to be disclosed, therefore aiming to protect privacy while disseminating sound statistical output [9]. This is intended to create acceptance and trust in statistical procedures and institutions. Various formal data protection models exist to achieve this. National statistical institutes typically create their own rules, following these guidelines.

These may not necessarily align with concepts used in academic research exactly and negotiation between both interest groups is often times not straight forward. Therefore, in the context of this paper and due to the applications within a dynamic spatial microsimulation framework, we aim to distinguish explicitly between concepts of Statistical Disclosure Control (SDC) and disclosure harm in academic research and the requirements posed by German official

statistics. This harmonization is not solely important for the MikroSim project, but also for any research with synthetically generated data based on data releases by official statistics within the EU and Germany in particular.

## 2.1 Formal Data Protection Models

Substantially, information can be divided into three general groups: Firstly, *identifying information* that may help to single out units or groups from the database either via unique identifiers or unique patterns of data points, so-called quasi-identifiers. Secondly, *sensitive attributes* that may have adverse effects for some unit if publicized, such as a person's political stance or their income. Both categories are not mutually exclusive, meaning that some information may be both identifying and sensitive, for example the exact address of a person. Lastly, all other data, that does not fit in these two categories, is referred to as *non-confidential information*, as it neither helps to identify some unit, nor is it particularly interesting to an adversary and therefore inconsequential for the SDC procedure [21, p. 35f].

Building on these three essential types of information, one may describe disclosure in a two-dimensional scheme, as illustrated in Fig. 1. The first dimension essentially describes the way some unit in the data set is being disclosed. Disclosure processes may be classified as primary, if some unit is unique in one single attribute or a group of attributes. Secondary disclosure often times appears incidentally, for example because all other units with the same quasi-identifiers are already known. [3, p. 368].

The second dimension describes three specific types of information disclosure that may either happen via primary or secondary disclosure (through the first dimension). It classifies instances as *identity disclosure*, where some unit's identity may be revealed or where it is possible to single it out with full certainty. Secondly, *attribute disclosure* describes scenarios, where the sensitive attribute of some unit is exposed, while the adversary does not necessarily know the unit's identity. Lastly, *inferential disclosure* defines instances where the adversary is not able to specify some unit's sensitive attribute with full confidence, but can narrow down the range its value lies in [21, p. 39f].

With this in mind, it is clear that the three types of disclosure pose different levels of potential disclosure harm to the affected units. Although an adversary with knowledge about the range of some unit's sensitive attribute (inferential disclosure) should generally be avoided, knowledge about the exact value (attribute disclosure) or even about the unit's identity (identity disclosure) certainly is a greater privacy violation and therefore potential harm [3, 17, p. 368]. Because SDC is almost always a delicate balance between risk and utility, consideration of disclosure harms are often times helpful to decide for a degree of privacy measures.

In order to prevent disclosure of attributes and identities within a given database, two common strategies exist that manipulate information such that disclosure is improbable or that attribute disclosure remains unreliable even if

identity disclosure succeeds, namely the enforcement of *indistinguishability* of data points and the introduction of additional *uncertainty*.

Indistinguishability is commonly used in tabularization approaches also undertaken by German official statistics. Two prominent concepts are important for the course of this paper:  $k$ -anonymity and  $\ell$ -diversity. Both concepts are fairly straight-forward and well known in the field of SDC. We refer interested readers to [21, p. 58ff], [10, p. 716ff] and [20].

One important note that needs to be considered carefully, is the type of disclosure attacks that  $k$ -anonymity and  $\ell$ -diversity are able to protect against, or more so which types of disclosure they can not prevent. By considering only equivalence classes (quasi-identifiers),  $k$ -anonymity successfully counters direct identity disclosure and therefore also attribute disclosure in those cases, where sensitive attributes are revealed through knowledge about the unit's identity. However, it does not consider direct attribute disclosure and the more common inferential disclosure attacks. Specifically, in cases where all units of one equivalence class have similar sensitive attribute values, inferential disclosure or even attribute disclosure may occur, even if the adversary may not be able to identify the unit within this group [21, p. 58].

A second strategy frequently used in SDC is the introduction of additional uncertainty, particularly regarding matching- and linking-attacks. In order to prevent such attacks, the fundamental idea is to either involve unclarity about whether one particular unit does exist in a given data release, i.e. uncertainty about identity, or about their true attributes, namely their quasi-identifiers and sensitive attributes. For one, information can be obscured by introducing random noise to quasi-identifiers directly. This, of course, results in an adversary not being able to match or link the released data set to their pre-existing knowledge about some (sub-) population's quasi-identifiers reliably. If noise is additionally added to sensitive attributes, then an adversary can also not be sure about the precise values, even if they manage to identify some units correctly. Thus, in theory the introduction of noise is capable of protecting against both identity and attribute disclosure at the expense of data utility [2, 8, 11].

Synthetization approaches may also be seen as methods introducing uncertainty. Particularly, synthetic units do not represent real-world units on an individual basis, but only in their entirety. Identity disclosure is therefore eliminated on a principle level. Inferential disclosure and to some extent even attribute disclosure could occur in rare edge case scenarios in theory. This may be the case if risk populations are tiny while synthetization models capture reality particularly well or are calibrated accordingly [1]. Such edge-cases need to be considered with special care, depending on the disclosure harm that stems from it.

## 2.2 Confidentiality in German Official Statistics

In contrast to formal data protection models, (German) official statistics do not differentiate between sensitive attributes and non-sensitive attributes, nor does it differentiate various levels of disclosure harms in a special manner. Inferences about individual respondents must generally be avoided. This results from the

requirement that "individual details about personal and factual circumstances [...] must be kept secret" [5] (§16 (1) *Federal Statistics Act* [BStatG]), regardless of how much harm they could cause. Within German official statistics, all information is considered to be equally in need of protection, regardless of the potential damage it may cause to the people affected. In order to achieve this, confidentiality measures are used with the aim of establishing anonymity. These can be data-manipulating or information-reducing and can be applied both before and after generating the tables to be published (pre-tabular vs. post-tabular). Pre-tabular confidentiality measures are also referred to as anonymization and imply the modification of microdata before the data are disseminated by official statistics [14, p. 7ff]. Figure 2 shows an overview of common methods, depending on the type and the time of the intervention.

The established levels of anonymity in German official statistics, some of which are legally defined, can be divided into broad three categories:

- **Formal anonymity:** Individual data without name and address [5] (§5a (3) BStatG). This data harbours the risk of easy re-identification, as only unique, direct identifiers have been removed.
- **Factual anonymity:** Individual details can only be assigned with a disproportionate amount of time, cost, and labour [5] (§16 (6) sentence 1 No. 1 BStatG). Typically, additional knowledge is required for re-identification. The assignment to a specific person is subject to great uncertainty. Establishing factual anonymity is not trivial, as a cost/benefit analysis must be carried out for statistical purposes. This analysis and the risk assessment are influenced by the technical, organizational and contractual conditions for data provision. This means, that *factual anonymity* is not established only by changing the data material. It is rather determined by a triad of the above-mentioned factors.
- **Absolute anonymity:** It is no longer possible to identify the individuals providing the information. Absolutely anonymous data can be published to anyone without further restrictions. This also includes microdata.

The result of this process is anonymity in the form of a certain level of protection. This in turn determines *who* may work with the data, *how* and *where*, and which data may be *published*. Data from German official statistics may only be published if they are considered absolutely anonymous, as there is no statistical confidentiality obligation [5] (§16 (1) sentence 3 No. 4 BStatG). In principle, the public therefore only has access to this type of data. The situation is different for (independent and academic) science. Here, the legislator has introduced a so-called *scientific privilege* which allows *access to* and *use of* data with a lower degree of anonymity.

How the scientific community can access the data depends on the way of access. Individual data may be transmitted to scientific institutions for analysis as long as it is at most factual anonymous [5] (§16 (6) sentence 1 No. 1 BStatG). Access to formally anonymous data, on the other hand, is only permitted within specially secured areas of the Federal Statistical Office and the statistical offices of the federal states for scientific purposes [5] (§16 (6) sentence 2 No. 2 BStatG).

In addition to conducting independent scientific research (at the level of the institution), part of the access regulation is an obligation to maintain statistical confidentiality (at the level of the researcher). However, research results that are to be published, e.g. in the context of scientific publications, must meet the requirements of absolute anonymity. A distinction must therefore be made between the privilege of the scientific community to work with such data and the possibility of using it for publications. Data or results that are not absolutely anonymous may not be published, even in the context of scientific publications. The transmission of factual anonymous data sets to the scientific community results in a responsibility to maintain confidentiality. It must be ensured, that only absolutely anonymous data originating from official statistics is published by researchers.

In the case of formally anonymous data within the specially secured areas of the Federal Statistical Office and the statistical offices of the federal states (so-called *safe centres*), the confidentiality check of the results is carried out by the official statistics staff themselves. The so-called *confidentiality check* is carried out for all generated results before they leave the specially secured areas of German official statistics. Typically, so-called post-tabular confidentiality methods are used in contrast to procedures that are applied before tabularization (see Fig. 2). These can be data-manipulating, however often times information is simply suppressed according to pre-defined rules. A common rule that is intended to prevent the re-identification of feature carriers is the *minimum case number rule*. It corresponds to the  $k$ -anonymity model and is usually at least  $k := 3$ .

The transmitted results, which are checked for confidentiality, fulfil the conditions of absolute anonymity. This also applies to the results generated on the basis of official statistics for the generation of the synthetic MikroSim population. The use of coefficients and probabilities to generate synthetic data constitutes publication in the sense of the aforementioned considerations. Since data sets from German official statistics were used to a large extent in MikroSim, it must be ensured that no individual values occurring in reality are exactly replicated within the simulation. While exact replications rarely may occur during a simulation by nature of stochastic processes, these must not appear systematically. Synthesization is typically used as a method of anonymization and is then classified as a pre-tabular, data-altering method: Synthesized for a specific data set in order to anonymize the original data. This makes it easy to check how well the synthesized data set has changed the original data set.

However, synthetization in the MikroSim project serves a completely different purpose. Primarily, it is not about anonymization. It is rather about generating new, much larger data sets that contain more information.

The outputs from German official statistics used as inputs for the simulation in MikroSim are of course absolutely anonymous. However, these input coefficients and parameters are then used to generate synthetic microdata that should reflect reality as closely as possible. The synthetization process integrates a wide variety of different data sources from statistical institutes, each of which by themselves are absolutely anonymous, as indicated in Fig. 4. The generated

microdata are further aligned and calibrated towards known marginals on small-area level in the sense of a *multi-source-estimation* [13,22]. Generally speaking, due to the interaction of the amount and complexity of information, there is the (theoretical) possibility that synthetic microdata may be generated that replicates reality closely and systematically, if not addressed accordingly during the data generation process. This may lead to re-identification of individuals in certain edge-cases. Therefore, synthesization as a method of confidentiality is not sufficient in this case by itself. The apparent paradox makes it necessary for the synthetically generated data, even if it was generated from absolutely anonymous data, to be checked for confidentiality yet again.

Random replication of *true* values may occur by pure chance, but are relatively unlikely due to the stochastic component of the simulation. In practice, this poses only little risk, because simulations are run multiple times within a Monte-Carlo simulation. This results both in multiple value sets for synthetic individuals and different individuals altogether (due to stochastic processes in household-compositions like partnerships, migrations, births, deaths, etc.). This effect is reinforced by the fact, that it is almost certainly impossible for a data user to know whether a specific value set of such synthetic individual corresponds to a real world individual in every detail. Since every value is determined in part via stochastic processes, even if quasi-identifiers may match by pure chance, sensitive attribute values of the real-world individual, unknown to an adversary, may not be similar to the synthetic individual at all. In this view however, the risk that simulations might systematically replicate reality must be quantified for the MikroSim data. This is to determine whether additional confidentiality measures have to be applied to the synthetic population and whether access to the data material must be restricted or regulated further.

In contrast to MikroSim, the original data set used in a standard publication is 1) directly comparable with the analysis result and 2) there is no risk of inadvertently replicating original values since there is no simulative implementation of the publication that is also accessible to the public. The generated outputs based on German official statistics, do not have the latter properties. This is because they are not (only) used individually for publications or similar purposes, but additionally also as input for the simulations and the base population data sets in MikroSim. The further use of the tested results in the form of a SimDC, where external researchers may be involved and might have access to certain parts of the data, is not a typical way of utilizing absolutely anonymous outputs. The explanations therefore show, that this circumstance must be examined and investigated to ensure that the MikroSim datasets in the SimDC do not accidentally create any re-identification risks for the respondents of the original data.

### 3 Disclosure Control in MikroSim

Within complex dynamic microsimulation frameworks, often times there exists no single database which contains all required information. Instead, various data

sources are integrated within a synthetization approach to form concise synthetic data sets. Coincidentally, synthetization of data may also be interpreted as introduction of uncertainty to the data, as discussed already in Sect. 2.

While partial synthetization protects mostly against attribute disclosure and only to some degree against identity disclosure, a fully synthetic database substantially eliminates the risk for identity disclosure completely, while also protecting against attribute disclosure to some degree. Inferential disclosure, while also highly unlikely in synthetic data, poses little threat whatsoever, even if realized. In the context of multiply run and fully synthetic simulation data (such as in dynamic microsimulations), aspects of SDC are comparable with those of multiple imputed or synthesized data tables [1, 7].

During the simulation process, stochastic procedures extend the data set and project it into the future for each individual in the database. While aggregate statistics of simulation results (such as the number of transitions within various subpopulations) are often times regionalized and aligned towards known values and distributions, the stochastic processes used for these transitions introduce further uncertainty into the outcomes [15, 19, 22]. However, depending on granularity of the initial synthetization approach as well as those of alignment and calibration procedures, rare events and transitions within small subpopulations may in theory be reproduced systematically [1]. For those cases, the Monte-Carlo variation between different simulation runs becomes the main factor of analysing disclosure risks, i.e. how often are rare events reproduced when a simulation is repeated *many* times and whether the units falling within sparsely populated equivalence classes are identical in each run.

Besides the Monte-Carlo effect, introduced by the simulation process itself, dynamic microsimulations inherit a multitude of additional uncertainties and assumptions that contribute to differing outcomes on every single simulation run [4, 16]. These sources of uncertainty on unit-level have vast consequences for the disclosure risk of the affected subpopulations. They should therefore be considered for the overall classification of disclosure risks in spatial dynamic microsimulations, such as MikroSim.

The MikroSim population is based on anonymized population register data and has been calibrated to known distributions from the German Census 2011. Depending on required granularity, different levels of georeferencing may be available, which in turn needs to be considered by the SDC measures. Correlations and marginal distributions are generated synthetically at various levels on the basis of the microcensus and known total values from the census. This basic data set is supplemented with additional information, depending on the thematic focus, e.g. demand for primary school places, care needs, modelling of living space, etc. All non-public input data of the base population, stemming from official statistics sources, are treated with common SDC procedures by the national and regional statistical institutes involved. The SDC-treated parameters and estimates are handed over to the project partners for the synthetization procedure [13].

A similar process is conducted for the transition models used within the simulation itself. As with the initial information for the synthetization process, model parameters are estimated on the data supply within the research data centre of the German Federal Statistical Office. At this point, they already passed the common input SDC procedures within the German official statistics framework discussed in Sect. 2.2 and undergo further output SDC procedures by the official statistics data providers within their own vicinity. Once these parameters pass the according measures, they are handed over and implemented in the MikroSim framework. In this sense, the common research data centre output is used as parameter and data input for the MikroSim framework.

During the simulation, these parameters are used to extend the data set cross-sectionally or project it longitudinally in the form of state transitions. Thereby, variables within the database are extended or updated conditionally on other variables. This process leads to a concise population database and simulation data outputs that are constructed from numerous input data sets, as displayed in Fig. 4 in the Appendix.

Within our simulation framework, rather than exporting microdata from within the simulation, results are generated by aggregating key indicators, such as number of state transitions, from the microlevel population. For example, such aggregate results may be as coarse as counting the number of births within a certain region within one simulation year. However, generally, results may be produced on arbitrary resolution, such as the number of births within a certain region by age and citizenship of the mother. The more detailed aggregation covariates are, the more cases may arise where only few events appear within any equivalence class, up to a point where micro-level data could be reconstructed from the aggregate data output.

To guarantee that rare events and particularly detailed aggregate outputs are protected under the  $k$ -anonymity concept, we implement a simple *stochastic rounding* procedure. Whenever small cell counts  $z_c < k$  appear for any equivalence class  $c$  within the generated output aggregates, the cell value is either rounded downwards to 0 or upwards to a pre-defined minimum cell-value  $k$  (typically 3, 5 or even 10).

Whether the value is rounded up- or downwards is determined by the relationship between  $z_c$  and  $k$ . More precisely,  $z_c$  is replaced by  $k$  with probability  $\mathcal{P} = \frac{z_c}{k}$  or by 0 with its counter-probability  $\mathcal{P}' = (1 - \mathcal{P}) = \frac{k-z_c}{k}$ , as denoted in Eq. 1.

$$f(z_c) = \begin{cases} 0 & \text{with } \mathcal{P}' = \frac{k-z_c}{k} \\ k & \text{with } \mathcal{P} = \frac{z_c}{k} \end{cases} \quad \forall 0 < z_c < k \quad (1)$$

The method results in asymptotically unbiased, true to expectation cell values over *many* simulation runs  $R$ , as denoted in Eq. 2. This means that, while aggregate results from each single simulation run  $r$  do not contain cell values less than  $k$  but more than 0 (i.e. cases that would violate  $k$ -anonymity), arithmetic means of cell values from a specific equivalence class  $c$  across many simulation runs  $R$  would be approximately equal to the expected cell value of  $z_c$ .

$$\lim_{R \rightarrow \infty} \frac{1}{R} \sum_{r=1}^R f(z_{c,r}) \approx \lim_{R \rightarrow \infty} \frac{1}{R} \sum_{r=1}^R z_{c,r} \approx E(z_c) \quad (2)$$

This process leads to individually privacy preserving simulation aggregates, which result in asymptotically unbiased results even for extremely rare events or small subpopulations and equivalence classes with high granularity.

Within an example simulation of mortality state transitions for the city of Trier, based on data for the year 2019, we aggregate simulated deaths within each run on age and sex, thus we define the equivalence class via the (age, sex) tuples for all possible combinations:  $c \in ((0, \text{female}), (1, \text{female}), \dots, (95, \text{male}))$ . Data for this simulation study was taken from the German periodic mortality tables for the year 2019 [18]. The population data is a synthetic micro-level database for the city of Trier, generated from publicly available information [6]. Age within the population is truncated at  $\geq 95$ .

The aggregated mortality cases are then treated with the stochastic rounding procedure and compared to the same mortality cases without rounding treatment. An example of the effects from stochastic rounding are depicted in Fig. 3. As can be easily seen on the left side of the figure, no cell values smaller than  $k := 3$  are reported within a single run after applying stochastic rounding. In particular, differences between the pre- and post-treatment simulations arise for both females and males between the ages 0 and 55.

Contrary to this, when repeating the simulations 1,000 times, averaging the number of deaths for each equivalence class across all  $R$  simulation runs, as shown on the right side of Fig. 3, the simulated and rounded values match the expected number of mortality cases for each equivalence class quite well. Importantly, for any equivalence class between the  $R$  simulation runs, different populations units make the transition, i.e. as long as the transition probability of a unit is less than 100% in every run, disclosure risk decreases with each simulation run. In this example simulation, maximum mortality probability is less than 30% (males, 95 years old).

Lastly, one major goal within the second phase of the MikroSim project is the establishment of a Simulation Data Centre (SimDC) for the European research community. For this purpose, a multi-layer confidentiality architecture is established, regarding the population database used within the simulation. It is strictly divided into three levels, which roughly correspond to the three anonymity levels of German official statistics, discussed in Sect. 2.2 (albeit, unlike within the research data centre in official statistics, the MikroSim base populations are **synthetic** data in all cases where disclosure of identity is factually impossible in any way): A) *Level-0* is the inner layer of the SimDC, where only few project members have direct access to the population data via a high security high-performance-computing cluster. This cluster is strictly separated (*air-gapped*) from other networks and computers. Here, the most granular and detailed data set is used, which inherits high data utility and detailed georeferences of the (synthetic) individuals. In terms of official statistics, this layer would correspond to the non-anonymized data. B) *Level-1* is the middle layer, which is available to all

other project partners and researchers with according usage contracts. The simulation results and aggregate outputs are provided via high-security interfaces on a separate High-Performance-Computing cluster. Direct access to or export of the microdata is not available in almost all cases. The data set used within simulations of this middle layer are a delicate compromise between disclosure risk and data utility. The level of anonymity is equivalent to the (at most) formal or factually anonymous data. C) *Level-2* is the outer layer of the SimDC, that is available for the general research community. This data is synthesized in a way that prioritizes low disclosure risks over utility and is mostly used for development of simulation modules or teaching purposes. This outer layer is the synthetic equivalence of what the research data centre releases as so-called *campus files*, which are considered to be absolute anonymous.

## 4 Conclusion and Outlook

In this paper, we outline a common SDC framework for spatial dynamic microsimulations with synthetic populations on high granularity and spatial resolution. The measures presented have already been implemented in the MikroSim framework. The concept of output confidentiality combined with an access concept dependent on the need for protection of the data has already been proven itself in official statistics and has been adapted and extended for the need of the MikroSim project and the SimDC. During the two project funding phases, direct project partners conducted detailed and novel analyses on the middle layer. Detailed investigation of the process and outputs showed, that research on methods as well as applications is possible while maintaining data protection requirements and confidentiality aspects. The measures to secure the synthetic population data therefore extend to both the data itself and access rights.

Regardless of this, before opening the SimDC to external research, we deem it necessary to demonstrate the effectiveness of the security measures taken in the project in practice. As there is no single original source for the MikroSim population, a direct comparison between synthetic and real-world data is not straightforward. The available official data, used to estimate the coefficients that were used for the synthetization of the base population, must not leave the protected area of German official statistics, while no raw data at micro level may be exported from MikroSim's inner layer either, even among the project partners, as long as disclosure risks cannot be ruled out. How such a comparison could be designed will be shown in an independent work.

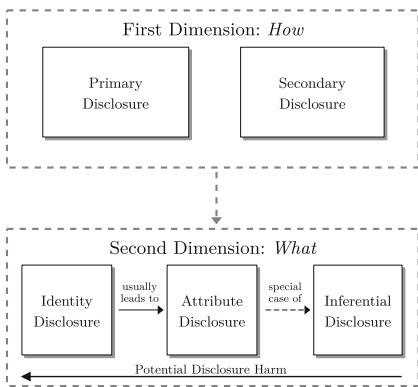
An audit for disclosure risks serves, on the one hand, the demonstration that the mandatory confidentiality checks from the regional and national statistical institutes have been carried out correctly from the perspective of official statistics and that only absolutely anonymous data is permitted when issuing outputs. On the other hand, we aim to show that the confidentiality methods used by German official statistics are suitable for fulfilling the intended purpose, namely to avoid disclosure risks, while maintaining data utility. Lastly, this demonstration is essential for further use of the database in the form of a SimDC, where external researchers are granted access to the MikroSim codebase and data, in which case both for ethical and legal reasons, simulation outputs must be absolutely anonymous.

**Acknowledgements.** This research was funded by the German Research Foundation (DFG) grant number FOR 2559.

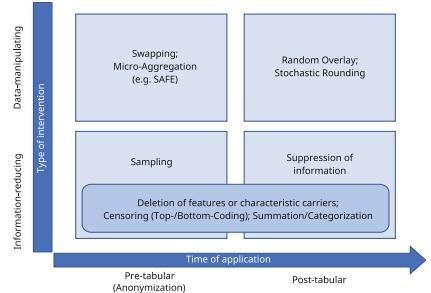
**Author Contributions.** Conceptualization: HB, MP, JW, RM; methodology: MP, JW; resources: RM; writing—original draft preparation: MP, JW; writing—review and editing: HB, MP, JW, RM; visualization: MP, JW; supervision: HB, RM; project administration: RM; funding acquisition: HB, RM. All authors have read and agreed to the published version of the manuscript.

**Disclosure of Interests.** The author Dr. Hanna Brenzel is leader of the Research Data Centre at the German Federal Statistical Office (Destatis). Prof. Dr. Ralf Münnich is project speaker for the MikroSim FOR 2559 project. All authors are part of the MikroSim research group funded by the German Research Foundation.

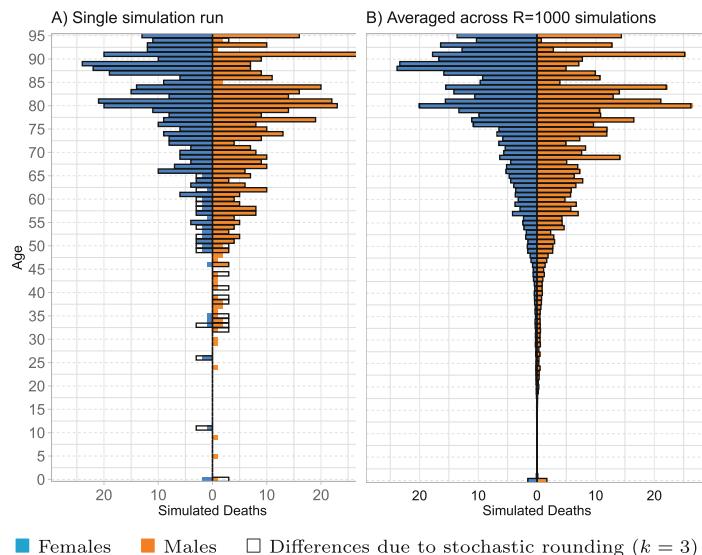
## Appendix



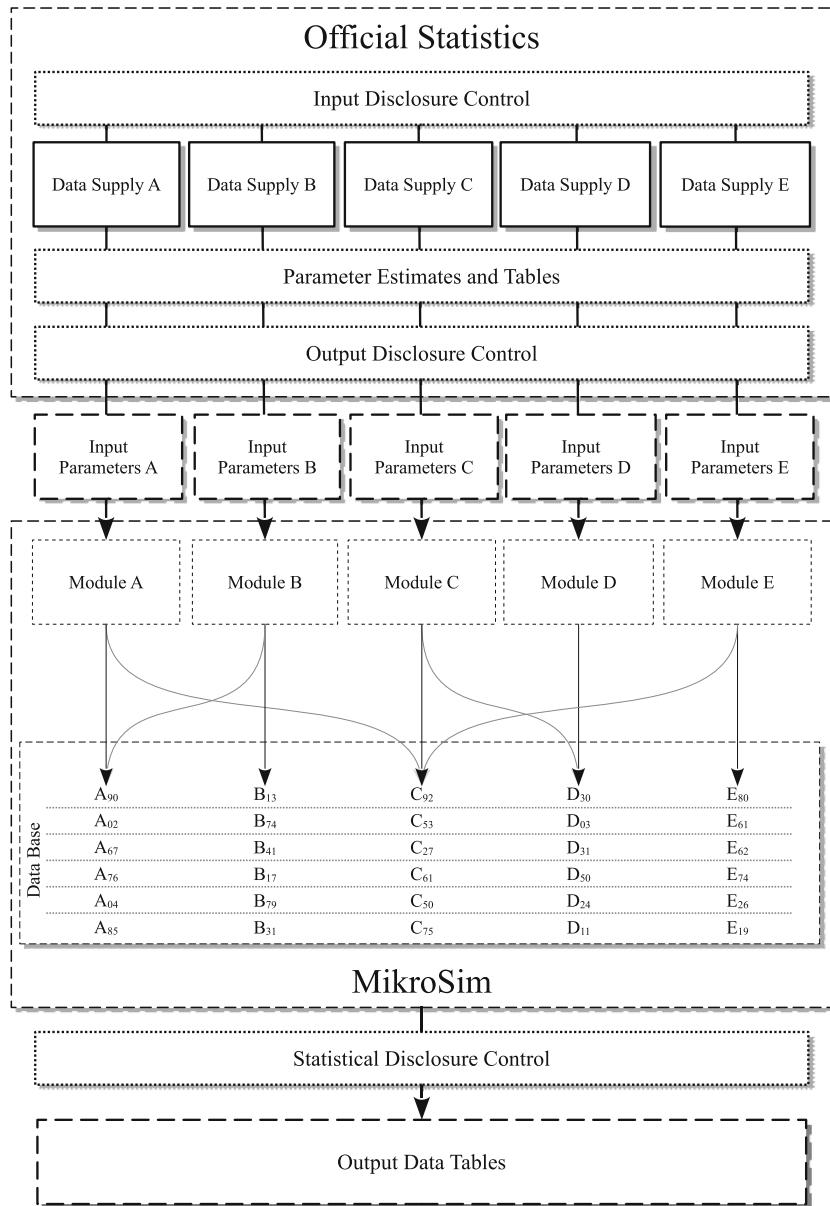
**Fig. 1.** Dimensions of Disclosure



**Fig. 2.** Confidentiality measures depending on time of application and type of intervention following [14, p. 7ff])



**Fig. 3.** Simulated Mortality of a synthetic population, Trier 2019



**Fig. 4.** Construction of the MikroSim base population under Statistical Disclosure Control

## References

1. Ahmed, A., et al.: Overview report on the generation of synthetic universes for microsimulations (2021). deliverable D12.6, InGRID-2 project 730998 - H2020
2. Antal, L., Enderle, T., Giessing, S.: Harmonised protection of census data in the ess (2017). [https://ec.europa.eu/eurostat/cros/content/methods-protecting-census-data\\_en](https://ec.europa.eu/eurostat/cros/content/methods-protecting-census-data_en)
3. Bach, F.: Statistical disclosure control in geospatial data: the 2021 Eu census example. In: Döllner, J., Jobst, M., Schmitz, P. (eds.) Service-oriented Mapping: Changing Paradigm in Map Production and Geoinformation Management, chap. 18 (2019). [http://deposit.dnb.de/cgi-bin/dokserv?id=6865c1fe9fb14e6ba6e87b3b4102234d&prov=M&dok\\_var=1&dok\\_ext=htm](http://deposit.dnb.de/cgi-bin/dokserv?id=6865c1fe9fb14e6ba6e87b3b4102234d&prov=M&dok_var=1&dok_ext=htm)
4. Bilcke, J., Beutels, P., Brisson, M., Jit, M.: Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. *Med. Decis. Mak.* **31**(4), 675–692 (2011). <https://doi.org/10.1177/0272989X11409240>
5. Bundesrepublik Deutschland: Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz - BStatG) [Federal Statistics Act] (2022). [https://www.gesetze-im-internet.de/bstatg\\_1987/](https://www.gesetze-im-internet.de/bstatg_1987/). Accessed 29 Dec 2022
6. Burgard, J.P., Shams, S., Pamplona, J.V.: Gesyland – synthetic replica of germany (2024). <https://www.gesundheitsforschung-bmbf.de/de/teilprojekt-trier-17291.php>, (version: 0.0.1)
7. Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the german iab establishment panel. *Trans. Data Priv.* **1**(3), 105–130 (2008)
8. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. *J. Priv. Confid.* **7**(3), 17–51 (2017). <https://doi.org/10.29012/jpc.v7i3.405>
9. Eurostat: European statistics code of practice (2018). <https://doi.org/10.2785/798269>, <https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142>
10. Ghinita, G., Tao, Y., Kalnis, P.: On the anonymization of sparse high-dimensional data. In: 2008 IEEE 24th International Conference on Data Engineering. Bd. [1]. No. 1, IEEE Service Center, Piscataway, NJ (2008)
11. de Jonge, E., de Wolf, P.-P.: Spatial smoothing and statistical disclosure control. In: Domingo-Ferrer, J., Pejic-Bach, M. (eds.) PSD 2016. LNCS, vol. 9867, pp. 107–117. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-45381-1\\_9](https://doi.org/10.1007/978-3-319-45381-1_9)
12. Li, J., O'Donoghue, C.: A survey of dynamic microsimulation models: uses, model structure and methodology. *Int. J. Microsimulation* **6**(2), 3–55 (2013)
13. Münnich, R., et al.: A population based regional dynamic microsimulation of Germany: the Mikrosim model. *Methods Data Anal.* **15**(2), 24 (2021)
14. Rothe, P.: Statistische geheimhaltung - der schutz vertraulicher daten in der amtlichen statistik. teil 1: Rechtliche und methodische grundlagen. Statistische Ämter des Bundes und der Länder: FDZ-Arbeitspapier (50), 3–13 (2019). <https://doi.org/10.29012/jpc.v7i3.405>
15. Schmaus, S.: Methoden regionalisierter dynamischer Mikrosimulationen. PhD dissertation, Universität Trier (2023). <https://doi.org/10.25353/ubtr-xxxx-512e-3257>
16. Sharif, B., Kopec, J., Wong, H., Finès, P., Sayre, E., Liu, R., Wolfson, M.: Uncertainty analysis in population-based disease microsimulation models. *Epidemiol. Res. Int.* **2012**(2012) 14 (2012). <https://doi.org/10.1155/2012/610405>

17. Skinner, C.: Statistical disclosure risk: separating potential and harm. *Int. Stat. Rev.* **80**(3), 349–368 (2012). <https://doi.org/10.1111/j.1751-5823.2012.00194.x>
18. Statistisches Bundesamt: Sterbetafel (Periodensterbetafel): Deutschland, Jahre, Geschlecht, Vollendetes Alter (2020). <https://www-genesis.destatis.de/genesis/online?operation=table&code=12621-0001>. Accessed 17 May 2024
19. Stephensen, P.: Logit scaling: a general method for alignment in microsimulation models. *IJM* **9**(3), 89–102 (2016). <https://doi.org/10.34196/ijm.00144>
20. Sweeten, G.: Scaling criminal offending. *J. Quant. Criminol.* **28**(3), 533–557 (2012). <https://doi.org/10.1007/s10940-011-9160-8>
21. Templ, M.: Statistical disclosure control for microdata: methods and applications in R (2017)
22. Weymeirsch, J., Ernst, J., Münnich, R.: Model recalibration for regional bias reduction in dynamic microsimulations. *Mathematics* **12**(10) (2024). <https://doi.org/10.3390/math12101550>, <https://www.mdpi.com/2227-7390/12/10/1550>

# **Machine Learning and Privacy**



# Combinations of AI Models and XAI Metrics Vulnerable to Record Reconstruction Risk

Ryotaro Toma<sup>(✉)</sup> and Hiroaki Kikuchi<sup>✉</sup>

Graduate School of Advanced Mathematical Sciences, Meiji University,  
4-21-1 Nakano, Tokyo 164-8525, Japan  
[{cs242022,kikn}@meiji.ac.jp](mailto:{cs242022,kikn}@meiji.ac.jp)

**Abstract.** Explainable AI (XAI) metrics have gained attention because of a need to ensure fairness and transparency in machine learning models by providing users with some understanding of the models' internal processes. Many services, including Amazon Web Services, the Google Cloud Platform, and Microsoft Azure run machine-learning-as-a-service platforms, which provide several indices, including Shapley values, that explain the relationship between the output of the black-box model and its private input features. However, in 2022, it was demonstrated that a Shapley-value-based explanation could lead to the reconstruction of private attributes, posing a privacy risk of information leakage from the model. It was shown that the leaked value would depend on the AI black-box model used. However, it was not clear which combinations of black-box model and XAI metric would be vulnerable to a reconstruction attack. The present study shows, both theoretically and experimentally, that Shapley values are indeed vulnerable to a reconstruction attack. We prove that Shapley values for a linear model can lead to a perfect reconstruction of records, that is, they can enable an accurate estimation of private values. In addition, we investigate the impact of various optimization algorithms used in attack models on the reconstruction risk.

**Keywords:** Shapley Values · Explainabilities · XAI · Reconstruction Attack

## 1 Introduction

Machine learning (ML) models have recently been used in important use cases, including finance, healthcare, E-commerce, and employment [1–3]. One well-known issue with artificial intelligence (AI) models is a lack of transparency. Many kinds of models, particularly deep neural networks and ensemble models, have complex internal structures regarded as “black boxes,” which prevent internal analysis of their operation and therefore how the models arrive at their decisions. For an application to be trustworthy, we need some transparency about the mapping between the input features and the outputs of the model used.

Explainable AI (XAI) technologies are therefore necessary to guarantee the transparency of models and to explain the relationship between their input features and their outputs [1, 4]. XAI helps to build trust by providing AI users with insights into how systems work and why they have made specific decisions. In addition to gaining transparency in AI models, XAI can mitigate biases in AI models and can address social concerns about their use.

Currently, XAI indices are available for most machine-learning-as-a-service (MLaaS) platforms, which provide ML models with some XAI indices that explain how the input features affect the outputs of the models. In particular, Shapley-value-based XAI methods [15, 16] are state-of-the-art and are offered for the major commercial MLaaS platforms, including Amazon Web Services [5], Google Cloud [18], and Microsoft Azure [6].

Unfortunately, XAI raises serious privacy concerns. Luo et al. [7] showed that private input record values can be inferred from an explanation based on Shapley values. They proposed an algorithm that estimates attribute values using a gradient-descent method with a sampled auxiliary dataset. This enables private records to be reconstructed with a significant probability of success. They demonstrated the feasibility of attacks using six open datasets plus three synthetic datasets and four black-box AI models.

However, reconstruction accuracy depends on the black-box models used, such as support vector machines (SVMs) or decision trees. The risk must also depend on the XAI metric used because there are many XAI technologies, including local interpretable model-agnostic explanations (LIME) [8], Shapley additive explanations (SHAP) [15], and Anchors [17]. Luo et al. investigated only the Shapley-value approach. It remains unclear if particular combinations of black-box models and XAI metrics may be particularly vulnerable to attack. This is the motivation for our investigation of the privacy risks in vulnerable combinations of black-box models and XAI metrics.

In this work, we claim that *a particular combination of black-box AI model and XAI metric is vulnerable in the context of record reconstruction risks*. We prove that a linear regression model using a Shapley-value XAI metric enables an attacker to reconstruct private records completely (no estimation error). Note that this is an inevitable vulnerability, obtained via theoretical analysis, which will hold for any kind of dataset when the exact Shapley values are given. However, in most platforms, an approximation to or a variation of the Shapley values is used because the exact computation of Shapley values suffers from an exponential computation cost with respect to the number of attributes and becomes infeasible for large-scale systems with more than 50 attributes.

To address these additional questions, we conducted experiments to investigate the reconstruction risks with respect to the differences in the algorithms used by potential attackers. We used three open datasets: Adult [9], Bank Marketing [10], and Credit Card Client [11], with Shapley values. We studied not only the black-box AI models used for training the baseline estimation but also the attack models used to estimate the private record values and the optimization algorithms used by the attack models, including SGD, Momentum, RMSProp,

and Adam. The experiments demonstrated that our theoretical result holds for common open datasets.

Our work makes the following contributions:

- We proved that the combination of a linear regression method and using Shapley values is vulnerable against a reconstruction attack, in that the attacker can reconstruct private input attributes exactly.
- We evaluated the record reconstruction risk on Shapley values with respect to the various optimization algorithms used in attacker models. Based on our experimental results, we suggest a possible mitigation approach that adjusts the learning rate with the aim of minimizing the estimation risk.

The remainder of the paper is organized as follows. In Sect. 2, we introduce some background definitions, Shapley values, and some related work. In Sect. 3, we present the problem statement and the threat model assumed in our study. We show some examples of record reconstruction attacks on a toy example, aiming to provide insight into attack methods. Our main theorem is given in Sect. 4, where we prove that using Shapley values with a linear regression model allows an attacker to identify the private inputs without error. In Sect. 5, we describe our experiments using open data with various parameter settings. We discuss the generalization of our methodology and its limitations in Sect. 6. We also suggest a possible approach to mitigating the effects of reconstruction attacks. We conclude our work in Sect. 7.

## 2 Preliminaries

### 2.1 Shapley Values

Shapley values [12], proposed by Shapley in 1953, are indices that quantify the contributions of each player in cooperative game theory. In this work, let  $\mathbf{s} = (s_1, \dots, s_n)$  be the Shapley values representing the local explanation for the output of model  $f(\mathbf{x})$  for  $n$  input features  $\mathbf{x} = (x_1, \dots, x_n)$ .

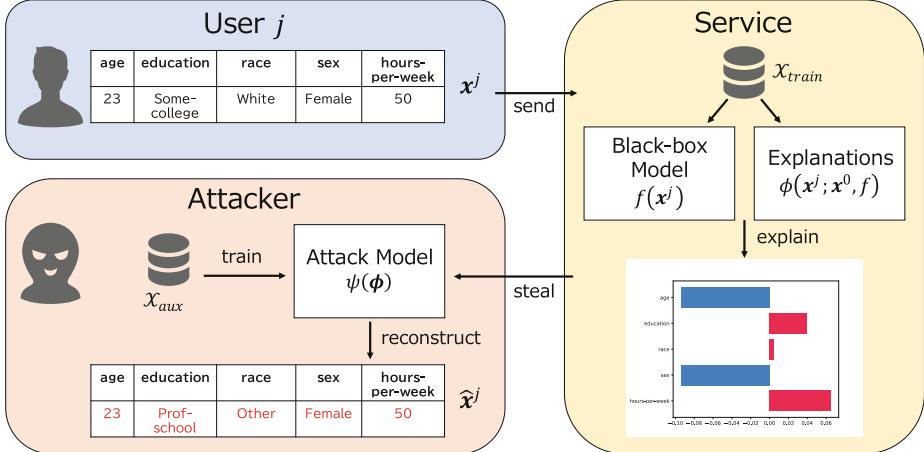
Let  $N = \{1, 2, \dots, n\}$  be the index set of features,  $S$  be a subset of  $N$ ,  $\mathbf{x}^0$  be a reference sample for calculating Shapley values, and  $\phi(\mathbf{x}; \mathbf{x}^0, f) = (s_1, \dots, s_n)$  be a mapping to calculate Shapley values. Then, the Shapley value  $s_i$  is:

$$s_i = \phi_i(\mathbf{x}; \mathbf{x}^0, f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (f(\mathbf{x}_{[S \cup \{i\}]}) - f(\mathbf{x}_{[S]})), \quad (1)$$

where  $\mathbf{x}_{[S]} = ((x_{[S]})_1, \dots, (x_{[S]})_n)$  denotes an input vector corresponding to  $S$  and is defined for  $i = 1, \dots, n$  as:

$$(x_{[S]})_i = \begin{cases} x_i & \text{if } i \in S, \\ x_i^0 & \text{otherwise.} \end{cases} \quad (2)$$

For example, given the input vector  $\mathbf{x} = (\mathbf{2}, \mathbf{5}, \mathbf{1}, \mathbf{3})$ , the reference sample  $\mathbf{x}^0 = (0, 3, 2, 1)$ , and the subset  $S = \{2, 3\}$ , the vector  $\mathbf{x}_{[S]}$  is  $\mathbf{x}_{[S]} = (0, \mathbf{5}, \mathbf{1}, 1)$ .



**Fig. 1.** Overview of the system model

## 2.2 Feature Inference Attack on Shapley Values

**System Model.** Luo et al. [7] assumed a system model by which the service provider trains their black-box models  $f$  based on a private training dataset  $\mathcal{X}_{train}$  and uses them for their service provided on MLaaS platforms. Figure 1 shows the overview of their system model. The **Attacker** has access to the **Service** with its store of explanation for target **User  $j$** . Given the explanation  $\phi(\mathbf{x}^j; \mathbf{x}^0, f)$ , the **Attacker** attempts to reconstruct the original record  $\mathbf{x}^j$  using the attack model  $\phi(\phi)$ .

**Threat Model.** An attacker can send data to the service and receive the Shapley values as an explanation of the data. In addition, the attacker can steal the Shapley values of other users from the service. Under this assumption, the attacker performs the feature inference.

## 2.3 Related Work

**Explainable AI Metrics.** There are many ways to explain black-box and white-box models. Explainabilities can be partitioned into global and local methods. Global methods explain overall model behavior and compute feature importance values [13, 14], whereas local methods explain the feature importance for each input [8, 15–17]. Major MLaaS platforms including Amazon SageMaker [5], Microsoft Azure [6], and Google Cloud Platform [18] offer SHAP [15], an approximation to Shapley values [12, 16] using local methods such as LIME [8].

Our study therefore focuses on Shapley values [12, 15, 16] as local explainabilities.

**Privacy Risk with Explainability.** Many previous studies have investigated the privacy risk with explainabilities against a variety of attacks, including membership inference [19–21], model extraction [19, 22, 23], feature inference [7, 24], and adversarial attack [19, 25]. In particular, Luo et al. [7] identified a feature inference attack with Shapley values and investigated the privacy risk from the explanations.

**Defenses Against Attack.** Some defensive approaches have been proposed, including differential privacy [29–31] and using synthetic data [29, 32, 33]. In addition, several studies have proposed privacy-preserving XAI methods [26–28]. For example, Patel et al. [26] proposed using a local explainability method such as LIME with differential privacy. Nevertheless, the privacy risk in the context of feature inference attacks remains unclear for explainabilities other than using Shapley values.

### 3 Problem Statement

#### 3.1 Record Reconstruction Attack

We can define a record reconstruction attack (or a feature inference attack) [7] as follows.

Let  $f$  and  $\psi$  be black-box and attack models, respectively. Let  $\mathcal{X}_{train}$ ,  $\mathcal{X}_{aux}$ , and  $\mathcal{X}_{test}$  be a training dataset for training  $f$ , an auxiliary dataset for training  $\psi$ , and a test dataset, respectively. Let  $\mathbf{x}^j = (x_1^j, \dots, x_n^j)$  be an input vector for the user  $j = 1, \dots, m$ . Let  $\mathbf{x}^0$  and  $\mathbf{s}^j = \phi(\mathbf{x}^j; \mathbf{x}^0, f)$  be the reference sample and the Shapley values of the input vector  $x^j$ , respectively. Given the explanation dataset  $\mathcal{S}_{aux}$  for all  $\mathbf{x}_{aux} \in \mathcal{X}_{aux}$  and the black-box model  $f$ , the attacker trains the attack model  $\psi : \mathcal{S}_{aux} \rightarrow \mathcal{X}_{aux}$ .

#### 3.2 Evaluation Metrics

We use two metrics for evaluating the record reconstruction risk, the attacker’s mean absolute error (MAE) and the success rate (SR).

**Attacker’s MAE.** The MAE is the average of a set of absolute errors. The attacker’s MAE of the estimated data  $\hat{\mathbf{x}}$  for the dataset  $\mathbf{x}$  with  $m$  rows and  $n$  columns is given as:

$$MAE(\hat{\mathbf{x}}, \mathbf{x}) = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n |\hat{x}_i^j - x_i^j|. \quad (3)$$

**Attacker’s SR.** The SR represents the ratio of correctly estimated features to all input features. We say that a feature estimation is a success if the estimated value is identical to the original values for discrete variables. For continuous variables, the estimation is a success if the absolute error is below a particular threshold value. The SR of the estimated data  $\hat{\mathbf{x}}$  for the dataset  $\mathbf{x}$  is given as:

$$SR(\hat{\mathbf{x}}, \mathbf{x}) = \frac{success(\hat{\mathbf{x}}, \mathbf{x})}{mn}, \quad (4)$$

where  $success(\hat{\mathbf{x}}, \mathbf{x})$  is the number of correctly estimated features.

## 4 The Record Reconstruction Risk with Linear Regression

In this section, we show the vulnerability of using the combination of a linear model and the Shapley values. First, we provide a simple example to give insight into the vulnerability.

### 4.1 Example of a Record Reconstruction Attack

Consider an example dataset of 10 rows and 5 columns. We synthesize the dataset  $x_1 = n_1$ ,  $x_2 = n_2$ ,  $x_3 = n_1 n_2$ ,  $x_4 = n_2 n_3$ , and  $y = x_1 - x_3 x_4$  using three independent identically distributed sequences  $n_1$ ,  $n_2$ , and  $n_3$ , as shown in Table 1.

Let  $\mathcal{X}_{test}$  and  $\mathcal{X}_{train}$  comprise rows 1 to 5 and rows 6 to 10 of the dataset, respectively. The Shapley value is the average of the values with respect to each row of  $\mathcal{X}_{train}$  in the reference sample, i.e.,  $s = \frac{1}{5} \sum_{j=6}^{10} \phi(\mathbf{x}; \mathbf{x}^j, f)$ .

Table 2 shows the Shapley values  $S_{test}$  for the input dataset  $\mathcal{X}_{test}$  with the linear black-box model  $f$  trained for the dataset  $\mathcal{X}_{train}$  in the example.

The attacker’s MAEs for the model  $f$  and the estimation algorithm  $\psi$  are then shown in Table 3.  $f$  and  $\psi$  are either a linear regression or a decision tree. They are implemented via scikit-learn and trained on  $\mathcal{X}_{train}$ . Note that the input features are correctly estimated without error when  $f$  and  $\psi$  are both linear models.

### 4.2 Theoretical Analysis of MAE

According to Eq. (1), the Shapley values are computed as the weighted average of the difference of two outputs  $f(\mathbf{x}_{[S \cup \{i\}]}) - f(\mathbf{x}_{[S]})$ . First, we show the linearity of this difference.

**Lemma 1.** Let  $f$  be a linear black-box model. For any  $i \in N$ ,  $S \subseteq N \setminus \{i\}$ , and reference sample  $(x_1^0, x_2^0, \dots, x_n^0)$ , the following holds:

$$f(\mathbf{x}_{[S \cup \{i\}]}) - f(\mathbf{x}_{[S]}) = \beta_i(x_i - x_i^0) \quad (5)$$

**Proposition 1.** Let  $f$  and  $\psi$  be a linear black-box model and a linear attack model, respectively. When  $n < |\mathcal{X}_{aux}|$ , the attacker’s MAE with  $\psi$  is 0.

This result implies that a record reconstruction could be perfect if the Shapley values from the linear model were to be used.

**Table 1.** Example dataset

	$x_1$	$x_2$	$x_3$	$x_4$	$y$
$\mathcal{X}_{test}$	1.8	0.1	0.3	-0.4	1.9
	0.4	1.5	0.6	1.0	-0.2
	1.0	0.8	0.7	0.7	0.5
	2.2	0.1	0.3	-0.1	2.2
	1.9	0.4	0.8	1.0	1.1
$\mathcal{X}_{train}$ ( $\mathcal{X}_{aux}$ )	-1.0	0.3	-0.3	-0.5	-1.2
	1.0	1.5	1.4	0.1	0.9
	-0.2	-0.2	0.0	0.0	-0.2
	-0.1	0.3	0.0	0.5	-0.1
	0.4	-0.9	-0.4	-1.3	-0.1

**Table 2.** Shapley values  $s_i \in \mathcal{S}_{test}$  when  $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4$ , and  $\mathbf{x}^5$  represent the reference sample

	$s_1$	$s_2$	$s_3$	$s_4$
$\mathbf{x}^1$	1.30	0.02	0.06	-0.04
$\mathbf{x}^2$	0.28	-0.29	0.18	0.34
$\mathbf{x}^3$	0.72	-0.13	0.21	0.26
$\mathbf{x}^4$	1.59	0.02	0.06	0.04
$\mathbf{x}^5$	1.37	-0.04	0.25	0.34

**Table 3.** The attacker’s MAEs for the combination of black-box model  $f$  and attack model  $\psi$ 

Black-box $f$	Attack $\psi$	Attacker’s MAE				
		$x_1$	$x_2$	$x_3$	$x_4$	average
Linear Regression	Linear Regression	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>
Linear Regression	Decision Tree	0.82	1.24	0.74	1.18	1.00
Decision Tree	Linear Regression	0.69	0.52	0.41	0.53	0.54
Decision Tree	Decision Tree	0.68	1.16	0.82	0.54	0.80
average		0.55	0.73	0.49	0.59	

## 5 Experiments

### 5.1 Dataset

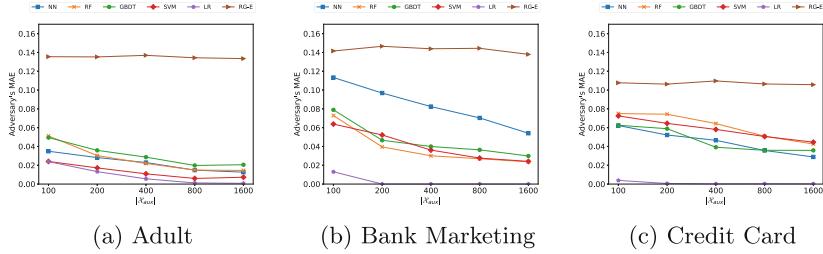
Table 4 shows the dataset used for the experiments. For these settings, we aim to clarify the record reconstruction risk of XAI metrics and the differences between the optimizers used in the attack algorithm.

### 5.2 Experiment 1: Risk with Respect to Black-Box Models and XAI Methods

We evaluated the record reconstruction risk for the explanation of black-box model  $f$  given by Shapley values. The black-box model  $f$  was one of five models: neural networks (NN), random forest (RF), gradient-boosting decision tree (GBDT), kernel SVM (kSVM), and linear regression (LR).

**Table 4.** Three open datasets used in the experiments

Dataset	No. of Records	Classes	No. of Features
Adult [9]	48,842	2	14
Bank Marketing [10]	45,211	2	16
Credit Card [11]	30,000	2	24



**Fig. 2.** The attacker’s MAEs for a record reconstruction attack using Shapley values with respect to the number of rows in the auxiliary dataset  $|X_{aux}|$  and the black-box model  $f$

We implemented NN using PyTorch [34] using an  $n$ -dimensional input layer, a  $c$ -dimensional output layer, and two hidden layers of  $2n$  neurons each. The activation function was softmax for the output layer and rectifier linear unit (ReLU) for the remainder. The other models, RF, SVM, GBDT, and LR, were implemented via scikit-learn. The number of trees and the maximal depth were 100 and 5, respectively, for RF, with those for GBDT being 100 and 3, respectively. We used default values for other parameters if not specifically mentioned. By “RG-E”, we denote a random guess prediction from the empirical distribution based on  $X_{aux}$  for comparison purposes.

### 5.3 Experiment 2: Risk with Respect to Optimization Algorithms

We investigated the record reconstruction risk of the various optimization algorithms that could be used by a potential attacker. An optimizer would be used to update the parameter  $\theta_\psi$  of the attack model  $\psi$  as:

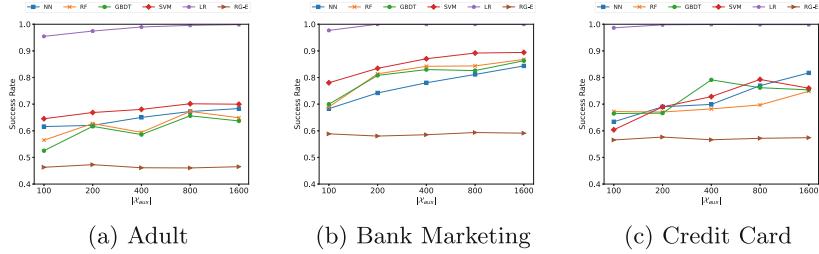
$$\theta_\psi \leftarrow \theta_\psi - \alpha \nabla_{\theta_\psi} \text{loss} \quad (6)$$

In this experiment, we investigated four types of optimization algorithms including SGD [35], Momentum [36], RMSProp [37], and Adam [38]. Using the settings in Luo et al. [7], we examined the attack model  $\psi$  deployed in neural networks with  $4n$  neurons in the hidden layer and  $n$  neurons in the output layer for  $n$  features, with the sigmoid function being used for all layers. We implemented model  $\psi$  using PyTorch, using default values for all parameters except  $\eta = 0.01$  for the learning rates of SGD and Momentum and  $momentum = 0.9$  for Momentum. (Momentum refers to an SGD for which  $momentum \neq 0$ ).

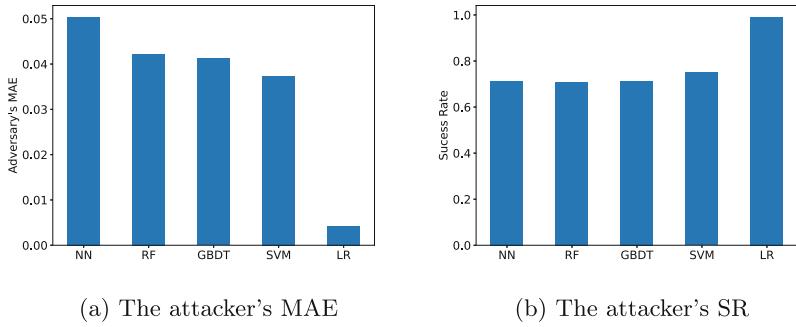
### 5.4 Results

Figures 2 and 3 show the record reconstruction risk for Shapley values with respect to the number of rows of the auxiliary dataset  $X_{aux}$ .

We found that the attacker’s MAE decreases and the SR increases as the number of rows  $|X_{aux}|$  increases. Note that the record reconstruction risks for



**Fig. 3.** The attacker’s SRs for a record reconstruction attack using Shapley values with respect to the number of rows in the auxiliary dataset  $|\mathcal{X}_{aux}|$  and the black-box model  $f$

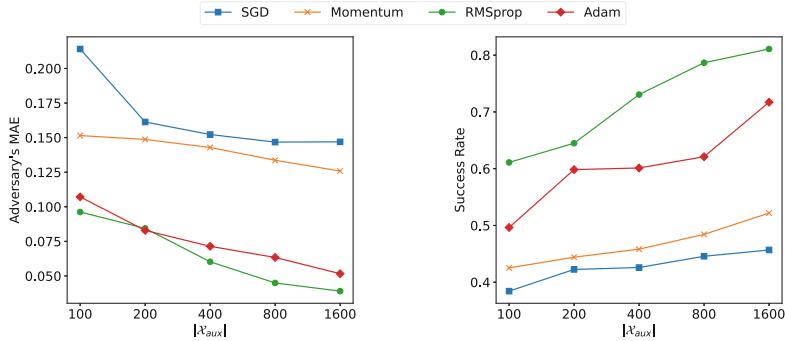


**Fig. 4.** The evaluation averages for each model  $f$

Shapley values are small but not zero when  $|\mathcal{X}|$  is small for (a) Adult and (b) Bank scenarios. These results appear to be in conflict with Proposition 1, which states that there should be no errors for a linear black-box model. However, these small errors may remain because the conditions in Proposition 1 do not hold here. Another possible reason is that the Shapley values were not calculated exactly as defined but approximated [15, 16]. Note also that the small MAE increases in proportion to the size of the auxiliary dataset.

Figure 4 shows the average of the attacker’s MAE and SR for each model  $f$  and dataset size  $|\mathcal{X}_{aux}|$  for an attack based on Shapley values. The linear model had the highest record reconstruction risk in terms of both the attacker’s MAE and SR.

Figure 5 shows the distributions of the attacker’s MAE and SR for the various optimization algorithms used in an attack. In common with the attacker’s MAE and SR, the accuracy of record reconstruction using SGD is the least and, with RMSProp, it is greatest. For all optimization algorithms, the attacker’s MAE decreased and SR increased with  $|\mathcal{X}_{aux}|$ . Adam and RMSProp tended to increase the record reconstruction risk, whereas SGD and Momentum tended to decrease it. It is known that Adam and RMSProp are methods that tune the learning rate when training the model. Therefore, it was tuning the learning rate that improved the accuracy of the record reconstruction.



**Fig. 5.** The attacker’s MAE and SR with respect to the size of dataset  $|X_{aux}|$  for the four types of optimizer available to an attacker

## 6 Discussion

### 6.1 Differences Between the Theorem and the Experiments

In Proposition 1, we proved that an attacker can reconstruct the original input features correctly when both the black-box model  $f$  and the attack model  $\psi$  are linear. However, our experimental results showed small errors remaining. We consider that the lack of a sufficient number of instances in some smaller datasets may be the source of the error. The premise of the proposition would then not hold and the linear regression would fail to estimate the exact values. We also note that using Shapley values involves approximations and a reconstruction attack using XAI values might therefore include few mistakes. We plan to explore these conjectures in future work.

### 6.2 Linearity of the Black-Box Models

The proof of Proposition 1 was based on the assumption of linearity in the black-box model  $f$ . The reconstruction risk could therefore be nonzero if the black-box model  $f$  is not linear. However, we believe that there is a positive correlation between the reconstruction risk and the additivity of the explanations. In 2017, Lundberg et al. [15] proposed a class of additive feature attribution methods that included methods based on Shapley values. There are several methods belonging to the class, except for Shapley values, but a property called “local accuracy” is not fully satisfied in these cases. We therefore believe that the property of local accuracy may be a key feature of the vulnerability when using Shapley values.

### 6.3 Effects of Encoding Methods for Qualitative Variables

In our experiments, we encoded the qualitative variables into discrete values using one-hot encoding. This could lead to a dimensionality issue. It is possible that, as the number of features increases and the accuracy of the approximations

in Shapley values decreases, this could result in a higher reconstruction risk than the original risk. We are also concerned that an attacker could gain more information from the explanations because the dimensionality of the explanation vector would be higher than that for the original. Again, we plan to conduct further experiments using other encoding methods.

#### 6.4 Optimization for Black-Box Models

In our experiments, we investigated the privacy risk with respect to optimization algorithms used in the attack model. It is possible that the optimizer in the black-box model affects the reconstruction risk when the black-box model is a neural network. We might consider that the privacy risk would be higher if the accuracy of the black-box model were increased.

#### 6.5 Mitigation

To decrease the record reconstruction risk, we suggest three defensive methods.

First, we could use one of several privacy-enhancing technologies, such as using synthetic data to train the black-box model or using differential privacy for the explanation values. In 2022, Patel et al. [26] proposed explainability via differential privacy. Using a privacy-preserving method should reduce the reconstruction risk.

Second, the access control on MLaaS platforms could be made more efficient. For example, a limitation on the number of requests could reduce the amount of information available to a potential attacker.

Finally, the quantization and masking of Shapley values could be useful ways of decreasing the record reconstruction risk.

### 7 Conclusion

We have examined the record reconstruction risk of XAI methods based on Shapley values using the attack algorithm of Luo et al. [7]. We also found that learning-rate tuning in the optimization algorithms used by an attacker increases the privacy risk, particularly for Adam and RMSProp optimizers. Using Shapley values can enable the exact reconstruction of private inputs when both black-box and attack models are linear.

To mitigate these risks, we recommend using synthetic data for training black-box models and applying differential privacy to explanation values. Limiting requests and access on MLaaS platforms could also enhance privacy protection.

In future work, we aim to explore the reconstruction risk with explanation vectors that use additive noise and develop new XAI methods to further reduce privacy risks.

**Acknowledgement.** Part of this work was supported by JSPS KAKENHI Grant Number 23K11110 and JST, CREST Grant Number JPMJCR21M1, Japan.

## A Algorithm in Luo et al. [7]

In the model investigated by Luo et al. [7], a user  $j$  sends a private input vector  $\mathbf{x}^j$  to a service and receives the output of the model  $f(\mathbf{x}^j)$  and explanations about  $n$  features  $\mathbf{s} = \phi(\mathbf{x}^j; \mathbf{x}^0, f) = (s_1, \dots, s_n)$ . The attacker has an auxiliary dataset  $\mathcal{X}_{aux}$ , which is distributed similarly to the training dataset  $\mathcal{X}_{train}$ . The attacker sends  $\mathbf{x}_{aux} \in \mathcal{X}_{aux}$  to the model  $f$  and receives the explanation dataset  $\mathcal{S}_{aux} = \{\phi(\mathbf{x}_{aux}; \mathbf{x}^0, f) | \mathbf{x}_{aux} \in \mathcal{X}_{aux}\}$ . The attacker then trains the attack model  $\psi : \mathcal{S}_{aux} \rightarrow \mathcal{X}_{aux}$  to minimize the loss  $L(\psi(\mathcal{S}_{aux}), \mathcal{X}_{aux})$ . Finally, the attacker estimates the original private input features  $\mathbf{x}^j$  as  $\hat{\mathbf{x}}^j = \psi(\mathbf{s})$  with the given Shapley values  $\mathbf{s}$ . This is described formally in Algorithm 1.

---

**Algorithm 1.** Estimating algorithm using the auxiliary dataset [7]

---

**Input:** Black-box model  $f$ , auxiliary dataset  $\mathcal{X}_{aux}$ , learning rate  $\alpha$ , attacked Shapley vector  $\mathbf{s}$

**Output:** Estimated private input features  $\hat{\mathbf{x}}$

- 1:  $\mathcal{S}_{aux} \leftarrow \phi(\mathcal{X}_{aux}; f)$
- 2:  $\theta_\psi \leftarrow \mathcal{N}(0, 1)$
- 3: **for** each epoch **do**
- 4:     **for** each batch **do**
- 5:          $loss \leftarrow 0$
- 6:          $B \leftarrow$  randomly select a batch of samples
- 7:         **for**  $j \in 1, \dots, |B|$  **do**
- 8:              $(\hat{\mathbf{x}}_{aux})^j \leftarrow \psi((\mathbf{s}_{aux})^j; \theta_\psi)$
- 9:              $loss \leftarrow loss + L((\hat{\mathbf{x}}_{aux})^j, (\mathbf{x}_{aux})^j)$
- 10:         **end for**
- 11:          $\theta_\psi' \leftarrow \theta_\psi - \alpha \nabla_{\theta_\psi} loss$
- 12:     **end for**
- 13: **end for**
- 14:  $\hat{\mathbf{x}} \leftarrow \psi(\mathbf{s}; \theta_\psi)$
- 15: **return**  $\hat{\mathbf{x}}$

---

## B Proofs

### B.1 Proof of Lemma 1

*Proof.* Denoting the model  $f$  as  $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$ , for any subset  $S$  and element  $i$ ,

$$\begin{aligned} f(\mathbf{x}_{[S \cup \{i\}]}) - f(\mathbf{x}_{[S]}) &= \beta_0 + \sum_{k \in S \cup \{i\}} \beta_k x_k + \sum_{k \in N \setminus (S \cup \{i\})} \beta_k x_k^0 \\ &\quad - \beta_0 + \sum_{k \in S} \beta_k x_k + \sum_{k \in N \setminus S} \beta_k x_k^0 \\ &= \beta_i (x_i - x_i^0). \end{aligned}$$

## B.2 Proof of Proposition 1

*Proof.* From Lemma 1, the  $i$ -th Shapley value  $s_i$  is calculated as follows:

$$\begin{aligned} s_i &= \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} f(\mathbf{x}_{[S \cup \{i\}]}) - f(\mathbf{x}_{[S]}) \\ &= \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \beta_i(x_i - x_i^0) \\ &= \lambda_i(x_i - x_i^0) \end{aligned}$$

where  $\lambda_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \beta_i$ . Therefore, the attack model  $\psi$  is given as the linear equations of  $x_1, \dots, x_n$  as follows:

$$\begin{aligned} \hat{x}_i &= \alpha_0 + \alpha_1 s_1 + \dots + \alpha_n s_n \\ &= \alpha_0 + \alpha_1 (\lambda_1(x_1 - x_1^0)) + \dots + \alpha_n (\lambda_n(x_n - x_n^0)) \\ &= \alpha_0 - \sum_{k=1}^n \alpha_k \lambda_k x_k^0 + \alpha_1 \lambda_1 x_1 + \dots + \alpha_n \lambda_n x_n \\ &= \gamma_0 + \gamma_1 x_1 + \dots + \gamma_n x_n \end{aligned}$$

where  $\gamma_i = \alpha_i \lambda_i$  and  $\gamma_0 = \alpha_0 - \sum_{k=1}^n \alpha_k \lambda_k x_k^0$ . Note that this is a linear polynomial of  $x_0, \dots, x_n$ . If  $\mathcal{X}_{aux}$  is large enough and the number of rows exceeds  $n+1$ , the coefficients  $\gamma_1, \dots, \gamma_n$  are correctly estimated by the least squares method. A linear regression will give the exact input variables, proving the proposition.

## References

1. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019)
2. Zest AI Insights. <https://www.zest.ai/insights/why-zaml-makes-your-ml-platform-better>. Accessed 19 Apr 2024
3. Sakai, A., et al.: Medical professional enhancement using explainable artificial intelligence in fetal cardiac ultrasound screening. *Biomedicines* **10**(3), 551 (2022)
4. Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to explain: an information-theoretic perspective on model interpretation. In: 35th International Conference on Machine Learning, Stockholm, Sweden, pp. 882–891. PMLR (2018)
5. Amazon SageMaker Documentation. <https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html>. Accessed 19 Apr 2024
6. Azure Machine Learning Documentation. <https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability>. Accessed 19 Apr 2024
7. Luo, X., Jiang, Y., Xiao, X.: Feature inference attack on Shapley values. In: 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS 2022), Los Angeles, CA, USA, pp. 2233–2247. Association for Computing Machinery (2022)

8. Ribeiro, M., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), San Francisco, California, USA, pp. 1135–1144. Association for Computing Machinery (2016)
9. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996). <https://doi.org/10.24432/C5XW20>
10. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.* **62**, 22–31 (2014)
11. Yeh, I., Lien, C.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.* **36**(2), 2473–2480 (2009)
12. Shapley, L.: 17. A value for n-person games. *Contrib. Theory Games (AM-28)* **II**, 307–318 (1953)
13. Covert, I., Lundberg, S., Lee, S.: Understanding global feature contributions with additive importance measures. In: 34th International Conference on Neural Information Processing Systems (NIPS 2020), Vancouver, BC, Canada, pp. 17212–17223. Curran Associates Inc. (2020)
14. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**(177), 1–81 (2019)
15. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. In: 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, California, USA, pp. 4768–4777. Curran Associates Inc. (2017)
16. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2016). <https://doi.org/10.1007/s10115-013-0679-x>
17. Ribeiro, M., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI 2018/IAAI 2018/EAAI 2018), New Orleans, Louisiana, USA, pp. 1527–1535. AAAI Press (2018)
18. Introduction to Vertex Explainable AI. <https://cloud.google.com/vertex-ai/docs/explainable-ai/overview>. Accessed 19 Apr 2024
19. Kuppa, A., Le-Khac, N.: Adversarial XAI methods in cybersecurity. *IEEE Trans. Inf. Forensics Secur.* **16**, 4924–4938 (2021)
20. Shokri, R., Strobel, M., Zick, Y.: On the privacy risks of model explanations. In: 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2021), pp. 231–241. Association for Computing Machinery, Virtual Event, USA (2021)
21. Liu, H., Wu, Y., Yu, Z., Zhang, N.: Please tell me more: privacy impact of explainability through the lens of membership inference attack. In: 2024 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, pp. 119–138. IEEE Computer Society (2024)
22. Yan, A., Hou, R., Liu, X., Yan, H., Huang, T., Wang, X.: Towards explainable model extraction attacks. *Int. J. Intell. Syst.* **37**(11), 9936–9956 (2022)
23. Yan, A., Huang, T., Ke, L., Liu, X., Chen, Q., Dong, C.: Explanation leaks: explanation-guided model extraction attacks. *Inf. Sci. Int. J.* **632**(C), 269–284 (2023)
24. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: 22nd ACM SIGSAC Conference

- on Computer and Communications Security (CCS 2015), Denver, Colorado, USA, pp. 1322–1333. Association for Computing Machinery (2015)
- 25. Baniecki, H., Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: a survey. *Inf. Fusion* **107**, 102303 (2024)
  - 26. Patel, N., Shokri, R., Zick, Y.: Model explanations with differential privacy. In: 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022), Seoul, Republic of Korea, pp. 1895–1904. Association for Computing Machinery (2022)
  - 27. Nguyen, T., Lai, P., Phan, H., Thai, M.: XRand: differentially private defense against explanation-guided attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 10, pp. 11873–11881 (2023)
  - 28. Bozorgpanah, A., Torra, V., Aliahmadipour, L.: Privacy and explainability: the effects of data protection on Shapley values. *Technologies* **10**(6), 125 (2022)
  - 29. Shlomo, N.: Integrating differential privacy in the statistical disclosure control toolkit for synthetic data production. In: Domingo-Ferrer, J., Muralidhar, K. (eds.) PSD 2020. LNCS, vol. 12276, pp. 271–280. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-57521-2\\_19](https://doi.org/10.1007/978-3-030-57521-2_19)
  - 30. Wang, G., Gehrke, J., Xiao, X.: Differential privacy via wavelet transforms. *IEEE Trans. Knowl. Data Eng.* **23**(8), 1200–1214 (2011)
  - 31. Ito, S., Miura, T., Akatsuka, H., Terada, M.: Differential privacy and its applicability for official statistics in japan – a comparative study using small area data from the Japanese population census. In: Domingo-Ferrer, J., Muralidhar, K. (eds.) PSD 2020. LNCS, vol. 12276, pp. 337–352. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-57521-2\\_24](https://doi.org/10.1007/978-3-030-57521-2_24)
  - 32. Slokom, M., Wolf, P., Larson, M.: When machine learning models leak: an exploration of synthetic training data. In: Domingo-Ferrer, J., Laurent, M. (eds.) PSD 2022. LNCS, vol. 13463, pp. 283–296. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_20](https://doi.org/10.1007/978-3-031-13945-1_20)
  - 33. Tritscher, J., Ring, M., Schlr, D., Hettinger, L., Hotho, A.: Evaluation of post-hoc XAI approaches through synthetic tabular data. In: Helic, D., Leitner, G., Stettinger, M., Felfernig, A., Raš, Z.W. (eds.) ISMIS 2020. LNCS (LNAI), vol. 12117, pp. 422–430. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59491-6\\_40](https://doi.org/10.1007/978-3-030-59491-6_40)
  - 34. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, pp. 8026–8037. Curran Associates Inc. (2019)
  - 35. Bottou, L.: On-line learning and stochastic approximations. *On-Line Learn. Neural Netw.* 9–42 (1999)
  - 36. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: 30th International Conference on Machine Learning (ICML 2013), Atlanta, GA, USA, pp. 1139–1147. JMLR.org (2013)
  - 37. Hinton, G.: Coursera Neural Networks for Machine Learning Lecture 6. [https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture\\_slides\\_lec6.pdf](https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf). Accessed 19 Apr 2024
  - 38. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)



# DISCOLEAF: Personalized DISecretization of COntinuous Attributes for LEArning with Federated Decision Trees

Saloni Kwatra<sup>(✉)</sup> and Vicenç Torra

Department of Computing Science, Umeå University, Umeå, Sweden  
[{salonik,vtorra}@cs.umu.se](mailto:{salonik,vtorra}@cs.umu.se)

**Abstract.** Federated learning is a distributed machine learning framework, in which each client participating to the federation trains a machine learning model on its data, and shares the trained model information with a central server, which aggregates, and sends the aggregated information back to the distributed clients. The machine learning model we choose to work with is decision trees, due to their simplicity, and interpretability. On that note, we propose a full-fledged federated pipeline, which includes discretization and learning with decision trees for the horizontally partitioned data. Our federated discretization approach can be plugged-in as a prepossessing step before any other federated learning algorithm. During discretization, we ensure that each client creates the number of discrete bins, according to their own data/choice. Hence, our approach is both federated and personalized. After discretization, we propose to apply the post randomization method to protect the discretized data with differential privacy guarantees. After protecting its database, each client trains a decision tree classifier on its protected database locally, and shares the nodes, containing the split attribute, and the split value with the central server. The central server obtains the most occurred split attribute, and combines the split values. This process goes on until all the nodes to be merged are leaf nodes. The central server then shares the merged tree with the distributed clients. Hence, our proposed framework performs personalized, privacy-preserving federated learning with decision trees by discretizing continuous attributes, and then masking them prior to the training stage. We call our proposed framework DISCOLEAF.

**Keywords:** Federated Learning · Discretization · Decision Trees · Differential Privacy · Decision Tree Aggregation

## 1 Introduction

FL is a setup in which distributed clients learn a machine learning model together without exposing their private data. FL enables collaborative learning, as each

client shares the parameters of the model trained with the central server (if it is present), or with other clients (if it is a peer-to-peer architecture) [13]. It is quite clear by now that sharing of information about models, like gradients for neural networks [23], split attributes, and split values for decision trees leak information about the data [6], which means clients must apply some protection methods on their database or add some privacy while training the machine learning models to preserve their privacy. In this work, we are interested in decision trees, as they are explainable in nature, and therefore can be deployed in practical domains, like medical, and finance. Training ensemble methods such as Random Forest and Gradient Boosting Decision Trees on the client side can lead to more accurate results. However, this comes at the cost of reducing model interpretability, making it difficult to understand how the model is making its predictions. Hence, we propose a federated framework for decision trees. This work is an inspiration from our previous work on federated learning with decision trees for the horizontally partitioned data, where each client trains a decision tree locally, and the central server merges the decision trees trained locally [9]. A key component of our framework is to preserve the privacy of users. To achieve this goal, each client discretizes their data first, and then apply Post Randomization Masking (PRAM) method before training the decision trees. We call our proposed framework DISCOLEAF. We use a federated discretization approach in our framework, followed by PRAM to generate a protected version of discretized data, and then the clients train decision trees on their protected data. As, each client discretizes their data in our solution, and we know that the process of discretization does not offer any formal privacy guarantees, we propose to apply PRAM after the discretization process. The key characteristic of our proposed framework is that it is a federated pipeline from start to end, even the discretization process is federated, so that we utilize the distribution of other clients to facilitate the FL. According to the survey paper [22], for decision tree-based federated learning, many factors, such as privacy, accuracy, and efficiency need to be considered. Our proposed solution aims to align with all these factors. We list the contributions of this paper as follows.

- We propose a privacy-preserving federated learning framework with Decision Trees for the horizontally partitioned data, DISCOLEAF.
- We propose a federated discretization approach in our work. This is the first paper to deal with this problem.
- We provide a solution that offers a sweet spot for two trade-offs, one is about personalization, and collaboration in FL, and the other is about privacy and utility.
- We utilise PRAM, which provides Local Differential Privacy (LDP) to the client’s data.

The subsequent sections of this paper are organized as follows: Sect. 2 reviews essential concepts, including EWD, PRAM, and decision tree aggregation. Section 3 elaborates on our contributions. In particular, we discuss our proposed framework, and its components, which include federated discretization. Section 4 discusses the data we use for our experiments and our experimentation settings.

Section 5 presents and discusses the results, and Sect. 6 concludes the paper with insights into future directions.

## 2 Background and Related Work

In this section, we explain all the relevant background theories needed to understand our proposed framework, DISCOLEAF.

### 2.1 Equal Width Discretization

By discretizing features, the complexity of the data is reduced, leading to a faster model training and inference [7]. Moreover, this transformation ensures compatibility with decision tree-based algorithms, leveraging their inherent capability to handle discrete data efficiently. In our work, we use Equal Width Discretization (EWD). EWD works by finding the range of the data by subtracting the lowest value from the highest value, then we decide how many bins we want to create. We decide it using Freedman Diaconis rule, as suggested in the work [17], in which Senavirathne *et al.* did a comprehensive analysis among different unsupervised discretization techniques from the perspective of utility and privacy. Freedman Diaconis rule provides a heuristic to determine the optimum number of bins, which takes into account the size, and the variability of the dataset using the formula below.

$$\text{Number of bins (b)} = \frac{\text{Range}(\mathcal{D})}{2 \times \text{IQR}(\mathcal{D}) \times n^{-1/3}} \quad (1)$$

Here,  $n$  is the number of data points, and  $\mathcal{D}$  is the dataset. After knowing the suitable value of  $b$ , the width of each bin is calculated by dividing the range by the number of bins. Then, each data point is assigned to the appropriate bin based on its value. The process of discretization does not give us any privacy guarantees. e.g., in case of unbalanced bins produced by EWD, it may be possible that a bin has only one value. So, we need a masking method to protect the discretized dataset obtained on using EWD, which we discuss in next Subsect. 2.2.

### 2.2 Post RAndomization Masking

Discretization of attributes do not provide privacy guarantees, i.e., the process of discretization does not restrict the number of points in each bin, like  $k$ -anonymity [15, 16, 18, 19], which restricts the number of point in each equivalence class to atleast  $k$ . Therefore, after discretization, we propose to apply PRAM, which is a perturbative method for categorical attributes, and allows us to provide personalized Local Differential Privacy (LDP) [4]. More concretely, each client constructs its own Markov matrix to apply PRAM locally. Let  $C = \{c_1, c_2, c_3, \dots, c_c\}$  be a set of categories. The Markov matrix specifies transition probabilities i.e., the probability that a value remains unchanged, or is changed to any of the other  $c-1$  values. This matrix thus satisfies  $\sum_{C_j \in C} P(c_i, c_j) = 1$ . For each client, on

applying PRAM, the original local discretized database  $DisDB$  is transformed into  $ProDisDB$ .

$$P(c_i, c_j) = P(ProDisDB = c_j | DisDB = c_i) \quad (2)$$

In our work, we use invariant PRAM, which ensures that the frequencies of categories do not change after applying invariant PRAM. We denote  $T_X(c_i)$  as frequency of category  $c_i$ , and  $T_X(c_k)$  is the smallest frequency. Each value  $P_{ij}$  of the transition matrix  $P$  can be filled using the Eq. 3

$$P(c_i, c_j) = \begin{cases} 1 - \frac{\theta T_X(c_k)}{T_X(c_i)} & \text{if } i = j \text{ and} \\ \frac{\theta T_X(c_k)}{(c-1)T_X(c_i)} & \text{if } i \neq j \end{cases} \quad (3)$$

Note that,  $\theta$  equals to zero means no perturbation, and  $\theta$  equals to one means full perturbation. Hence,  $\theta$  permits a user to control the level of noise in the perturbed dataset [20].

### 2.3 PRAM and Differential Privacy

As we have seen, PRAM models the randomised response mechanism via a transition matrix  $P$ . This process satisfies LDP for a given  $\epsilon$ . From the  $\theta$  parameter, we build a transition matrix  $P$ , which satisfies LDP. It is known (see e.g. [20]) that a matrix  $P$  provides LDP with parameter  $\epsilon$  when

$$\max_{i=1}^c \max_{j=1}^c \max_{l=1}^c \frac{P(X' = c_l | c_i)}{P(X' = c_l | c_j)} \leq e^\epsilon. \quad (4)$$

Then, given a transition matrix  $P$ , we have that the PRAM satisfies LDP for all  $\epsilon$ , such that  $\epsilon \geq \epsilon_0$ , where  $\epsilon_0$  is defined by

$$\epsilon_0 = \log\left(\max_{i=1}^c \max_{j=1}^c \max_{l=1}^c \frac{P(X' = c_l | c_i)}{P(X' = c_l | c_j)}\right). \quad (5)$$

It is easy to infer that we need that all the probabilities in the matrix  $P$  should be non-zero. Otherwise,  $\epsilon_0$  is infinite, which means no privacy at all from the perspective of differential privacy [20]. Besides, it is possible to find an expression for  $\epsilon_0$  for invariant PRAM.

Given  $P$ , to find  $\epsilon_0$  means to find the  $i, j, l$  that maximize the fraction. This maximization can be achieved when numerator is maximized and denominator is minimized. While this is not possible, there are only two cases that can lead to this maximum value.

Let  $i_0$  denote the category of the largest frequency. The maximum value of the numerator is achieved when  $l = i = i_0$ . Hence, the numerator should be  $p_{i_0 i_0}$ . Then, the smallest denominator corresponds also to the largest frequency, so  $i_0$ , but not in the diagonal. So, we need to select the category with the second largest frequency, say  $i_1$ . Then, we have  $p_{i_0 i_0}/p_{i_1 i_0}$ . Alternatively, we can select the smallest denominator with the second largest numerator. That is,  $p_{i_1 i_1}/p_{i_0 i_1}$ .

Any other combination will be smaller as e.g. the category with the third largest frequency  $i_2$  will be such that their  $p_{i_2 i_2}$  is smaller (or equal) than the other two and all possible denominators are larger (or equal) than the other two. Because of that,

$$\epsilon_0 = \log \max \left( \frac{p_{i_0 i_0}}{p_{i_1 i_0}}, \frac{p_{i_1 i_1}}{p_{i_0 i_1}} \right)$$

For each  $\theta$ , representing the level of perturbation, we construct a transition matrix for a specific category using Eq. (3). From each transition matrix, we can compute the value of  $\epsilon_0$  using Eq. (5). It is an important point to note that, even for the same  $\theta$ , different variables can yield different values of  $\epsilon_0$ . This is because the construction of the transition matrix depends on both the number of categories in each variable and the frequencies of those categories. Therefore, two different variables with the same  $\theta$  can have distinct  $\epsilon_0$  values.

In our experiments, we calculate the value of  $\epsilon_0$  for each variable at various values of  $\theta$ . Conversely, one can compute  $\theta$  for a given value of  $\epsilon_0$ . Generally, higher values of  $\theta$  (perturbation level) correspond to lower  $\epsilon_0$  values. Variables in the original dataset that are highly correlated and have similar numbers of categories tend to have similar  $\epsilon_0$  values, which could aid an attacker in a re-identification attack. Our experiments demonstrate that variables with more categories tend to have higher values of  $\epsilon_0$  (indicating lower privacy) compared to variables with fewer categories. Additionally, variables containing rare categories tend to have higher  $\epsilon_0$  values. To protect rare categories and variables with a larger number of categories (such as “education-num” and “occupation” attributes in the Adult dataset), we may need to increase the perturbation level controlled by the parameter  $\theta$ . We illustrate these trends in Fig. 3 in the results Sect. 5.

## 2.4 Decision Tree Aggregation

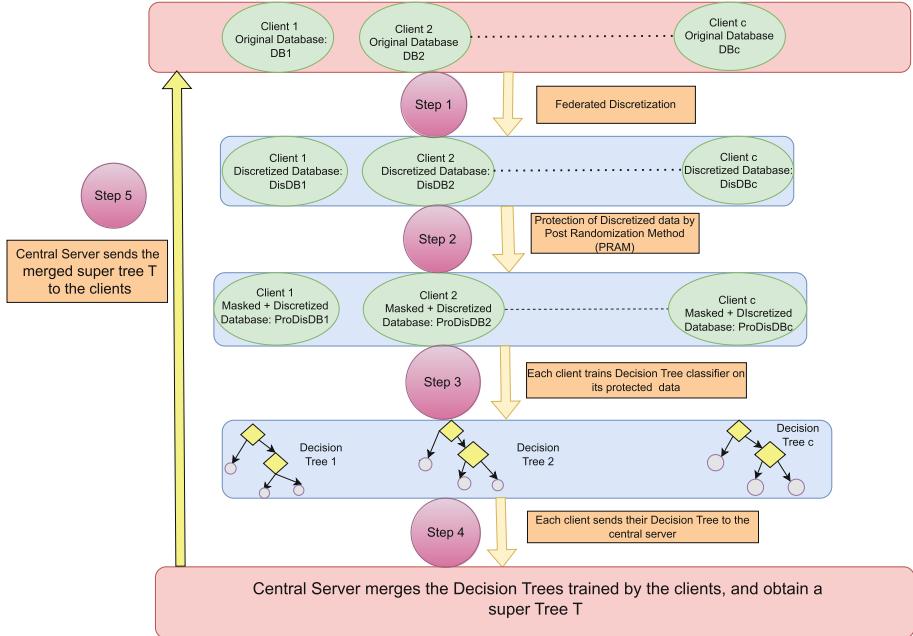
Aggregation is the core step in FL process. It is done by a central server in a client server architecture of FL. Decision trees are hard to aggregate in comparison with other numerical models, like neural networks, support vector machines, logistic regression, and others, which is why most of the existing works on decision tree-based federated learning do not utilise any algorithm to aggregate decision trees. For training Gradient Boosting Decision Trees (GBDT) in a horizontally partitioned manner, some existing approaches use hashing, particularly Locality-Sensitive Hashing (LSH) [11] to know the similar samples from the distributed clients, and then incorporate the weights of the similar samples while computing the gradients in the training process of GBDT, while others use homomorphic encryption on the gradients [3], and the decision paths [2] for training GBDT in a vertically partitioned data.

Fan *et al.* [5] proposed an approach to aggregate decision trees, and it was also discussed in the review article on tree aggregation [10] among rest other

tree aggregation approaches. Kwatra *et al.* [9] used the decision tree aggregation algorithm in the FL set up for the first time for the horizontally partitioned data. The survey on decision tree-based federated learning [22] states that in the existing literature, there is more work on vertical FL than horizontal FL for training decision trees in federated environment. Hence, our objective is also to perform federated learning with decision trees when the data across the clients is horizontally partitioned. That is why, to fulfil our objective, we use the aggregation algorithm for decision trees by Fan *et al.* [5]. This aggregation algorithm constructs the merged decision tree by identifying the most frequently occurring split attribute at each level. It determines the number of branches for the merged tree based on the split values of the most frequent attribute at each level. The merged decision tree continues to grow until all nodes from the trees being merged become leaves. For the prediction, each client traverses the merged decision tree obtained from the aggregation technique. If a client reaches a base or stopping condition (indicating that the merged tree node is a leaf node), the prediction function returns the predicted class label associated with that node. Otherwise, for the split attribute of the tree node, the prediction function compares its split value with the feature value of the test instance, and continue traversing the merged tree's children recursively until the base condition is reached.

### 3 Proposed Work

We propose a framework to perform FL with decision trees with an objective to preserve the privacy of users. We call our proposed framework, DISCOLEAF. Our approach starts by converting continuous data into categories, making it easier to analyze. The uniqueness of our approach is that each client in the federated learning setup gets to decide how many bins they want to use or each client apply Freedman Diaconis rule on its data to compute the adequate number of bins. This makes the process personalized for them. We execute the federated discretization process in a Secure Multi Party Computation (SMPC) manner. We discuss this in detail in Sect. 3.1. To ensure privacy of the discretized data, we use PRAM, which provides LDP to the client's data. Participants can choose the level of privacy protection/perturbation. Thereafter, each participant trains a decision tree on their protected data. Each client then share details about the tree with a central server, which combines all the trees from different participants to create a single model. Overall, DISCOLEAF lets participants train models privately and efficiently, making FL more secure. We depict the flow of our proposed FL framework in Fig. 1.



**Fig. 1.** Process Flow of our proposed framework DISCOLEAF

### 3.1 Federated Discretization

We refer to a paper [1], which does secure computation of the median or any  $k^{th}$  ranked element in a multi party scenario. For ease of understanding, we explain the protocol for a two-party scenario. We assume that two parties, namely  $P_1$ , and  $P_2$ , each having an input of size  $\frac{n}{2}$  aim to compute the value of the median, which is,  $\frac{n}{2}^{th}$ -ranked element, and there is one more assumption for simplicity that all input values are different. This protocol operates in multiple steps. In each step, each party computes the median value of their remaining input, and then the two parties compare their median inputs. If  $P_1$ 's median value is smaller than  $P_2$ , then  $P_1$  adjust their input by discarding the values, which are less than or equal to their median, and  $P_2$  discards the values, which are greater than their median. This protocol goes on until the length of the input is one, which is basically  $\log(n)$  steps, where  $n$  is the number of inputs. For a secure comparison, this protocol encodes the comparison function as a binary circuit, which compares the bits encoding the two inputs, and then apply Yao's protocol to it for secure computation [12].

As, such SMPC protocol [1] can be used to compute the number of any rank, we utilise it to compute the largest of all the largest (global maximum), and the smallest of all the smallest values (global minimum) from the union of the data of all the clients. This protocol is repeated for each attribute to perform the discretization of a database in a federated manner for each client. We aim to compute global maximum, and global minimum so that each client benefits from the collaboration with other clients. Once the global maximum, and global

minimum is known from the union of the data of each client, each client can have the parameter number of bins, denoted by  $b$  of its own choice. The lower the value of  $b$ , lower is the privacy. After knowing the suitable value of  $b$ , each client calculates the width of each bin by dividing the range (global maximum minus global minimum) by the number of bins, and then the client assigns its each data point to the appropriate bin based on its value. Also, this protocol is secure against a malicious adversary, who sends incorrect inputs, as their protocol verifies that the parties send consistent inputs. E.g., if a party  $P_1$  sends an input 60 for the age attribute in the adult dataset, and the output is that the party  $P_1$  should delete all the inputs less than 60, then the party can never send a number less than 60. In this manner, this protocol verifies the inputs sent by the party in each invocation. Hence, we offer personalized privacy to the clients participating in the federated discretization process using the described SMPC protocol.

## 4 Data and Experimental Settings

We perform our experiments on Adult, and Skin Segmentation datasets. Adult dataset has 32,561 records, and 14 attributes. Skin Segmentation dataset has 24,198 records, and 4 attributes. The adult dataset contains various demographic features such as age, education, occupation, marital status, race, sex, capital gain, capital loss, and hours worked per week. The target variable in this dataset is typically the income level, which is binary: “>50K” or “≤50K”. The “Skin Segmentation” dataset is used for image classification tasks. It consists of RGB (Red, Green, Blue) color values of pixels in images, where each pixel is labeled as either skin or non-skin. There is diversity in the values of attributes for the adult, and skin segmentation dataset. Hence, discretization of attributes would make sense, which is why we chose these two datasets. For the preprocessing of Adult dataset, we used `preprocessing.LabelEncoder()` from `sklearn` library in Python to convert the categorical variables, such as race, sex, native country, workclass, education, occupation, relationship, and marital status into numerical variables. We dropped the features capital gain and capital loss before the model training, as these two features don’t improve the learning of the model. Hence, they can be dropped.

For constructing the decision trees at the client side, we use Gini index [14], as an entropy measure. We use the number of clients equal to 10. The depth of decision trees as 2, 3, 4, and 5. In FL, distributed clients have heterogeneous datasets. Hence, we create non-independent and identically distributed (i.i.d.) partitions for training decision tree model at each client locally. For creating non-i.i.d. partitions, we use an optimization problem discussed in a work by Torra [21], where the solution of the optimization problem returns the probability of records of a particular class label at each client. Since, we are dealing with binary classification in an FL scenario, it is important that each client has imbalanced number of records from the two classes to depict the non-i.i.d. scenario. In our experimental setting, firstly, we divide the data among each client, and then each client divides its data into training data, and testing data. Each client

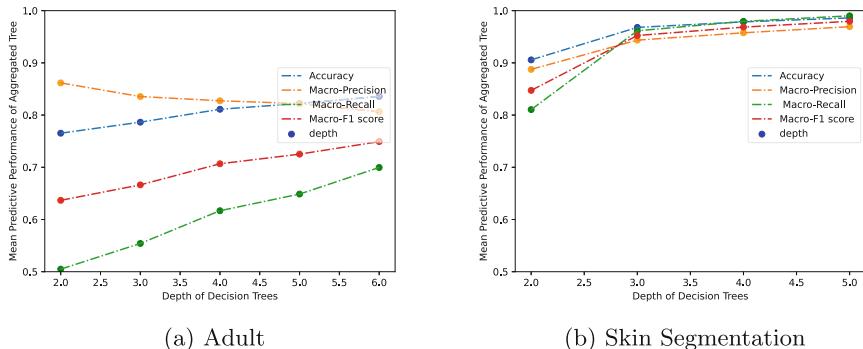
trains decision tree classifier on its own training data, and participate to the federation to obtain the aggregated/merged decision tree. After this, each client can use the aggregated decision tree to evaluate the performance on its own test data.

## 5 Results and Analysis

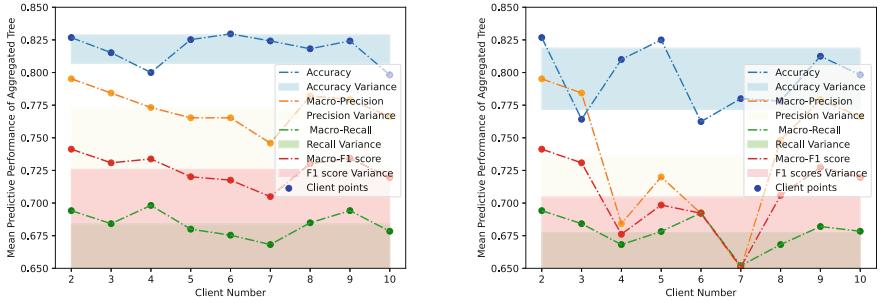
In this section, we present our results obtained on the datasets, Adult, and Skin segmentation. We use Accuracy, Precision, Recall, and F1-score for our performance evaluation.

Firstly, we show results without applying PRAM, and then we show the results on applying PRAM after the discretization step. We also show some interesting analysis regarding the level of perturbation denoted by  $\theta$  for PRAM. For the adult dataset, if we set  $\theta = 0.3$ , the accuracy drops to 0.7627, which is almost useless, as the model is predicting the same label for very test record, as the adult dataset is an imbalanced dataset, i.e. of the adult dataset 24,720 records out of the total 32,561 records (which is 75.9%) belong to the same class. Hence,  $\theta$  should be set lesser than 0.3. Also, the number of bins can vary the utility & privacy tradeoff. If the number of bins is set to a value too lower than the value calculated by the Freedman Diaconis rule, then the utility (accuracy) drops, although the privacy increases. Hence, to maintain good utility the parameter number of bins should be set around the value recommended by the Freedman Diaconis rule. We describe our result figures as follows.

- Figure 2 shows the performance of decision trees, when the depth of the decision trees is varied. It shows that the performance of the decision trees improves slightly, as we increase the depth of the decision tree for the Adult dataset. For the Skin Segmentation dataset also, the performance improves with the depth of the decision tree. From the results, we can conclude that the top features contribute most to the predictive capability of the aggregated decision tree model. Hence, this approach of performing FL with decision trees is scalable, if each client applies feature selection methods prior to the training process of decision trees.

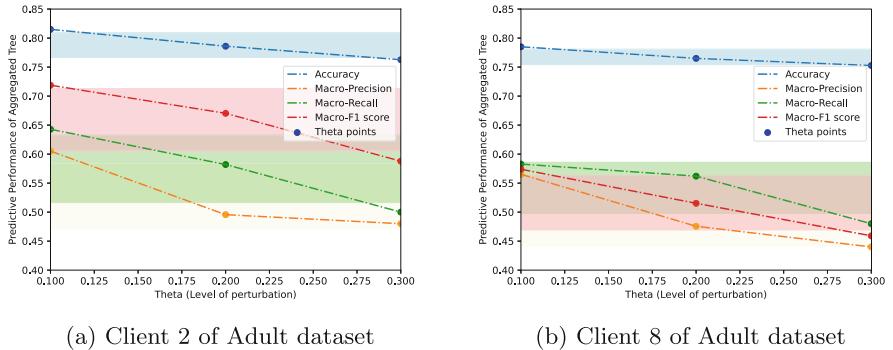


**Fig. 2.** Performance Metrics vs. the depth of the Decision Tree



(a) Adult dataset with i.i.d. setting

(b) Adult dataset with non-i.i.d. setting

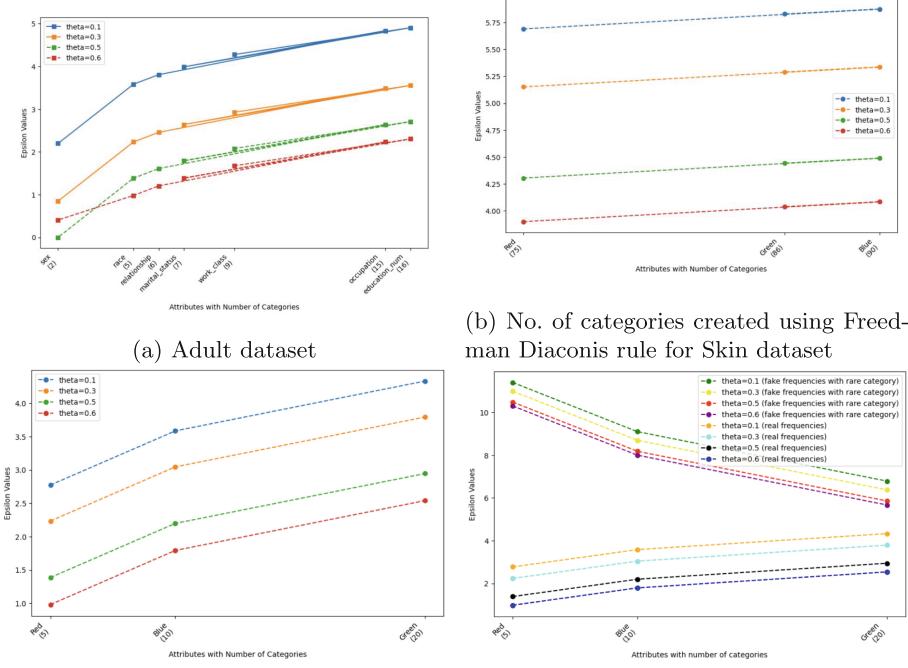
**Fig. 3.** Epsilon Vs. Number of Categories for different perturbation levels

(a) Client 2 of Adult dataset

(b) Client 8 of Adult dataset

**Fig. 4.** Performance Metrics vs. Client number without applying PRAM

- Figure 4 shows the performance on each client’s testing data, without applying PRAM for both i.i.d., and non-i.i.d. setting. The range of the shaded region around the line of each performance metrics is mean  $\pm$  performance metric. It shows that there is more variance in the non-i.i.d. setting, when each client has different proportions of samples from each class. The results of client 2 and client 8 are different from each other. This is because of imbalanced proportion of instances from both the classes in the non-i.i.d. partitions. Specifically, client 2 has more number of samples from the minority class label than client 8. We chose these two clients to show the impact of non-i.i.d. data distribution on the performance of federated framework.
- Figure 5 shows the performance on client’s testing data, after applying PRAM for non-i.i.d. setting. Here, the  $X$  shows the level of perturbation,  $\theta$ . We experimented with  $\theta = \{0.1, 0.2, 0.3\}$ . For better interpretability of our experimental results, we kept the value of  $\theta$  same for all the clients at once. Nevertheless, each client can choose the value of  $\theta$  according to its own personal privacy requirements. We show that as the value of  $\theta$  increases, the performance



**Fig. 5.** Performance Metrics vs. Theta after applying PRAM in a non-i.i.d. setting for Adult dataset

degrades. At  $\theta = 0.3$ , the accuracy drops to almost 0.75, which is of no good for the imbalanced dataset like Adult dataset. Hence, we stopped at  $\theta = 0.3$ .

- Figure 3 shows that the value of  $\epsilon_0$  versus the number of categories in each attribute. We observe that the values of  $\epsilon_0$  for a particular value of  $\theta$  are higher for the variables with more number of categories in comparison with the variables with less number of categories. We show it for both, the adult as well as the skin segmentation dataset. Since,  $\theta$  signifies the level of perturbation, a higher value of  $\theta$  results in a lower value of  $\epsilon$ , which indicates a higher level of privacy.
- Figure 5a shows the trend between the values of  $\epsilon_0$ , and the frequency of categories for the Adult dataset for different perturbation levels ( $\theta$ ). Education-num attribute has the highest number of categories (16), which is followed by the occupation attribute (15). Hence, these two variables have high values of  $\epsilon_0$  in contrast with the sex attribute, which has 2 categories, and the race attribute, which has 5 categories.
- For the Skin segmentation dataset in Fig. 5b, the Freedman Diaconis rule creates more number of categories for different values of  $\theta$ , which results in high value of  $\epsilon_0$  for a particular  $\theta$ , thus indicating low privacy.

- Freedman Diaconis rule usually creates a large number of categories, as shown in Fig. 5b. Therefore, we reduced the number of categories for the Skin segmentation dataset in Fig. 5c. We found out that reducing the number of categories results in lower values of  $\epsilon_0$ , with respect to the ones provided by the Freedman Diaconis rule. Hence, we obtain high privacy (low value of  $\epsilon_0$ ) when we reduce down the number of categories in a variable. We observed similar trends for both the datasets we experimented with.
- In Fig. 5d, we compared the values of  $\epsilon_0$  for real frequencies, and fake frequencies of each category in each attribute. The real frequencies of categories are distributed evenly, which means that no category is too sensitive or an outlier. In fake frequencies, we introduced a rare category with frequencies 5, 50, and 500 for each variable. This shoted up the values of  $\epsilon_0$ , which means lower privacy levels for the fake frequencies compared to the real frequencies, as shown in Fig. 5d. Hence, our experiments demonstrate that the rarer the category is among other categories for each variable, the higher the value of  $\epsilon_0$  for a particular  $\theta$ , indicating lower privacy. Conversely, higher the frequency of category, lower is the value of  $\epsilon_0$ , indicating higher privacy.

## 6 Conclusion and Future Work

We propose a complete federated pipeline, from discretization of attributes to training decision trees while preserving the privacy of clients. In our proposed framework, we offer privacy to each client’s data using PRAM, which provides DP guarantees. In our current implementation, all clients train decision trees locally, which is then aggregated by the central server. Our FL approach is one-shot, i.e., the training is done at the local devices only once, which makes it communication efficient. Hence, our proposed FL approach with decision trees is both privacy-preserving as well as communication efficient. A line of future work is exploring other multi-variate, and federated discretization approaches. Recently, Guidotti *et al.* [8] proposed a generative tree model, called GENTREE, to construct shallow decision trees. Another idea for future is to explore GENTREE from privacy, and federated perspective.

**Acknowledgement.** This study was partially funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Partial support from VR in the project “Privacy-aware secure explainable data-driven models in federated learning (VR 2023-05531)” is also acknowledged.

## References

1. Aggarwal, G., Mishra, N., Pinkas, B.: Secure computation of the median (and other elements of specified ranks). *J. Cryptol.* **23**(3), 373–401 (2010)
2. Chen, X., et al.: Fed-EINI: an efficient and interpretable inference framework for decision tree ensembles in vertical federated learning. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 1242–1248. IEEE (2021)

3. Cheng, K., et al.: Secureboost: a lossless federated learning framework. *IEEE Intell. Syst.* **36**(6), 87–98 (2021)
4. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. *Found. Trends® Theor. Comput. Sci.* **9**(3–4), 211–407 (2014)
5. Fan, C., Li, P.: Classification acceleration via merging decision trees. In: Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, pp. 13–22 (2020)
6. Gambs, S., Gmati, A., Hurfin, M.: Reconstruction attack through classifier analysis. In: Cappens-Boulahia, N., Cappens, F., Garcia-Alfaro, J. (eds.) DBSec 2012. LNCS, vol. 7371, pp. 274–281. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-31540-4\\_21](https://doi.org/10.1007/978-3-642-31540-4_21)
7. Garcia, S., Luengo, J., Sáez, J.A., Lopez, V., Herrera, F.: A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *IEEE Trans. Knowl. Data Eng.* **25**(4), 734–750 (2012)
8. Guidotti, R., Monreale, A., Setzu, M., Volpi, G.: Generative model for decision trees. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 21116–21124 (2024)
9. Kwatra, S., Torra, V.: A  $k$ -anonymised federated learning framework with decision trees. In: Garcia-Alfaro, J., Muñoz-Tapia, J.L., Navarro-Arribas, G., Soriano, M. (eds.) DPM/CBT -2021. LNCS, vol. 13140, pp. 106–120. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-93944-1\\_7](https://doi.org/10.1007/978-3-030-93944-1_7)
10. Kwatra, S., Torra, V.: A survey on tree aggregation. In: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6. IEEE (2021)
11. Li, Q., Wen, Z., He, B.: Practical federated gradient boosting decision trees. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 4642–4649 (2020)
12. Lindell, Y., Pinkas, B.: An efficient protocol for secure two-party computation in the presence of malicious adversaries. In: Naor, M. (ed.) EUROCRYPT 2007. LNCS, vol. 4515, pp. 52–78. Springer, Heidelberg (2007). [https://doi.org/10.1007/978-3-540-72540-4\\_4](https://doi.org/10.1007/978-3-540-72540-4_4)
13. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282. PMLR (2017)
14. Raileanu, L.E., Stoffel, K.: Theoretical comparison between the Gini index and information gain criteria. *Ann. Math. Artif. Intell.* **41**, 77–93 (2004)
15. Samarati, P.: Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001)
16. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information:  $k$ -anonymity and its enforcement through generalization and suppression (1998)
17. Senavirathne, N., Torra, V.: Rounding based continuous data discretization for statistical disclosure control. *J. Ambient. Intell. Humaniz. Comput.* **14**(11), 15139–15157 (2023)
18. Sweeney, L.: Achieving  $k$ -anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**(05), 571–588 (2002)
19. Sweeney, L.:  $k$ -anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowledge-Based Systems* **10**(05), 557–570 (2002)
20. Torra, V.: A Guide to Data Privacy. Springer, Cham (2022)

21. Torra, V.: A systematic construction of non-IID data sets from a single data set: non-identically distributed data. *Knowl. Inf. Syst.* **65**(3), 991–1003 (2023)
22. Wang, Z., Gai, K.: Decision tree-based federated learning: a survey. *Blockchains* **2**(1), 40–60 (2024)
23. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. In: *Advances in Neural Information Processing Systems*, vol. 32 (2019)



# Node Injection Link Stealing Attack

Oualid Zari<sup>1(✉)</sup>, Javier Parra-Arnau<sup>2,3</sup>, Ayşe Ünsal<sup>1</sup>, and Melek Önen<sup>1</sup>

<sup>1</sup> EURECOM, Biot, France

{oualid.zari,ayse.unsal,melek.onen}@eurecom.fr

<sup>2</sup> Karlsruhe Institute of Technology, Karlsruhe, Germany

<sup>3</sup> Universitat Politècnica de Catalunya, Barcelona, Spain

javier.parra@upc.edu

**Abstract.** We present a stealthy privacy attack that exposes links in Graph Neural Networks (GNNs). Focusing on dynamic GNNs, we propose to inject new nodes and attach them to a particular target node to infer its private edge information. Our approach significantly enhances the  $F_1$  score of the attack compared to the current state-of-the-art benchmarks. Specifically, for the Twitch dataset, our method improves the  $F_1$  score by 23.75%, and for the Flickr dataset, remarkably, it is more than three times better than the state-of-the-art. We also propose and evaluate defense strategies based on differentially private (DP) mechanisms relying on a newly defined DP notion. These solutions, on average, reduce the effectiveness of the attack by 71.9% while only incurring a minimal utility loss of about 3.2%.

**Keywords:** graph neural networks · link stealing · differential privacy

## 1 Introduction

Graph-structured data is prevalent in applications like social networks, biological systems, and recommendation engines. Graph Neural Networks (GNNs) are powerful for analyzing such data but they pose significant privacy risks as the graph structure is often sensitive. For instance, social network links can reveal common interests or personal beliefs, leading to privacy breaches [3].

This paper advances edge privacy in GNNs by introducing the Node Injection Link Stealing (NILS) attack and a tailored Differential Privacy (DP) defense. The adversary infers links among target nodes in a GNN trained for node classification. The GNN processes graph structure and node features to produce class membership predictions which are accessed via an inference API.

Previous works like the Linkteller attack [23] and studies on feature correlation [9] have shown that attackers can infer graph links by analyzing node features and GNN outputs. Our NILS attack introduces a stronger adversary exploiting GNN dynamics. The adversary adds a node, connects it to the target node, and queries the model with malicious features, inferring neighbors and stealing graph connections using various strategies.

The NILS attack is akin to activities in social networks, where sending a friend request is analogous to injecting a new node into the network. By connecting this new node to a target node, the attacker can observe changes in the network's behavior, similar to how content recommendations or interactions change when a new connection is established. This method allows the attacker to infer hidden links within the network, posing a significant threat to user privacy.

We also study potential defense strategies based on DP mechanisms. Specifically, we propose a new privacy notion, *one-node-one-edge privacy*, and evaluate existing DP-based defense strategies under this definition.

To summarize, we make the following contributions:

- We propose a novel attack (denoted NILS) for inferring private links in a graph structure by injecting a new node, linking it to a target node, and employing various strategies to analyze the changes in the GNN's output;
- We provide a comprehensive evaluation of the proposed attack's effectiveness on various datasets, demonstrating its superior performance compared to existing work such as LinkTeller [23] and link-stealing [9];
- We explore the application of DP mechanisms as a means to mitigate the effectiveness of our proposed attack, evaluating the trade-off between privacy preservation and model utility. To this end, we introduce a new notion of privacy and evaluate defense strategies under this new notion.

## 2 Background

We present a brief introduction to GNNs and formulate the concept of DP.

### 2.1 Graph Neural Networks

*GNNs Overview.* GNNs [18] are powerful machine learning models designed for graph-structured data. They effectively capture complex patterns in graphs, excelling in tasks like node classification [21], link prediction [26], and graph classification [24]. Node classification, the focus of this paper, involves assigning labels to nodes based on their features and graph structure.

A graph  $G = (V, E)$  consists of nodes  $V$  and edges  $E$ . Nodes represent data points like users in social networks or proteins in biological networks, while edges represent relationships or interactions. Graphs are represented using an adjacency matrix  $A \in \mathbb{R}^{n \times n}$ , where  $n$  is the number of nodes, and  $A_{ij} = 1$  if there is an edge between nodes  $i$  and  $j$ , and  $A_{ij} = 0$  otherwise. Nodes have feature vectors represented by the matrix  $X \in \mathbb{R}^{n \times d}$ , containing information like demographic data in social networks.

GNNs use a message-passing mechanism [18] for nodes to exchange and aggregate information from their neighbors, capturing local and global graph structure. For instance, Graph Convolutional Networks (GCNs) [11], a well-known GNN model, use graph convolutional layers formulated as:

$$H^{(0)} = X, \quad H^{(l+1)} = \sigma(\hat{A}H^{(l)}W^{(l)}), \quad H^{(L)} = P.$$

Here,  $H^{(0)}$  is the node feature matrix  $X$ ;  $H^{(l)}$  is the hidden node representation at layer  $l$ ;  $P$  represents prediction scores for each node class;  $W^{(l)}$  is the learnable weight matrix;  $\sigma(\cdot)$  is an activation function (e.g., ReLU); and  $\hat{A}$  is the normalized adjacency matrix.

*GNNs with Dynamic Graphs.* GNNs handle dynamic graphs as in social networks or recommendation systems where graphs evolve over time. New nodes and edges are introduced, requiring updates to the adjacency matrix  $A \in \mathbb{R}^{n \times n}$  and the feature matrix  $X \in \mathbb{R}^{n \times d}$ . The matrices expand to  $A' \in \mathbb{R}^{(n+1) \times (n+1)}$  and  $X' \in \mathbb{R}^{(n+1) \times d}$ , incorporating new nodes' connections and features. Once updated, the GNN performs inference on the modified graph using the message-passing mechanism described earlier.

## 2.2 Differential Privacy

Differential Privacy (DP) is a framework for ensuring the privacy of individual records in a database. In the context of GNNs, it helps protect the privacy of graph structures. For detailed definitions and mechanisms, please refer to the Appendix.

## 3 Related Work

GNNs have gained significant attention for their effectiveness in handling graph-based data across various applications [1, 12, 18]. As GNN adoption increases, concerns about privacy and adversarial attacks also rise [20]. Several privacy-preserving methods have been developed to mitigate these attacks [15, 17].

*Privacy Attacks on GNNs.* Privacy attacks on GNNs target graph nodes, attributes, or edges. Node privacy attacks, like membership inference attacks (MIA) [22], determine if a node was part of the training set. Attribute inference attacks [4] reveal sensitive node attributes. This work focuses on edge privacy violations, where attacks like link stealing, re-identification, or inference aim to uncover graph edges.

Early works [4, 9, 23] demonstrated the feasibility of link-stealing attacks. In [9], the adversary uses prior knowledge about the graph to infer links, applying clustering methods to predict connections among nodes. In [4], node embeddings trained to preserve the graph structure are used to recover edges by training a decoder. The Linkteller attack [23] involves probing node features and studying their GNN output predictions to infer links.

Existing link-stealing attacks have weaknesses. The attack in [9] requires a powerful adversary with access to features, a shadow dataset, and the ability to train shadow GNNs. Its performance declines in the inductive setting, where training and inference occur on different graphs. The Linkteller attack [23] has non-stealthy perturbations, especially with discrete data, making it easier to detect. Its effectiveness decreases against deeper GNNs.

This paper proposes a novel link-stealing attack, NILS, which addresses these limitations by exploiting the dynamic nature of GNNs through malicious node injection. NILS outperforms previous attacks [9, 23]. Concurrent with our research, [14] proposes a link inference attack using node injection. The attack injects multiple target nodes and nodes with zero features, training an attack model to infer links. While effective on high-homophily graphs [23], this method assumes the adversary has access to a partial graph and does not address low-homophily graphs.

*Differential Privacy Mechanisms for Graphs.* DP has been studied and applied to graphs to preserve sensitive information. Various DP mechanisms protect node and edge information [2]. Node-level DP [8] protects individual nodes from attacks like MIA [22]. Edge-level DP [13] protects edge information, preventing link stealing attacks [4, 9, 23].

Research has focused on achieving node-level and edge-level DP in graph models. Approaches allow the publication of graph statistics with edge-level DP guarantees, such as degree subgraph count and degree distributions [10]. While beneficial for graph analysis, these statistics are inadequate for GNN training, which requires access to the raw graph structure. Other approaches use input perturbation DP to release graphs while ensuring edge-level DP [23].

In designing DP solutions, specific privacy threats and adversary strengths must be considered. For the NILS attack, the adversary injects a node to a specific node to discover sensitive edge information, violating edge privacy. We propose a customized DP notion addressing this attack, leveraging the LapGraph algorithm [23] to achieve desired DP guarantees.

## 4 Node Injection Link Stealing Attack

GNNs are vulnerable to various privacy attacks aiming to learn about their underlying graph structure. They inherit attacks from standard neural networks, such as membership inference attacks (MIA) [22], where the adversary tries to determine if a sample is included in the training dataset. This paper focuses on the *link stealing attack*, where an adversary, without access to the adjacency matrix, aims to learn whether a particular edge exists. We introduce our node injection link stealing (NILS) attack that exploits the dynamic nature of GNNs.

### 4.1 Threat Model

*Environment.* We consider a GNN application where a server trains the GNN using a specific dataset and provides access through a black-box API. This API allows users to interact with the pre-trained GNN model without accessing its internal components. Users can submit prediction queries using node IDs and add new nodes using a *connect* query. The API processes input data and returns output predictions, ensuring the model’s computations remain hidden. Users only know the set of node IDs.

*Adversary’s Goal and Knowledge.* The adversary,  $\mathcal{A}$ , acts as a GNN user aiming to determine the neighbors of a specific *target node*,  $v_t$ , from a set of *target nodes*,  $V_{\mathcal{A}}$ . If  $\mathcal{A}$  aims to identify all links,  $V_{\mathcal{A}}$  would include all nodes in the graph  $V$ .  $\mathcal{A}$  may perform multiple node injections, but this approach’s practicality is debatable. In social networks, the adversary’s background knowledge, like users’ interests, guides the selection of target nodes more likely to be connected. We choose target nodes uniformly at random.  $\mathcal{A}$  obtains prediction scores of  $V_{\mathcal{A}}$  by querying their IDs through the API and can use the *connect* query to connect a node  $v_m$  to  $v_t$ .

## 4.2 Node Injection Link Stealing Attack

The NILS attack exploits the dynamic nature of GNNs. The adversary  $\mathcal{A}$  can *connect* new nodes and query prediction scores of nodes  $V_{\mathcal{A}}$  in the graph. By adding a new node  $v_m$ ,  $\mathcal{A}$  can discover neighbors of  $v_t$ . The attack is depicted in Fig. 1, outlined in Algorithm 1 and involves the following steps:

1.  $\mathcal{A}$  queries the prediction scores of target nodes  $V_{\mathcal{A}}$  and receives prediction matrix  $P$ .
2.  $\mathcal{A}$  generates malicious features for node  $v_m$  based on  $P$ .
3.  $\mathcal{A}$  sends a *connect* query to inject node  $v_m$ , specifying the features  $x_m$  and the ID of target node  $v_t$ .
4. The server adds node  $v_m$  to the graph and links it to  $v_t$ .
5.  $\mathcal{A}$  queries the server again for the new prediction matrix  $P'$  of  $V_{\mathcal{A}}$ .
6.  $\mathcal{A}$  computes the  $L_1$  distance between  $P(v)$  and  $P'(v)$  for each node  $v$  in  $V_{\mathcal{A}}$ . A significant change in prediction scores indicates a high probability of being a neighbor of  $v_t$ . If the difference exceeds a threshold  $R$ ,  $\mathcal{A}$  infers that node  $v$  is a neighbor of  $v_t$ .

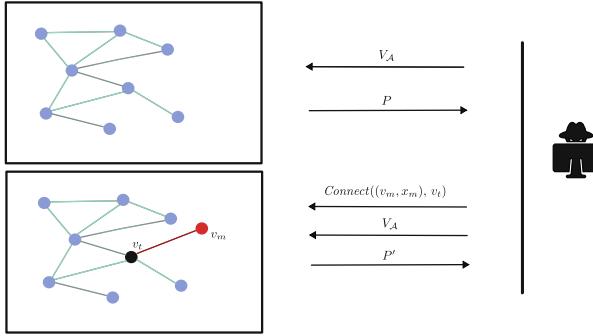
*Selection of the Decision Threshold R.* The threshold  $R$  is determined through parameter tuning, aiming for an optimal trade-off between precision and recall in identifying true neighbors of the target node, represented by the  $F_1$  score. Various values of  $R$  are evaluated, and the one yielding the highest  $F_1$  score is selected as optimal. In practice, the adversary can select  $R$  by estimating the graph’s density  $\hat{d}$  and picking the top  $\hat{d}$  nodes with the highest changes before and after injection.

## 4.3 Strategies for Generation of Malicious Node’s Features

To evaluate how injecting the malicious node  $v_m$  affects GNN predictions, we study five strategies to generate the malicious node’s features  $x_m$ . These strategies vary in sparsity and stealthiness, allowing us to assess their effectiveness in altering the model’s predictions. The proposed strategies are:

1. **All-ones strategy:** Generates a dense feature vector of all ones:

$$x_m = \mathbf{1}.$$



**Fig. 1.** Adversary-Server Interaction: At inference,  $\mathcal{A}$  first queries prediction scores  $P$  of target nodes  $V_{\mathcal{A}}$ . Next, the server sends predictions  $P$  to  $\mathcal{A}$ . Then,  $\mathcal{A}$  sends *Connect* query to inject malicious node  $v_m$ , with features  $x_m$ , to target node  $v_t$ . Finally,  $\mathcal{A}$  queries prediction scores  $P'$  of target nodes  $V_{\mathcal{A}}$ .

This can cause significant prediction changes but is less stealthy due to its density.

2. **All-zeros strategy:** Creates a sparse feature vector of all zeros:

$$x_m = \mathbf{0}.$$

This subtly alters GNN output, causing smaller prediction changes while being more stealthy.

3. **Identity strategy:** Uses a feature vector identical to the target node's:

$$x_m = x_t.$$

This confuses model predictions for neighboring nodes, with variable stealthiness depending on similarity to the target node's features.

4. **Max attributes strategy:** Computes the element-wise maximum of each attribute in the target nodes' feature matrix, excluding the target node's class:

$$x_{m,k} = \max_{i \in V_{\mathcal{A}}, \text{ with } C_i \neq C_t} X_{i,k}, \quad \text{for } k = 1, \dots, d.$$

Here,  $C_i$  and  $C_t$  represent the classes of node  $i$  and the target node, respectively. This strategy causes significant prediction changes but may be less stealthy due to exaggerated features.

5. **Class representative strategy:** Selects the feature vector of the node with the highest confidence score for a different class than the target node's:

$$x_m = x_{i^*} \text{ with } i^* = \arg \max_{\substack{i \in V_{\mathcal{A}}, \\ C_i \neq C_t}} p_{i,j}.$$

In this equation,  $i^*$  is the index of the node with the highest confidence score for a class different from the target node's. This strategy leverages model predictions to alter neighboring node predictions, potentially increasing stealthiness.

**Algorithm 1:** Node Injection Link Stealing Attack

---

**Input:** set of nodes  $V_A$  and target node  $v_t$ .  
**Output:** the identified neighbors of  $v_t$  by the adversary.

```

 $P = \text{GNN}(V_A, X_{V_A})$                                 ▷ Step 1
Generate malicious features  $x_m$  of node  $v_m$           ▷ Step 2
Connect node  $v_m$  to  $v_t$ .                                ▷ Step 3-4
 $P' = \text{GNN}(V_A \cup v_m, X_{V_A} \cup x_m)$            ▷ Step 5
for each node  $v$  in  $V_A$  do
     $D(v) = \|P(v) - P'(v)\|_1$                           ▷ Step 6
    if  $D(v) \geq R$  then
        |  $v$  is a neighbor of  $v_t$ 
    end
    else
        |  $v$  is not a neighbor of  $v_t$ 
    end
end

```

---

Additionally, we introduce the LinkTeller **Influence** strategy, which incorporates the feature perturbation strategy from [23]. This involves perturbing the target node's features by adding a small real value  $\alpha$ :

$$x_m = x_t + \alpha.$$

We compare the Influence strategy's performance with other strategies to determine if attack performance gains are due to node injection or malicious feature crafting. However, this strategy may be easily detected if  $x_t$  has discrete features, as  $x_m$  becomes real-valued.

## 5 Evaluation of the Attack

We present the evaluation results of our proposed attack, starting with a brief summary of the experimental setup and then analyzing its performance on various datasets. Detailed experimental setup information is provided in the appendix. We also address one limitation of the attack related to the depth of the GNN, discussed in the appendix.

### 5.1 Summary of Experimental Setup

We evaluated our attack on several real-world datasets, including the Flickr dataset [25], and two Twitch datasets (TWITCH-FR and TWITCH-RU) [16], following the approach in [23]. For the transductive setting, we used three citation network datasets: Cora, Citeseer, and Pubmed [19]. The models were trained using Graph Convolutional Networks (GCNs) with hyperparameters selected through a grid search strategy. Evaluation metrics include precision, recall, and the  $F_1$  score. Detailed information on datasets, models, and evaluation metrics is provided in the appendix.

## 5.2 Analysis of Strategies for Malicious Node's Features

We analyze the impact of different strategies for generating the features  $x_m$  of the malicious node  $v_m$  on the success of our attack.

The success rates, shown in Table 1, indicate that the All-ones, Max attributes, and Class representative strategies are the most effective in causing significant changes in the predictions of the target node's neighbors. Injecting nodes with high-valued or class-specific features effectively disrupts the model's output predictions.

Conversely, the All-zeros and Identity strategies exhibit lower success rates. While these strategies offer stealthiness, their impact on graph structure and predictions is less pronounced, highlighting a trade-off between attack effectiveness and stealthiness.

For the Influence strategy, NILS shows modest improvement over the LinkTeller baseline for the Twitch-FR dataset (Table 1), suggesting the effectiveness of node injection. However, for the Twitch-RU dataset, NILS underperforms compared to LinkTeller. The most significant improvement is seen in the Flickr dataset, where NILS increases the  $F_1$  score of LinkTeller from  $0.32 \pm 0.13$  to  $0.89 \pm 0.10$ , showcasing the advantage of node injection.

The Max attributes approach significantly enhances the  $F_1$  score beyond the LinkTeller baseline [23]. For the Twitch datasets, it improves the  $F_1$  score by 23.75% on average. For the Flickr dataset, it records a remarkable increase, raising the  $F_1$  score from 0.32 to 1.0, a 212.5% improvement over LinkTeller [23].

These findings underscore the importance of considering both the effectiveness and stealthiness of malicious feature generation strategies in link inference attacks on GNNs.

**Table 1.**  $F_1$  scores for different attack methods and datasets.

Method	Twitch-FR	Twitch-RU	Flickr
Class Rep.	$0.94 \pm 0.01$	$0.83 \pm 0.06$	$0.96 \pm 0.06$
Max Attr.	$0.99 \pm 0.00$	$0.98 \pm 0.02$	<b><math>1.00 \pm 0.00</math></b>
All-ones	<b><math>0.99 \pm 0.00</math></b>	<b><math>0.97 \pm 0.01</math></b>	$0.99 \pm 0.02$
All-zeros	$0.58 \pm 0.02$	$0.48 \pm 0.01$	$0.71 \pm 0.07$
Identity	$0.81 \pm 0.02$	$0.69 \pm 0.01$	$0.95 \pm 0.07$
Influence NILS	$0.81 \pm 0.02$	$0.70 \pm 0.01$	$0.89 \pm 0.10$
Influence LinkTeller [23]	$0.80 \pm 0.02$	$0.74 \pm 0.01$	$0.32 \pm 0.13$

## 5.3 Comparison with the Baselines

We evaluate the performance of the NILS attack compared to the LinkTeller attack using an identical experimental setup. Our focus is on analyzing the optimal attacks for both approaches, accurately estimating the number of neighbors

of the target nodes. The results in Table 2 show that NILS outperforms LinkTeller on both Twitch datasets (TWITCH-FR and TWITCH-RU). Additionally, NILS exhibits a substantial improvement over LinkTeller on the Flickr dataset, achieving nearly double the precision and recall values. NILS demonstrates stable performance across varying node degrees, with only a marginal decrease for high-degree target nodes. This slight decrease in performance for high-degree nodes is due to the reduced influence of each neighboring node. When the target node has a high degree, the impact of each individual neighbor is diluted in the aggregated information of the GCN layer. Overall, NILS consistently outperforms LinkTeller.

We also compare NILS with the link-stealing attacks introduced in [9], where different attack strategies rely on various types of background knowledge, such as node attributes and shadow datasets. Specifically, in Attack-2, the adversary has access to both the features and prediction scores of the nodes, creating LSA2-attr and LSA2-post attacks. LSA2-attr calculates distances between node attributes, while LSA2-post computes distances between prediction scores. These attacks align closely with our threat model, making them relevant for comparison. As shown in Table 3, NILS outperforms both LSA2-post and LSA2-attr attacks but performs nearly equivalent to LinkTeller. These results demonstrate that NILS maintains effectiveness in both transductive and inductive settings.

**Table 2.** Performance of NILS and LinkTeller across TWITCH-FR, TWITCH-RU, and Flickr under low, unconstrained, and high constraint settings.

Dataset	Method	low		unconstrained		high	
		precision	recall	precision	recall	precision	recall
TWITCH-FR	NILS (Ours)	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	99.1 $\pm$ 0.8	99.6 $\pm$ 0.35	99.9 $\pm$ 2.6	100.0 $\pm$ 0.0
	LinkTeller	92.5 $\pm$ 5.4	92.5 $\pm$ 5.4	84.1 $\pm$ 3.7	78.2 $\pm$ 1.9	83.2 $\pm$ 1.4	80.6 $\pm$ 6.7
TWITCH-RU	NILS (Ours)	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	96.4 $\pm$ 0.4	98.3 $\pm$ 0.7	99.9 $\pm$ 0.1	99.4 $\pm$ 0.1
	LinkTeller	78.8 $\pm$ 1.9	92.6 $\pm$ 5.5	71.8 $\pm$ 2.2	78.5 $\pm$ 2.4	89.7 $\pm$ 1.7	65.7 $\pm$ 3.9
Flickr	NILS (Ours)	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	99.1 $\pm$ 1.7	95.8 $\pm$ 5.0	93.7 $\pm$ 3.1	78.9 $\pm$ 1.9
	LinkTeller	51.0 $\pm$ 7.0	53.3 $\pm$ 4.7	33.8 $\pm$ 13.3	32.1 $\pm$ 13.3	18.2 $\pm$ 4.5	18.5 $\pm$ 6.1

**Table 3.** Performance of NILS, LinkTeller [23], and link-stealing attacks [9] across Cora, Citeseer, and Pubmed.

Method	Cora		Citeseer		Pubmed	
	precision	recall	precision	recall	precision	recall
NILS (Ours)	99.7 $\pm$ 0.2	99.6 $\pm$ 0.3	97.4 $\pm$ 0.2	98.2 $\pm$ 0.1	99.7 $\pm$ 0.0	100.0 $\pm$ 0.0
LinkTeller	99.5 $\pm$ 0.1	99.5 $\pm$ 0.1	99.7 $\pm$ 0.0	99.7 $\pm$ 0.0	99.7 $\pm$ 0.0	99.7 $\pm$ 0.0
LSA2-post	86.7 $\pm$ 0.2	86.7 $\pm$ 0.2	90.1 $\pm$ 0.2	90.1 $\pm$ 0.2	78.8 $\pm$ 0.1	78.8 $\pm$ 0.1
LSA2-attr	73.6 $\pm$ 0.1	73.6 $\pm$ 0.1	80.9 $\pm$ 0.1	80.9 $\pm$ 0.1	82.4 $\pm$ 0.1	82.4 $\pm$ 0.1

## 6 Defense

This section introduces DP in the context of GNNs to protect the privacy of the graph, preventing an adversary from discovering whether there is a link between two nodes. We define the neighboring relation of graphs and revise the definition of DP accordingly.

### 6.1 DP for Graphs

DP was originally defined for microdata, where two databases are neighbors if they differ by one record. For graphs, this notion must be adapted since two graphs may differ by either one edge or one node. In the literature, two adaptations of DP for graphs are proposed: edge-level DP [13] and node-level DP [8]. A graph  $\mathcal{G} = (V, E)$  is represented by adjacency matrix  $A$ , where  $A_{ij} = 1$  if there is a link between node  $i$  and node  $j$ , and  $A_{ij} = 0$  otherwise, where  $i, j \in \{1, \dots, |V|\}$ .

Edge-level DP considers graphs  $\mathcal{G}$  and  $\mathcal{G}'$  as neighbors if they differ by a single edge, while node-level DP considers them neighbors if they differ by a single node and all its incident edges.

### 6.2 One-Node-One-Edge-Level DP

The adversary defined in Sect. 4.2 adds a malicious node to a graph and connects it to a target node through a single edge. Countering such an adversary with node-level DP would increase noise and decrease model accuracy unnecessarily. Therefore, we define a new notion of neighborhood between graphs and the corresponding DP mechanism.

**Definition 1 (*One-node-one-edge-level adjacent graphs*).**  $\mathcal{G}$  and  $\mathcal{G}'$  are one-node-one-edge-level adjacent if one can be obtained from the other by adding a single node with one edge only.

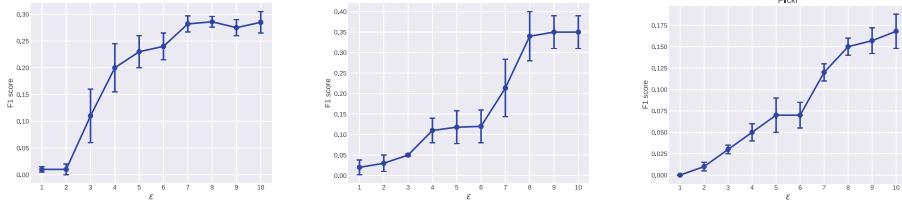
The adjacency matrices of such neighboring graphs differ by one row and one column, and the difference in  $L_1$ -norm is always one.

**Definition 2 (( $\varepsilon, \delta$ )-One-node-one-edge-level DP).** A randomized mechanism  $\mathcal{M}$  satisfies  $(\varepsilon, \delta)$ -one-node-one-edge-level DP with  $\varepsilon, \delta \geq 0$  if, for all pairs of one-node-one-edge-level adjacent graphs  $\mathcal{G}, \mathcal{G}'$  and for all measurable  $\mathcal{O} \subseteq \text{Range}(\mathcal{M})$ , the following holds:

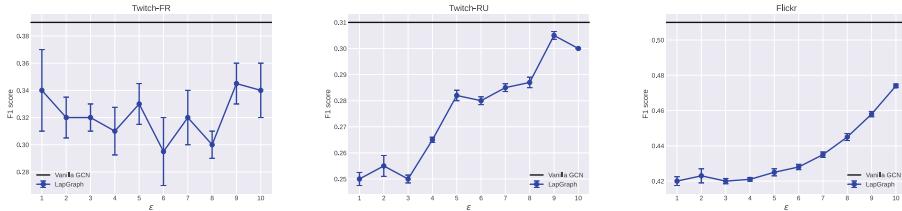
$$\Pr\{\mathcal{M}(\mathcal{G}) \in \mathcal{O}\} \leq e^\varepsilon \Pr\{\mathcal{M}(\mathcal{G}') \in \mathcal{O}\} + \delta.$$

### 6.3 Countermeasures for Our Attack

To defend against the NILS attack, we propose using the LapGraph mechanism introduced in [23]. While output perturbation [6] can also satisfy one-node-one-edge-level DP, it significantly deteriorates the GNN output accuracy due to the



**Fig. 2.**  $F_1$  score of the attack for different values of  $\epsilon$ .



**Fig. 3.**  $F_1$  score utility of the GCN for different values of  $\epsilon$ .

large  $L_1$ -global sensitivity of the prediction matrix. Instead, the LapGraph algorithm perturbs the adjacency matrix using the Laplace mechanism and binarizes it by replacing the top- $N$  largest values by 1 and the remaining values by 0. Here,  $N$  represents the estimated number of edges in the graph, computed using the Laplace mechanism.

Leveraging the post-processing property of DP<sup>1</sup>, the edge information remains protected even if the adversary observes the GNN’s predictions. Each time a user connects a new node, a new adjacency matrix is generated using LapGraph, accumulating the privacy budget by the sequential composition property of DP [7].

Although LapGraph was proposed to meet edge-level DP, it can also satisfy one-node-one-edge-level DP. Let  $f_A$  be the query function returning the adjacency matrix of a graph  $G$ . For one-node-one-edge neighboring graphs  $G$  and  $G'$ , the adjacency matrices  $A$  and  $A'$  have different dimensions. We append one zero-row and one-zero column to  $A$ , resulting in  $\bar{A}$ . The  $(n+1)$ -th columns (or rows) of  $\bar{A}$  and  $A'$  always differ in one element, yielding an  $L_1$ -global sensitivity of 1. Thus, the LapGraph mechanism provides stronger protection for the same level of utility compared to edge-level DP.

<sup>1</sup> The post-processing property allows arbitrary data-independent transformations to DP outputs without affecting their privacy guarantee [27].

## 6.4 LapGraph Evaluation

We evaluate the effectiveness of LapGraph [23] in reducing the success of the NILS attack while ensuring our one-node-one-edge-level DP notion. We also investigate the utility of GCN models trained with LapGraph protection.

*Evaluation Setup.* We use the same training hyperparameters and normalization techniques as in the vanilla case, where DP is not applied. Initially, we protect the training graph with LapGraph. Subsequently, we apply LapGraph each time the graph changes due to node injection by the adversary. Following [23], we compute the  $F_1$  score for our NILS attack and the classification task's  $F_1$  score for the GCN. This allows us to measure LapGraph protection and GCN utility across various privacy budgets  $\varepsilon$ . We report the results averaged over 5 runs with different random seeds for LapGraph.

*Evaluation Results.* Figure 2 presents the  $F_1$  score of the attack for various  $\varepsilon$  values. We observe that applying LapGraph reduces the effectiveness of NILS. The  $F_1$  score becomes almost zero when the privacy budget  $\varepsilon$  is small. However, for large  $\varepsilon$ , LapGraph provides moderate protection, but the attack's  $F_1$  score remains significantly lower than in the non-private case where DP is not applied.

In the LinkTeller [23] attack, where LapGraph is applied only once to ensure edge-level DP, LapGraph offers limited protection when  $\varepsilon$  is large, allowing LinkTeller to achieve a success rate nearly as high as in the non-private case. Conversely, in our scenario, where LapGraph is also applied after the adversary's node injection, LapGraph provides stronger protection.

Applying LapGraph during inference makes it more challenging for the adversary to distinguish between the target node's neighbors and non-neighbors, as the prediction scores of all target nodes change after each inference query. Consequently, the distances between the prediction scores  $P$  and  $P'$ , before and after the node injection, become noisier due to LapGraph's application.

To provide insights about the privacy-utility tradeoff of LapGraph, we present in Fig. 3 the utility of the GCNs for different values of the privacy budget. We observe that the utility increases when  $\varepsilon$  increases, as expected. Large values of  $\varepsilon \geq 7$  give a better utility close to that in the non-private vanilla case. Therefore, carefully choosing an  $\varepsilon$  will give fairly good utility and a certain level of protection against the NILS attack.

## 7 Conclusion

In this paper, we have presented a powerful new NILS attack-a link-stealing attack using node injection against GNNs. Our results have demonstrated the superior performance of NILS compared to previous attacks, further emphasizing the vulnerabilities of GNNs regarding edge information leakage. We have also evaluated NILS against differentially private GNNs, ensuring a one-node-one-edge-level DP notion specifically designed to protect against our proposed attack.

## A Appendix

### A.1 Differential Privacy

The original definition of Differential Privacy (DP) [5, 7] was introduced in the context of microdata, i.e., databases containing individual records. A central aspect of DP is the concept of *neighborhood*, originally defined for that data structure.

**Definition 3 (Neighboring Databases).** Let  $\mathcal{D}$  be the class of possible databases. Any two databases  $D, D' \in \mathcal{D}$  that differ in one record are called neighbors. For two neighboring databases,  $d(D, D') = 1$ , where  $d$  denotes the Hamming distance.

**Definition 4 ( $(\varepsilon, \delta)$ -Differential Privacy [5, 7]).** A randomized mechanism  $\mathcal{M}$  satisfies  $(\varepsilon, \delta)$ -DP with  $\varepsilon, \delta \geq 0$  if, for all pairs of neighboring databases  $D, D' \in \mathcal{D}$  and for all measurable  $\mathcal{O} \subseteq \text{Range}(\mathcal{M})$ ,

$$\mathbb{P}\{\mathcal{M}(D) \in \mathcal{O}\} \leq e^\varepsilon \mathbb{P}\{\mathcal{M}(D') \in \mathcal{O}\} + \delta.$$

In simple terms, the output of a DP mechanism should not reveal the presence or absence of any specific record in the database, up to an exponential factor  $\varepsilon$  and additional  $\delta$ . When each record corresponds to an individual, DP ensures their information remains confidential. A lower  $\varepsilon$ , known as the *privacy budget*, provides stronger protection.

The most popular DP mechanism is the Laplace mechanism, which relies on *global sensitivity*, defined as follows:

**Definition 5 (Global Sensitivity [7]).** The  $L_p$ -global sensitivity of a query function  $f : \mathcal{D} \rightarrow \mathbb{R}^d$  is defined as

$$\Delta_p(f) = \max_{D, D' \in \mathcal{D}} \|f(D) - f(D')\|_p,$$

where  $D, D'$  are any two neighboring databases.

**Definition 6 (Laplace Mechanism [7]).** Given any function  $f : \mathcal{D} \rightarrow \mathbb{R}^d$ , the Laplace mechanism is defined as:

$$\mathcal{M}_L(D, f(\cdot), \varepsilon) = f(D) + (Y_1, \dots, Y_d),$$

where  $Y_i$  are i.i.d. random variables drawn from a Laplace distribution with zero mean and scale  $\Delta_1(f)/\varepsilon$ .

## A.2 Experimental Setup

*Datasets.* We evaluated our attack on several real-world datasets used in related research. We include the Flickr dataset [25], where nodes represent images and edges connect nodes with shared properties. Node features contain word representations. We also use two Twitch datasets (TWITCH-FR and TWITCH-RU) [16] to evaluate NILS and Twitch-ES for training GNNs in an inductive setting, as done in [23]. Twitch datasets map follow connections between users and aim to classify if a streamer uses explicit language, using features like preferred games and location. For the transductive setting, where training and testing occur on the same graph, we use three citation network datasets: Cora, Citeseer, and Pubmed [19]. These involve predicting the topic of publications based on textual features and citation relationships.

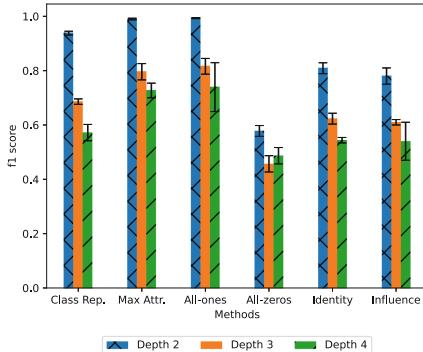
*Models.* We follow the approach in [23] for training models and selecting hyperparameters. The authors trained Graph Convolutional Networks (GCNs) using various configurations, including normalization techniques, the number of hidden layers, input and output units, and dropout rates. A grid search strategy identified optimal hyperparameters, evaluating performance on a validation set. By using the same training procedures and hyperparameter tuning strategies, we ensure consistency across studies.

*Evaluation Metrics.* We employ precision, recall, and the  $F_1$  score as our primary evaluation metrics, following [23]. These metrics are suitable for addressing the imbalanced binary classification problem, where the minority class (connected nodes) is of central interest. We select target nodes  $V_A$  such that  $|V_A| = 500$ , using uniform random sampling. We explore scenarios where target nodes exhibit either low or high degrees, as discussed in [23, Section V.D.]. Results are averaged over three runs with different random seeds, along with the corresponding standard deviation.

## A.3 Limitations: Depth of the GNN

We examine the impact of increasing the GNN depth on the success rate of the attack for the Twitch-FR dataset. Our findings in Fig. 4 show that as GNN depth increases, the attack’s success rate decreases. This reduction is due to the dilution of the injected node’s influence within the target node’s neighborhood. As GNN depth increases, the model aggregates information from a larger neighborhood, diluting the influence of the injected malicious node and diminishing the attack’s effectiveness.

Compared to LinkTeller [23], as shown in Table 4, NILS outperforms LinkTeller across various GCN depths. For the Twitch-FR dataset, NILS demonstrates higher precision and recall values at GCN depth 3 (precision:  $85.1 \pm 1.2$ , recall:  $81.6 \pm 1.2$ ) compared to LinkTeller at depth 2 (precision:  $84.1 \pm 3.7$ , recall:  $78.2 \pm 1.9$ ). These results highlight the effectiveness of our node injection strategy, consistently outperforming LinkTeller across different GCN depths.



**Fig. 4.** Success rates of the attack for different depths and malicious features generation strategies for Twitch-FR dataset

**Table 4.** Success rates of the attack for different depths in comparison with LinkTeller [23]. We use the all-ones strategy and Twitch-FR dataset.

Dataset	Method	Depth-2		Depth-3	
		precision	recall	precision	recall
TWITCH-FR	NILS (Ours)	$99.13 \pm 0.8$	$99.57 \pm 0.35$	$85.06 \pm 1.2$	$81.56 \pm 1.2$
	LinkTeller	$84.1 \pm 3.7$	$78.2 \pm 1.9$	$50.1 \pm 5.1$	$46.6 \pm 5.0$
TWITCH-RU	NILS (Ours)	$96.45 \pm 0.4$	$98.34 \pm 0.7$	$78.78 \pm 3.8$	$76.35 \pm 9.3$
	LinkTeller	$71.8 \pm 2.2$	$78.5 \pm 2.4$	$45.7 \pm 2.2$	$50.0 \pm 2.8$

## References

- Atwood, J., Towsley, D.: Diffusion-convolutional neural networks. In: 30th International Conference on Neural Information Processing Systems (NIPS) (2016)
- Brunet, S., Canard, S., Gambs, S., Olivier, B.: Novel differentially private mechanisms for graphs. IACR Cryptology ePrint Archive, p. 745 (2016)
- Cadwalladr, C., Graham-Harrison, E.: Revealed: 50 million Facebook profiles harvested for Cambridge analytica in major data breach. The Guardian (2018)
- Duddu, V., Boutet, A., Shejwalkar, V.: Quantifying privacy leakage in graph embedding. In: 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous) (2021)
- Dwork, C.: Differential privacy. In: 33rd International Colloquium on Automata, Languages and Programming, ICALP (2006)
- Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography (2006)
- Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
- Hay, M., Li, C., Miklau, G., Jensen, D.: Accurate estimation of the degree distribution of private networks. In: IEEE International Conference on Data Mining (2009)
- He, X., Jia, J., Backes, M., Gong, N.Z., Zhang, Y.: Stealing links from graph neural networks. In: USENIX Security Symposium (2021)

10. Karwa, V., Slavkovic, A.B.: Differentially private graphical degree sequences and synthetic graphs. In: Privacy in Statistical Databases - UNESCO Chair in Data Privacy, International Conference, PSD (2012)
11. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations (ICLR) (2017)
12. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations (ICLR) (2017)
13. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. *Science* **311**(5757), 88–90 (2006)
14. Li, K., et al.: Towards practical edge inference attacks against graph neural networks. In: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2023 (2023)
15. Mueller, T.T., Usynin, D., Paetzold, J.C., Rueckert, D., Kaassis, G.: SoK: differential privacy on graph-structured data. *CoRR* abs/2203.09205 (2022)
16. Rozemberczki, B., Sarkar, R.: Twitch gamers: a dataset for evaluating proximity preserving and structural role-based node embeddings. *arXiv preprint arXiv:2101.03091* (2021)
17. Sajadmanesh, S., Shamsabadi, A.S., Bellet, A., Gatica-Perez, D.: GAP: differentially private graph neural networks with aggregation perturbation. *CoRR* abs/2203.00949 (2022)
18. Scarselli, F., Gori, M., Tsoli, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. *IEEE Trans. Neural Netw.* **20**(1), 61–80 (2009)
19. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., Eliassi-Rad, T.: Collective classification in network data. *AI Mag.* **29**(3), 93 (2008)
20. Sun, L., et al.: Adversarial attack and defense on graph data: a survey. *arXiv preprint arXiv:1812.10528* (2018)
21. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: 6th International Conference on Learning Representations (ICLR) (2018)
22. Wu, B., Yang, X., Pan, S., Yuan, X.: Adapting membership inference attacks to GNN for graph classification: approaches and implications. In: IEEE International Conference on Data Mining (ICDM) (2021)
23. Wu, F., Long, Y., Zhang, C., Li, B.: Linkteller: recovering private edges from graph neural networks via influence analysis. In: 2022 IEEE Symposium on Security and Privacy (SP) (2022)
24. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: 7th International Conference on Learning Representations (ICLR) (2019)
25. Zeng, H., Zhou, H., Srivastava, A., Kannan, R., Prasanna, V.K.: Graphsaint: graph sampling based inductive learning method. In: 8th International Conference on Learning Representations (ICLR) (2020)
26. Zhang, M., Chen, Y.: Link prediction based on graph neural networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
27. Zhu, K., Fioretto, F., Van Hentenryck, P.: Post-processing of differentially private data: a fairness perspective. In: 31st International Joint Conference on Artificial Intelligence, IJCAI (2022)



# Assessing the Potentials of LLMs and GANs as State-of-the-Art Tabular Synthetic Data Generation Methods

Marko Miletic and Murat Sariyar<sup>(✉)</sup>

Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland  
murat.sariyar@bfh.ch

**Abstract.** The abundance of tabular microdata constitutes a valuable resource for research, policymaking, and innovation. However, due to stringent privacy regulations, a significant portion of this data remains inaccessible. To address this, synthetic data generation methods have emerged as a promising solution. Here, we assess the potentials of two state-of-the-art GAN and LLM tabular synthetic data generators using different utility & risk measures and propose a robust risk estimation for individual records based on shared nearest neighbors. LLMs outperform CTGAN by generating synthetic data that more closely matches real data distributions, as evidenced by lower Wasserstein distances. LLMs also generally provide better predictive performance compared to CTGAN, with higher  $F_1$  and  $R^2$  scores. Interestingly, this does not necessarily mean that LLMs better capture correlations. Our proposed risk measure, Shared Neighbor Identifiability (SNI), proves effective in accurately assessing identification risk, offering a robust tool for navigating the risk-utility trade-off. Furthermore, we identify the challenges posed by mixed feature types in distance calculation. Ultimately, the choice between LLMs and GANs depends on factors such as data complexity, computational resources, and the desired level of model interpretability, emphasizing the importance of informed decision-making in selecting the appropriate generative model for specific applications.

**Keywords:** Tabular Data · Synthetic Data Generation · LLM · GAN

## 1 Introduction

Tabular microdata amassed by diverse entities, encompassing statistical agencies, research institutions, and healthcare organizations represents a rich repository of information central for research, policymaking, and innovation [1]. However, a considerable portion of this data remains unpublished and inaccessible to the broader community due to stringent privacy and confidentiality concerns. These concerns stem from a complex interplay of legal mandates and ethical obligations aimed at safeguarding individual privacy. For instance, data protection regulations like the General Data Protection Regulation (GDPR) in the European Union impose rigorous requirements on how personal data can be processed and shared [2]. In this context, the development of sophisticated

methods for generating synthetic data has emerged as a promising solution to balance the need for data utility with the imperative of preserving privacy [3]. Synthetic data, which are artificially generated to mimic the statistical properties of real data, are claimed to provide researchers and policymakers with valuable insights without exposing actual sensitive information [4]. Ensuring that synthetic data maintains the utility of the original data while mitigating the risk of re-identification and preserving the underlying statistical relationships is a complex and ongoing area of research [5].

In recent years, significant advancements with various approaches have been made in the realm of Synthetic Data Generation (SDG) methods [6]. Within the statistical research community, research and development have focused on Copulas [7], Bayesian Networks [8], and multiple imputation methods such as Latent Class Analysis [9] or Stochastic Regression Imputation [10]. In the context of machine learning, techniques like SMOTE [11] and ROSE [12] are proposed for balancing data. More advanced approaches include artificial neural network architectures such as Variational Autoencoders [13], Generative Adversarial Networks (GANs [14]), and Large Language Models (LLMs [15]). These approaches exemplify various modeling strategies, including discriminative and generative modeling, as well as a combination of both, as seen most prominently in GANs. Discriminative models seek to learn conditional distributions, whereas generative models tackle the broader task of understanding the joint distribution across all variables. Depending on the selected modeling approach, challenges may arise especially due to computational complexity and data fidelity concerns.

It is sometimes claimed that synthetic data by design is not susceptible to attacks such as linkage, as there is no direct link between the original subject in the dataset and its synthetically generated counterpart [16]. Following this narrative, many studies focus on distance metrics between real and synthetic records, which represents only a proxy for the inherent disclosure risk in the synthetic records. On the other hand, this claim has been relativized in recent studies, which emphasize that privacy risks in synthetic data publishing are severely underestimated and that state-of-the-art generative models either do not prevent inference attacks or do not retain data utility [17]. In other words, they do not provide a better trade-off between risk and utility than the traditional anonymization techniques. In addition to that, they frequently exhibit a higher unpredictability as the complexity of the methods employed do not allow the anticipation of which traits a synthetic dataset will retain, and which will be lost.

There are several characteristics of tabular microdata, which must be accounted for when designing an SDG method. First, tabular data can contain various feature types, e.g., categorical (including binary), numerical (continuous as well as discrete), date & time, geospatial, and textual. This is challenging, as, for instance, copulas can only model non-linearly correlated continuous variables. Second, preprocessing such as normalization and scaling for numerical values, label or one-hot encoding for categorical data, missing value imputation, and outlier removal can incur information loss, making it difficult to disentangle SDG and pre-processing effects on data utility. Third, in contrast to text-based and image-based data, tabular data are rather order-invariant, which renders certain position-reliant methods such as sequence models and convolutional neural networks less effective. Additionally, tabular data often exhibit complex dependencies, correlations,

and hierarchical structures that must be accurately captured to ensure the utility and fidelity of the synthetic data.

Here, we assess the potentials of two state-of-the-art GAN and LLM tabular synthetic data generators using different utility and risk measures and propose a robust risk estimation for individual records based on shared nearest neighbors. Both method classes offer greater flexibility in modeling the data distribution than their statistical counterparts. GANs usually employ neural networks with a few thousand parameters only, leading to fast computation for directly modeling the joint data distribution. Issues like mode collapse and instable training are addressed by several extensions [18]. LLMs on the other hand, have millions/billions of parameters in order to model a probabilistic language model for the prediction of a feature or masked word. Here, not the syntactical level is approached but a semantic joint distribution. In philosophical terms, ontology is approximated by semantics. Central issues here are therefore related to the representation of the data, e.g., how to code tabular data as text and what kind of padding to use. Due to the number of parameters to learn, computational issues arise as well, for which several strategies such as token sequence compressions are used [19].

In the next section, we will first describe the nine datasets for benchmarking the methods for synthetic data generation, after which we introduce CTGAN as our GAN architecture and Tabula as our LLM framework, which can be used with pre-trained as well as randomly initialized models. Even though the LLM we are considering here is not ‘large’ in the sense of having more than 1 billion parameters, we will use this designation for simplicity. The end of the Methods section describes our metrics for general and specific utility as well as for the risk measurement. In the Result section, benchmarks on the datasets for different epochs of the SDG methods are provided and implications are formulated. In the Discussion section, insights are summarized and routes for further research are outlined.

## 2 Methods

In the following sections, we provide an overview of the characteristics of the nine datasets used for this study, after which we detail the synthetic data generation methods applied to these datasets. Following this, we outline the utility metrics and the risk measures employed. Finally, the evaluation process is described.

### 2.1 Datasets

Nine tabular datasets from various domains with different characteristics, variable types, and task types were sourced from the UCI Machine Learning Repository and Kaggle, as displayed in Table 1. The enumerated values for the numerical and categorical columns, obtained after preprocessing, include the target variable. Categorical columns encompass discrete numerical variables due to the working mechanism of the algorithms applied later. The URLs for these datasets are listed in Table A1 of the appendix. Six of the datasets are commonly used for classification tasks, encompassing both binary and multiclass target variables. The other three datasets are used for regression tasks and therefore

contain continuous target variables. The diversity in these datasets enables a comprehensive evaluation for both classification and regression models. SDG methods that perform well on varied datasets are more likely to generalize effectively to new, unseen data, which is crucial for robustness. The sample sizes of the datasets range from fewer than 1,000 to slightly over 30,000. Each dataset was split into 80% training and 20% testing sets. All models were trained using the same training samples. To showcase an out-of-the-box approach with real-world data, only following preprocessing steps were applied: a) Removal of missing data to ensure consistency in sample sizes for subsequent utility calculations and to maintain the integrity and validity of the datasets concerning the requirements of the different SDG methods, and b) Removal of redundant and noise variables, which are either correlated with other variables or do not correlate at all with the target variable.

**Table 1.** Detailed overview of the characteristics of the datasets used for the study.

Abbreviation	Name	# Train-split (80%)	# Test-split (20%)	# Numerical (cont.)	# Categorical	Task Type
AB	Abalone	3,341	836	8	1	Regression
AD	Adult	24,129	6,033	6	9	Binclass
BD	Buddy	13,885	3,472	4	4	Multiclass
BP	Body Performance	10,714	2,679	10	2	Multiclass
CH	Churn	8,000	2,000	6	5	Binclass
CP	Cell Phone	1,600	400	14	7	Multiclass
DB	Diabetes	614	154	8	1	Binclass
IN	Insurance	1,070	268	3	4	Regression
KI	King	17,290	4,323	11	8	Regression

## 2.2 Synthetic Data Generation Methods

**Conditional Tabular GAN (CTGAN).** Tabular data presents unique challenges for GAN design. Especially, the mixture of continuous and non-continuous variables poses four difficulties in terms of estimating the joint distribution: First, GANs are sensitive to the scaling and normalization of input features, which is crucial for stabilizing the training process. Second, different columns may require different scaling techniques due to varying ranges and distributions. Improper scaling can lead to poor convergence, and gradient computation issues. Third, GANs need to effectively handle missing data during training, ensuring that the generated synthetic data maintains the same pattern of missingness as the original data. Fourth, highly imbalanced categorical columns further complicate the training process, as dominant categories can cause mode collapse, leading to insufficient training for minor classes.

To mitigate these challenges, a novel approach called Conditional Tabular GAN (CTGAN) was introduced [18]. CTGAN offers two significant contributions: mode-specific normalization, which addresses non-Gaussian and multimodal distributions, and training-by-sampling, which effectively manages imbalanced categorical columns by sampling in such a manner that infrequent values have a higher probability to be represented during the training process. In an empirical evaluation, CTGAN outperformed Bayesian networks, Tabular Variational Autoencoders, and other GAN variants. CTGAN was also selected because alternatives such as PATE-GAN and ADS-GAN [20], while privacy-aware, either fail to fully adhere to privacy constraints or exhibit unstable training procedures, despite claims of mitigating this issue through techniques such as Wasserstein with Gradient Penalty [21].

**Tabula as an LLM Framework for Tabular Data.** Transformer-based LLMs have emerged as powerful tools for auto-regressive next-token prediction, a capability recently leveraged for generating highly realistic tabular data. Notably, these models facilitate conditioning on any subset of features with minimal computational overhead. The GReaT Framework pioneered a transformer-decoder-based network architecture tailored explicitly for modeling and generating tabular data [22]. At the core of this methodology lies the fusion of tabular and textual data modalities through a sophisticated textual encoding scheme, enabling seamless integration and synthesis of information from diverse sources. Tabula builds on the GReaT Framework and aims to reduce training time as well as to increase data quality through token sequence compression strategies and a novel token padding strategy known as middle padding. Tabula allows to use pre-trained as well as randomly initialized language models for creating a foundational tabular data generator. Despite the potential advantages of bigger LLMs in various contexts, there are several reasons to avoid using them for tabular data generation, e.g., they are resource intensive, they require vast amounts of training data, and they are harder to manage locally.

Here, we utilize Tabula with a randomly initialized 82M parameter DistilGPT2 model and the left padding strategy. It was shown that pre-trained models on a small scale (parameters  $\ll$  1 billion) have to unlearn too much as to provide a high-quality tabular data generator [19]. Internal investigations suggest that middle padding, while reducing training time, is detrimental to the data utility, as it tries to solve the issue of tokenizing numbers with many digits in a column, which can be mitigated by pre-processing and can be better addressed by a corresponding tokenizer. Following Tabula, we do not employ feature order permutation during the training process, because the target variable is always used to condition the sampling process instead of an arbitrary subset of the features.

### 2.3 Utility Metrics and Risk Measures

The utility of synthetic data can be classified into general and specific utility. General utility refers to statistical similarity, such as statistical distance between two probability distributions or correlation matrices that are determinable prior to specific analyses and are therefore use case agnostic. On the other hand, specific utility in this context assesses

the performance of machine learning or statistical models in downstream tasks. In this study, classification as well as regression tasks are considered.

**General Utility.** Numerical variables naturally form metric spaces while categorical data only have a metric space if the data is represented as vectors, e.g., as word embeddings, or if they are binary. Of interest for us are probabilistic metric spaces where the distance is related to probability distributions. Probability metrics satisfy the conditions of non-negativity, symmetry and the triangle inequality. For instance, the Wasserstein distance is suitable for measuring the probability distance between continuous numerical and binary variables, while the Hellinger distance is appropriate for distributions defined over discrete bins [23]. The Wasserstein distance, also known as earth movers' distance, computes the optimal transport plan, i.e., the plan that minimizes the total transportation cost between two distributions [24]. We use the implementation in the Python Optimal Transport Library (POT) for the calculation of the Wasserstein distance [25].

We compute the Wasserstein distance in two ways: first, considering all variables where categorical ones are one-hot encoded to establish a metric space; and second, solely for variables with an inherent metric space, excluding categorical ones. Additionally, we compute the Hellinger distance for all categorical columns. Our objective is to sum the Wasserstein distance (for numerical variables) and the Hellinger distance (for categorical variables) and compare this sum to the Wasserstein distance calculated using one-hot encoded variables. We compute the Wasserstein distance based on min-max normalized data.

To comprehensively evaluate the model's efficacy in capturing individual and joint feature distributions across nine datasets, we conduct a qualitative comparison. This involves computing and visualizing the absolute differences between correlation matrices derived from real and synthetic tabular data. To ensure consistency across variable types, we employ specific metrics: For numerical-numerical correlations, we use the Pearson correlation coefficient. In cases involving both categorical and numerical variables, we utilize the correlation ratio. For categorical-categorical correlations, the Uncertainty coefficient is employed. Additionally, all categorical variables have been one-hot encoded to facilitate accurate analysis.

**Specific Utility.** To evaluate the efficacy of machine learning models, we use random forests for both classification and regression tasks. We utilize the 20% of the original dataset as test data for evaluation purposes. For classification tasks, we employ the  $F_1$  score, while for regression tasks, we utilize the  $R^2$  score. Hyperparameter tuning for all random forest models is performed through cross-validated grid-search on the training split of the original dataset (see Table A4 in the appendix for details). Each optimized random forest model is then trained on the synthetic dataset, and its performance is evaluated on the reserved test set from the original data.

**Risk Measures.** To approximate the inherent re-identification risk in synthetic data, we first utilize the concept of  $\epsilon$ -identifiability [26]. This metric quantifies the likelihood that the  $i$ -th record in the original data has a smaller weighted distance to its nearest synthetic observation ( $\hat{r}_i$ ) than to its nearest real observation ( $r_i$ ), with  $\epsilon$  representing the threshold of closeness. The use of Euclidean distance alone is insufficient for identifiability purposes because the frequency of certain features or values varies. Therefore, discrete

entropy has been proposed to define the weight vector, representing the uncertainty of each feature. The underlying principle is that lower uncertainty results in higher identifiability. However, any other weight definition can also be applied within this framework. In summary, a synthetic dataset  $\hat{D}$  is  $\varepsilon$ -identifiable from the original dataset  $D$  if the following condition holds ( $\delta$  is the indicator function):

$$I(D, \hat{D}) = \frac{1}{N} \sum_{i \in N} \delta(\hat{r}_i < r_i) < \varepsilon$$

For all occurrences where the measurement  $\hat{r}_i$  is smaller than  $r_i$  we calculate the proportion of these occurrences relative to the total number of events. Additionally, we propose a more robust risk metric called Shared Neighbor Identifiability (SNI), which considers the local density around each synthetic data point using the Shared Nearest Neighbors (SNN [27]). This proposal addresses the issue that  $\varepsilon$ -identifiability is problematic only when the original record and the nearest synthetic records are located in a sparse region of the feature space. In DBSCAN [28], the density of a point is determined by counting the number of points within a specified radius, Eps. Points with a density above a threshold, MinPts, are classified as core points, while noise points are defined as non-core points lacking a core point within the Eps radius. SNN extends this concept by using the number of shared neighbors as a similarity measure, thereby avoiding issues that arise when relying solely on Euclidean distance. The SNN clustering algorithm consists of the following steps: (1) Construct a k-nearest neighbor graph, connecting each data point to its k nearest neighbors. (2) Compute the number of shared neighbors for each pair of points. Points with SNN similarity below a certain threshold are classified as noise. (3) Calculate the SNN density for each point by counting the number of points with an SNN similarity equal to or greater than the user-specified parameter, Eps. (4) Identify core points with an SNN density above another user-specified parameter, MinPts. (5) Form clusters by grouping core points within a radius, Eps, of each other. (6) Discard noise points that are not within a radius of Eps of any core point. (7) Assign remaining non-noise, non-core points to the nearest core point.

SNI leverages SNN for assessing re-identification risks in synthetic data. We configure SNN with an Eps value of 3 and MinPts of 2, ensuring that noise points are primarily isolated data points with no “near” neighbors. Synthetic and original data are combined, and SNN clustering is applied. In the subsequent post-processing step, any cluster containing fewer than five data points, comprising both original and synthetic records, is flagged as at risk. To maintain simplicity and draw an analogy with k-anonymity, we calculate the percentage of synthetic data deemed at risk, i.e., those belonging to a risky cluster, relative to the total number of synthetic records. Such a measure is also motivated by our concept of relational identity [29].

## 2.4 Evaluation Design

For each model and epoch, i.e., one complete pass of the training data, we execute the SDG process to produce 100 synthetic datasets per model. This procedure enables a more precise assessment of the variability and expected values of the utility metrics and risk

measures by calculating their mean and standard deviation. This approach addresses the inherent unpredictability in the preservation of signals and the potential loss of information within synthetic datasets. We perform this comprehensive analysis using multiple metrics, including the Wasserstein distance, Hellinger distance, summarized statistical distance, correlation distance,  $\epsilon$ -identifiability, and shared neighbor identifiability.

### 3 Results

In Sect. 3.1, we discuss the results regarding general utility, focusing on the Wasserstein distance and the mean absolute distance of the correlations in the original and synthetic datasets. Section 3.2 covers the results on specific utility using random forest estimators. Section 3.3 addresses the  $\epsilon$ -identifiability, and in Sect. 3.4, we present the results of our proposed SNI measure.

#### 3.1 General Utility

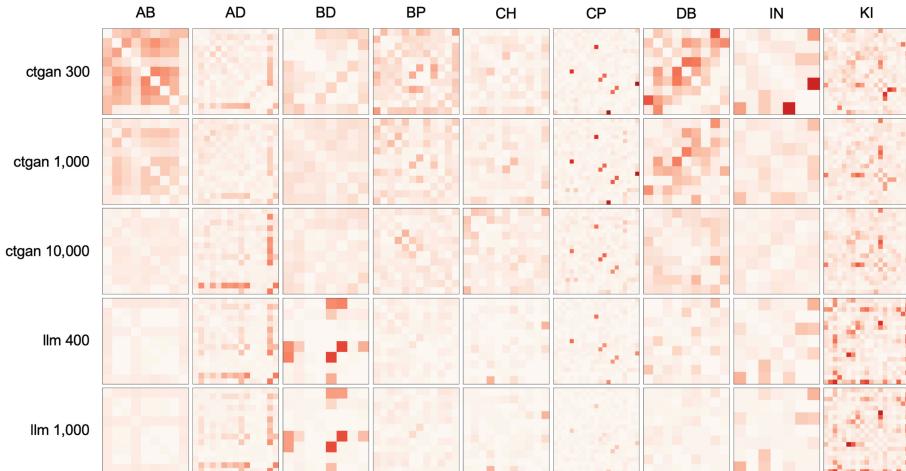
Table 2 shows the results of the Wasserstein distance metric. LLM consistently outperforms the CTGAN models across most datasets, demonstrating lower Wasserstein distances and indicating a closer match to the real data distributions. LLM achieves low Wasserstein distances with fewer epochs and maintains stable performance, showcasing its efficiency and robustness in generative modeling. As the number of epochs increases, the Wasserstein distance generally decreases for CTGAN, indicating improved model performance with more training. However, there is some inconsistency, such as on CP, where the distance slightly increases from 3.362 (300 epochs) to 3.84 (10,000 epochs). CTGAN exhibits more variation in standard deviations, especially at higher epochs, indicating less stability than LLM. Comparison with Table A2 in the appendix shows that similar conclusions can be drawn. Differences arise because, despite scaling, the metrics used are different. Aggregating these metrics appears less appropriate than using a single unified metric.

**Table 2.** Comparison of general utility in terms of Wasserstein distance (all categorical variables are one-hot encoded) on multiple datasets and epochs.

model	epochs	AB	AD	BD	BP	CH	CP	DB	IN	KI
ctgan	300	.141 ± .013	2.345 ± .012	1.128 ± .010	.364 ± .008	1.350 ± .020	3.362 ± .029	.451 ± .039	.936 ± .037	3.67 ± .012
ctgan	1,000	.088 ± .014	2.161 ± .011	1.009 ± .011	.230 ± .008	1.070 ± .018	3.810 ± .035	.416 ± .040	.728 ± .037	3.489 ± .011
ctgan	10,000	.067 ± .015	2.094 ± .015	.848 ± .011	.281 ± .009	.758 ± .016	3.384 ± .042	.033 ± .040	.744 ± .039	3.519 ± .015
llm	400	.037 ± .014	1.394 ± .018	.308 ± .014	.093 ± .012	.307 ± .025	2.296 ± .070	.104 ± .026	.408 ± .048	4.105 ± .019
llm	1,000	.029 ± .013	1.309 ± .016	.265 ± .013	.110 ± .011	.317 ± .024	1.960 ± .066	.090 ± .025	.283 ± .034	3.932 ± .019

Figure 1 shows the results for the mean absolute L2 correlation matrix distance between synthetic and original datasets (see also Table A3 in the appendix). For the CTGAN model, there's a consistent decrease in mean absolute differences as training epochs increase from 300 to 10,000. This indicates improved capability in capturing data associations, especially evident in all datasets except IN, where there are some

instabilities observed in the dataset. Conversely, the LLM model shows mixed results. While some datasets exhibit slight improvements with more epochs, others show no significant changes or even worsened performance. Overall, the LLM model consistently presents higher mean absolute differences compared to CTGAN, suggesting it may be less effective in capturing data associations.



**Fig. 1.** Visualization depicting the mean absolute distances of the correlations between original and synthetic datasets. Shades of red indicate larger distances, while shades of white represent smaller distances. (Color figure online)

### 3.2 Specific Utility

In Table 3, the machine learning efficiency is depicted, with the original model consistently demonstrating high scores across all datasets. LLM models generally outperform CTGAN models, notably evident in the CH dataset, where the LLM at 1,000 epochs achieves a score of  $.848 \pm .004$ , surpassing CTGAN's  $.800 \pm .005$  at 10,000 epochs. The impact of epochs on performance is evident, with increased epochs generally resulting in better performance, particularly notable in the CTGAN model. Small datasets pose challenges for CTGAN but increasing epochs aids in mitigating modeling difficulties.

### 3.3 $\epsilon$ -Identifiability

Table 5 illustrates the  $\epsilon$ -identifiability values across different models and epochs, representing the likelihood of individuals being identified by synthetic observations. For CTGAN, there is a trend of increasing  $\epsilon$ -identifiability values, suggesting a higher risk of individuals being identified as epochs increase. Conversely, LLM generally exhibits higher  $\epsilon$ -identifiability values across epochs compared to CTGAN, indicating a potentially greater risk of identification with this model. When combined with the findings

**Table 3.** Mean Random Forest Accuracy. For classification tasks the mean  $F_1$  value is reported and for regression tasks the mean  $R^2$  value is reported.

model	epochs	AB	AD	BD	BP	CH	CP	DB	IN	KI
original	-	.554 ± .002	.856 ± .001	.849 ± .001	.852 ± .001	.852 ± .001	.882 ± .009	.749 ± .013	.872 ± .001	.854 ± .004
ctgan	300	.337 ± .011	.825 ± .003	.790 ± .006	.438 ± .009	.787 ± .007	.162 ± .027	.580 ± .039	.293 ± .125	.657 ± .010
ctgan	1,000	.353 ± .012	.832 ± .003	.820 ± .005	.513 ± .008	.789 ± .007	.256 ± .026	.680 ± .023	.577 ± .038	.769 ± .006
ctgan	10,000	.446 ± .011	.839 ± .003	.831 ± .005	.541 ± .008	.800 ± .005	.664 ± .023	.725 ± .024	.746 ± .028	.784 ± .006
llm	400	.515 ± .009	.851 ± .002	.841 ± .004	.676 ± .007	.835 ± .005	.775 ± .013	.739 ± .026	.734 ± .031	.239 ± .011
llm	1,000	.525 ± .008	.851 ± .002	.842 ± .003	.666 ± .007	.848 ± .004	.814 ± .018	.747 ± .025	.808 ± .022	.324 ± .013

from Table 2, this demonstrates the utility-risk trade-off in generative modeling. LLMs consistently outperform CTGAN models, as indicated by lower Wasserstein distances, reflecting a closer alignment with real data distributions. However, the observed high  $\epsilon$ -identifiability values underscore the potential limitations of using  $\epsilon$ -identifiability as a sole risk measure (see Table 4), suggesting the need for further exploration into more suitable risk assessment metrics.

**Table 4.**  $\epsilon$ -identifiability, i.e., the proportion of records that are at risk of being identified by synthetic observations using k-nearest neighbor.

model	epochs	AB	AD	BD	BP	CH	CP	DB	IN	KI
ctgan	300	.079 ± .003	.244 ± .003	.324 ± .004	.205 ± .004	.432 ± .005	.418 ± .011	.141 ± .013	.202 ± .010	.214 ± .003
ctgan	1,000	.146 ± .005	.204 ± .002	.366 ± .004	.242 ± .004	.403 ± .005	.437 ± .010	.197 ± .015	.274 ± .010	.237 ± .003
ctgan	10,000	.240 ± .005	.283 ± .003	.419 ± .004	.275 ± .004	.367 ± .004	.435 ± .010	.415 ± .016	.310 ± .012	.277 ± .003
llm	400	.563 ± .007	.631 ± .003	.579 ± .004	.529 ± .004	.658 ± .004	.451 ± .011	.546 ± .020	.605 ± .012	.292 ± .003
llm	1,000	.615 ± .007	.604 ± .003	.603 ± .003	.567 ± .004	.653 ± .004	.681 ± .009	.639 ± .017	.636 ± .012	.312 ± .003

### 3.4 Shared Neighbor Identifiability

Table 5 presents the SNI-identifiability values, representing the proportion of records vulnerable to identification by synthetic observations using shared nearest neighbor analysis. Once again, across different epochs and models, CTGAN generally demonstrates lower SNI-identifiability values compared to LLM. As the number of epochs increases, both CTGAN and LLM exhibit a trend towards higher SNI-identifiability values, suggesting an elevated risk of identification with prolonged training. Interestingly, SNI values consistently appear lower than the corresponding  $\epsilon$ -identifiability values and are more aligned with utility values. For instance, on the KI dataset, LLM performs inferiorly to CTGAN, corresponding with lower risk values. Overall, SNI appears to better capture the identification risk and is more suitable for addressing the trade-off between risk and utility (see Table A5 in the appendix for a record example).

**Table 5.** SNI-identifiability, i.e., the proportion of records that are at risk of being identified by synthetic observations using shared nearest neighbor.

model	epochs	AB	AD	BD	BP	CH	CP	DB	IN	KI
ctgan	300	.038 ± .004	.113 ± .002	.111 ± .003	.079 ± .003	.159 ± .003	.163 ± .012	.085 ± .014	.107 ± .010	.114 ± .002
ctgan	1,000	.058 ± .003	.107 ± .002	.118 ± .001	.082 ± .002	.150 ± .003	.167 ± .009	.110 ± .014	.129 ± .011	.121 ± .002
ctgan	10,000	.087 ± .004	.117 ± .002	.128 ± .001	.097 ± .002	.135 ± .003	.161 ± .010	.165 ± .013	.128 ± .010	.131 ± .003
llm	400	.152 ± .009	.163 ± .002	.180 ± .003	.136 ± .002	.196 ± .004	.156 ± .007	.172 ± .020	.178 ± .005	.106 ± .002
llm	1,000	.200 ± .009	.173 ± .002	.206 ± .002	.178 ± .003	.209 ± .005	.232 ± .013	.229 ± .019	.200 ± .016	.112 ± .002

## 4 Discussion

LLMs outperform CTGAN by generating synthetic data that more closely matches real data distributions, as evidenced by lower Wasserstein distances. LLMs also generally provide better predictive performance compared to CTGAN, with higher  $F_1$  and  $R^2$  scores. Interestingly, this does not necessarily mean that LLMs better capture correlations. Our hypothesis is that the dependencies are more complex than linear relationships, so not much can be inferred from the correlation matrix. While LLMs offer superior utility, they come at the cost of higher  $\epsilon$ -identifiability and SNI values, indicating greater identification risk. These observations highlight the importance of considering both utility and risk when evaluating generative models and the need for careful selection of risk measures to ensure robust and practical applications in real-world scenarios.

The observed differences in performance between LLMs and CTGAN can be partly attributed to the parameter sizes of the models. LLMs typically have a larger number of parameters, enabling them to capture more complex patterns and dependencies within the data. This also explains why they do not benefit as significantly from an increased number of training epochs compared to CTGAN models. CTGAN models show more pronounced performance gains with additional epochs because the initial models might be underfitting, and longer training helps them converge towards a better approximation of the data distribution. The conditioning and mode-specific normalization in CTGAN is highly batch-dependent, making it crucial to process more batches to achieve stability. In contrast, LLMs, with their attention mechanisms, are less affected by differing batch characteristics, allowing faster convergence.

In our results, we observed a consistent relationship between lower Wasserstein distance and higher specific utility, suggesting that Wasserstein distance can serve as an indicator of model performance, particularly when enriching specific analyses with synthetic datasets. A significant advantage of specific utility metrics is their ability to assess the generalization capability by using the traditional validation pipeline of machine learning algorithms. Therefore, in our final analysis, whenever possible, we will also rely on these more specialized metrics to provide a comprehensive evaluation of model performance and generalization ability.

Our proposed new risk measure, SNI, consistently remains lower than the corresponding  $\epsilon$ -identifiability values. Moreover, it effectively captures the trade-off between risk and utility; in most cases, as utility increases, more risk must be accepted. This observation underscores the efficacy of SNI in accurately assessing potential risks. For example, on the KI dataset, LLM shows inferior performance to CTGAN, as reflected

in lower risk values. Overall, the SNI metric emerges as a robust tool for addressing the delicate balance between identification risk and utility in generative modeling scenarios. Further refinements will involve exploring various density measurements, aimed at improving the accuracy and applicability of SNI across a wide range of datasets and modeling techniques.

In our analysis, it became apparent that the mix of variable types poses a significant challenge in distance calculation. This renders the metric spaces much more complex than can be captured solely by Euclidean distance, for example. While one-hot encoding of discrete variables can embed them into a metric space, challenges persist, particularly in scaling and normalization. It is important to recognize that this is not a minor issue but manifests itself in various aspects. For instance, research on Wasserstein GANs suggests that their effectiveness stems from not accurately reflecting Wasserstein distance [30]. This implies that we often do not fully understand the behavior of our metrics in high-dimensional space, especially when variables exhibit mixed feature types. We encounter a similar problem in a modified form in the context of record linkage, where attempts are made to capture feature value differences using string metrics, which, however, no longer work adequately for numerical values.

Computing word embeddings could provide a solution to the challenges posed by mixed feature types in the metric space, offering greater flexibility in modeling. Word embeddings allow for the representation of categorical features in a continuous vector space, enabling a unified encoding of all variable types. By capturing semantic relationships between words based on contextual usage, word embeddings offer a more nuanced understanding of categorical feature values. This enhanced representation enables the better capture of complex patterns and dependencies within the data, potentially leading to improved model performance. This may be one of the reasons why our LLMs performed well, as we encountered no issues with mixed feature types when using word embeddings. However, transitioning word embeddings into other model classes besides neural networks remains a challenge.

In summary, when evaluating generative models such as LLMs and CTGAN, researchers and practitioners can use Wasserstein distance as a preliminary indicator of model performance. Lower Wasserstein distances generally indicate better performance in capturing the underlying data distribution, leading to higher utility in downstream tasks. LLMs and CTGANs have different strengths and are suited to different use cases based on their performance characteristics. LLMs, with their robust parameterization and ability to capture complex patterns, excel in scenarios where high-fidelity SDG is critical, such as in healthcare or finance. Additionally, LLMs are advantageous when there is a need to model mixed feature types, as they can leverage techniques like word embeddings to mitigate the challenges posed by diverse data types. On the other hand, CTGANs exhibit pronounced performance gains with increased training epochs, making them suitable for applications where extended training can be accommodated. Furthermore, CTGANs may be preferred in scenarios where computational resources are limited, as they can achieve competitive performance with fewer parameters compared to LLMs. Overall, the choice between LLMs and CTGANs should be informed by the specific requirements of the application, including data complexity, computational resources, and the desired level of model interpretability.

**Acknowledgments.** This study was funded by BRIDGE, a joint programme of the Swiss National Science Foundation SNSF and Innosuisse (grant number 211751).

## Appendix

**Table A1.** URLs of real-world datasets for the study

Dataset	URL
AB	<a href="https://archive.ics.uci.edu/dataset/1/abalone">https://archive.ics.uci.edu/dataset/1/abalone</a>
AD	<a href="https://archive.ics.uci.edu/dataset/2/adult">https://archive.ics.uci.edu/dataset/2/adult</a>
BD	<a href="https://www.kaggle.com/datasets/akash14/adopt-a-buddy">https://www.kaggle.com/datasets/akash14/adopt-a-buddy</a>
BP	<a href="https://www.kaggle.com/datasets/kukuroo3/body-performance-data">https://www.kaggle.com/datasets/kukuroo3/body-performance-data</a>
CH	<a href="https://www.kaggle.com/datasets/shubh0799/churn-modelling">https://www.kaggle.com/datasets/shubh0799/churn-modelling</a>
CP	<a href="https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification">https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification</a>
DB	<a href="https://www.kaggle.com/datasets/mathchi/diabetes-data-set">https://www.kaggle.com/datasets/mathchi/diabetes-data-set</a>
IN	<a href="https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender">https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender</a>
KI	<a href="https://www.kaggle.com/datasets/harlfoxem/housesalesprediction">https://www.kaggle.com/datasets/harlfoxem/housesalesprediction</a>

**Table A2.** General utility in terms of the sum between Wasserstein distance (solely for numerical features) and Hellinger distance for categorical data on multiple datasets and epochs.

model	epochs	AB	AD	BD	BP	CH	CP	DB	IN	KI
ctgan	300	.071 ± .006	.101 ± .002	.186 ± .002	.088 ± .002	.142 ± .003	.694 ± .010	.275 ± .018	.148 ± .009	.190 ± .002
ctgan	1,000	.048 ± .006	.097 ± .002	.142 ± .002	.080 ± .002	.128 ± .003	.768 ± .013	.246 ± .018	.078 ± .009	.175 ± .002
ctgan	10,000	.031 ± .006	.087 ± .002	.113 ± .002	.083 ± .003	.108 ± .003	.715 ± .013	.161 ± .018	.074 ± .009	.172 ± .002
llm	400	.018 ± .006	.075 ± .003	.033 ± .003	.024 ± .004	.037 ± .014	.382 ± .045	.076 ± .019	.049 ± .016	.278 ± .004
llm	1,000	.016 ± .006	.071 ± .003	.030 ± .004	.026 ± .004	.042 ± .014	.343 ± .041	.062 ± .022	.029 ± .018	.305 ± .005

The experiments were conducted on a server running Ubuntu 22.04 LTS and utilizing Cuda Toolkit 12.4 drivers. Virtual environments were created with Conda. The hardware setup included two NVIDIA H100 PCIe 80 GB graphics cards, two Intel Xeon Silver 4416+ processors, and 1008 GB of RAM.

**Table A3.** Mean absolute  $L^2$  correlation matrix distance.

model	epochs	AB	AD	BD	BP	CH	CP	DB	IN	KI
ctgan	300	1.939 ± .020	1.191 ± .016	.732 ± .011	1.581 ± .013	.693 ± .012	1.973 ± .029	1.857 ± .050	1.252 ± .040	2.349 ± .017
ctgan	1,000	1.104 ± .018	1.021 ± .014	.658 ± .012	1.171 ± .014	.704 ± .014	2.056 ± .032	1.474 ± .048	.625 ± .043	1.975 ± .019
ctgan	10,000	.440 ± .012	1.740 ± .014	.529 ± .011	.920 ± .016	.885 ± .019	1.517 ± .034	.974 ± .059	.371 ± .037	2.065 ± .022
ilm	400	.385 ± .065	1.595 ± .144	1.295 ± .011	.527 ± .369	.552 ± .057	1.384 ± .184	.704 ± .160	.670 ± .038	3.404 ± .313
ilm	1,000	.358 ± .080	1.425 ± .122	1.147 ± .027	.550 ± .172	.524 ± .044	.761 ± .122	.467 ± .116	.604 ± .052	3.163 ± .319

**Table A4.** Utilized random forest estimator hyperparameters for each dataset.

	n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features	bootstrap
Search Grid	[50, 100, 200]	[None, 10, 20]	[2, 5, 10]	[1, 2, 4]	['auto', 'sqrt']	[True, False]
AB	100	20	2	2	sqrt	True
AD	50	20	5	4	sqrt	False
BD	200	20	5	2	sqrt	True
BP	200	20	10	1	sqrt	False
CH	50	None	10	4	sqrt	True
CP	200	None	5	4	sqrt	True
DB	200	10	5	4	sqrt	True
IN	100	10	2	2	sqrt	False
KI	200	20	2	1	sqrt	False

**Table A5.** Record that is  $\varepsilon$ -identifiable but not SNI-identifiable.

observation	preg	Plas	Pres	Skin	Insu	Mass	Pedi	Age	class
$x_i$	1.0	139.0	46.0	19.0	83.0	28.7	0.654	22.0	0.0
$r_i$	3.0	113.0	50.0	10.0	85.0	29.5	0.626	25.0	0.0
$\hat{r}_i$	0.0	137.0	58.0	15.0	85.0	24.6	0.766	21.0	0.0

## References

1. Carvalho, T., Moniz, N., Faria, P., Antunes, L.: Survey on privacy-preserving techniques for microdata publication. ACM Comput. Surv. **55**, 309:1–309:42 (2023)
2. Sariyar, M., Schlünder, I.: Reconsidering anonymization-related concepts and the term “identification” against the backdrop of the European legal framework. Biopreserv. Biobank. **14**, 367–374 (2016). <https://doi.org/10.1089/bio.2015.0100>
3. Hu, J., Savitsky, T.D., Williams, M.R.: Private tabular survey data products through synthetic microdata generation. J. Surv. Stat. Methodol. **10**, 720–752 (2022). <https://doi.org/10.1093/jssam/smac001>

4. Buczak, A.L., Babin, S., Moniz, L.: Data-driven approach for creating synthetic electronic medical records. *BMC Med. Inform. Decis. Mak.* **10**, 59 (2010). <https://doi.org/10.1186/1472-6947-10-59>
5. Almasi, M.M., Siddiqui, T.R., Mohammed, N., Hemmati, H.: The risk-utility tradeoff for data privacy models. In: 2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS), pp. 1–5 (2016)
6. Raghunathan, T.E.: Synthetic data. *Annu. Rev. Stat. Appl.* **8**, 129–140 (2021). <https://doi.org/10.1146/annurev-statistics-040720-031848>
7. Li, Z., Zhao, Y., Fu, J.: SynC: a copula based framework for generating synthetic data from aggregated sources. In: 2020 International Conference on Data Mining Workshops (ICDMW), pp. 571–578 (2020). <https://doi.org/10.1109/ICDMW51313.2020.00082>
8. Kaur, D., et al.: Application of Bayesian networks to generate synthetic health data. *J. Am. Med. Inform. Assoc.* **28**, 801–811 (2021). <https://doi.org/10.1093/jamia/ocaa303>
9. Fonseca, J., Bacao, F.: Tabular and latent space synthetic data generation: a literature review. *J. Big Data* **10**, 115 (2023). <https://doi.org/10.1186/s40537-023-00792-7>
10. Neves, D.T., Alves, J., Naik, M.G., Proen  a, A.J., Prasser, F.: From missing data imputation to data generation. *J. Comput. Sci.* **61**, 101640 (2022)
11. Mukherjee, M., Khushi, M.: SMOTE-ENC: a novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Appl. Syst. Innov.* **4**, 18 (2021)
12. Zhang, J., Chen, L.: Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. *Comput. Assist. Surg.* (2019)
13. Razghandi, M., Zhou, H., Erol-Kantarci, M., Turgut, D.: Variational autoencoder generative adversarial network for synthetic data generation in smart home. In: IEEE International Conference on Communications, ICC 2022, pp. 4781–4786 (2022). <https://doi.org/10.1109/ICC45855.2022.9839249>
14. Little, C., Elliot, M., Allmendinger, R., Samani, S.S.: Generative adversarial networks for synthetic data generation: a comparative study (2021). <http://arxiv.org/abs/2112.01925>
15. Fang, X., et al.: Large Language Models (LLMs) on tabular data: prediction, generation, and understanding – a survey (2024). <http://arxiv.org/abs/2402.17944>
16. de Melo, C.M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., Hodgins, J.: Next-generation deep learning based on simulators and synthetic data. *Trends Cogn. Sci.* **26**, 174–187 (2022). <https://doi.org/10.1016/j.tics.2021.11.008>
17. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data – anonymisation groundhog day. Presented at the 31st USENIX Security Symposium (USENIX Security 2022) (2022)
18. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN (2019). <http://arxiv.org/abs/1907.00503>. <https://doi.org/10.48550/arXiv.1907.00503>
19. Zhao, Z., Birke, R., Chen, L.: TabuLa: harnessing language models for tabular data synthesis (2023). <https://doi.org/10.48550/ARXIV.2310.12746>
20. Yoon, J., Drumright, L.N., van der Schaar, M.: Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE J. Biomed. Health Inform.* **24**, 2378–2388 (2020). <https://doi.org/10.1109/JBHI.2020.2980262>
21. Gao, X., Deng, F., Yue, X.: Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty. *Neurocomputing* **396**, 487–494 (2020). <https://doi.org/10.1016/j.neucom.2018.10.109>
22. Borisov, V., Se  ler, K., Leemann, T., Pawelczyk, M., Kasneci, G.: Language models are realistic tabular data generators (2023). <http://arxiv.org/abs/2210.06280>. <https://doi.org/10.48550/arXiv.2210.06280>
23. Simpson, D.G.: Minimum Hellinger distance estimation for the analysis of count data. *J. Am. Stat. Assoc.* **82**, 802–807 (1987). <https://doi.org/10.1080/01621459.1987.10478501>

24. Piccoli, B., Rossi, F.: On properties of the generalized Wasserstein distance. *Arch. Ration. Mech. Anal.* **222**, 1339–1365 (2016). <https://doi.org/10.1007/s00205-016-1026-7>
25. Flamary, R., et al.: POT: python optimal transport. *J. Mach. Learn. Res.* **22**, 1–8 (2021)
26. Wang, W., Ying, L., Zhang, J.: On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Trans. Inf. Theory* **62**, 5018–5029 (2016). <https://doi.org/10.1109/TIT.2016.2584610>
27. Liu, R., Wang, H., Yu, X.: Shared-nearest-neighbor-based clustering by fast search and find of density peaks. *Inf. Sci.* **450**, 200–226 (2018). <https://doi.org/10.1016/j.ins.2018.03.031>
28. Khan, K., Rehman, S.U., Aziz, K., Fong, S., Sarasvady, S.: DBSCAN: past, present and future. In: The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), pp. 232–238 (2014). <https://doi.org/10.1109/ICA-DIWT.2014.6814687>
29. Sariyar, M., Holm, J.: On the concepts of identity and similarity in the context of biomedical record linkage. In: Public Health and Informatics, pp. 472–476. IOS Press (2021). <https://doi.org/10.3233/SHTI210203>
30. Stanczuk, J., Etmann, C., Kreusser, L.M., Schönlieb, C.-B.: Wasserstein GANs work because they fail (to approximate the Wasserstein distance) (2021). <http://arxiv.org/abs/2103.01678>. <https://doi.org/10.48550/arXiv.2103.01678>

## **Case Studies**



# Escalation of Commitment: A Case Study of the United States Census Bureau Efforts to Implement Differential Privacy for the 2020 Decennial Census

Krishnamurty Muralidhar<sup>1</sup> and Steven Ruggles<sup>2</sup>

<sup>1</sup> University of Oklahoma, Norman, OK 73019, USA

[krishm@ou.edu](mailto:krishm@ou.edu)

<sup>2</sup> University of Minnesota and Director of IPUMS, Minneapolis, MN 55455, USA

[ruggles@umn.edu](mailto:ruggles@umn.edu)

**Abstract.** In 2017, the United States Census Bureau announced that because of high disclosure risk in the methodology (data swapping) used to produce tabular data for the 2010 census, a different protection mechanism based on differential privacy would be used for the 2020 census. While there have been many studies evaluating the result of this change, there has been no rigorous examination of disclosure risk claims resulting from the released 2010 tabular data. In this study we perform such an evaluation. We show that the procedures used to evaluate disclosure risk are unreliable and resulted in inflated disclosure risk. Demonstration data products released using the new procedure were also shown to have poor utility. However, since the Census Bureau had already committed to a different procedure, they had no option except to escalate their commitment. The result of such escalation is that the 2020 tabular data release offers neither privacy nor accuracy.

**Keywords:** Census · disclosure · privacy

## 1 Introduction

*Escalating commitment (or escalation) refers to the tendency for decision makers to persist with failing courses of action.* (Brockner 1992).

To date, there has never been a documented case of reidentification (the ability to identify a real individual based on summary data released to the public) using the decennial tabular data released to the public. Yet, in 2017, the Chief Scientist of the Census Bureau (CB) issued the following statement (Abowd 2017): “This [database reconstruction] theorem is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared.” No evidence to support this statement was offered at this time. This was followed by the announcement in September 2017 (Garfinkel 2017): “The 2020 Census disclosure avoidance system will use differential

privacy to defend against a reconstruction attack.” There was still no evidence being offered that there was even a need for a change since reconstruction does not necessarily lead to reidentification.

Starting in 2018, there has been a string of reports from CB regarding the reconstruction and reidentification experiments regarding the 2010 Census tabular data release. The culminating claim is that approximately 92% of the data is accurately reconstructed and over 75% of the respondents in the 2010 Census data have been identified (Hawes 2022). Unfortunately, CB did not release any details of the reconstruction and reidentification procedure until 2021 (Abowd 2021) which meant that the specific claims of reconstruction and reidentification could not be corroborated. In the meantime, the 2020 decennial census data collection had already been completed and the focus shifted to an assessment of the accuracy of the output from the differentially private procedure (DPP) implementation for the 2020 census data. Although a few studies have addressed the 2010 reconstruction and reidentification claims (Francis 2022; Muralidhar 2022; Muralidhar and Domingo-Ferrer 2023a, 2023b; Ruggles 2018, 2024), a comprehensive examination of the claims is still missing.

## 2 Reconstruction and Reidentification in 2010 Tabular Data Release

A closer examination of the claims made by the CB as it relates to reconstruction and reidentification reveals the lack of rigor in reconstruction, randomness in putative identification, and a logical fallacy in the confirmation of identification.

### 2.1 Reconstruction

In 2010, Census data was collected at the household level. Apart from one question regarding type of housing, the remaining questions collect information on Person 1 (the owner/renter) followed by other residents of the household. For all residents, the information collected included Sex, Age, Race, and Ethnicity. For residents other than Person 1 (who is defined as the Householder), the relationship of the resident to Person 1 (14 categories) was also collected. The CB applied a disclosure limitation procedure (swapping) to protect vulnerable respondents, yielding a protected database. All tables were created using only the protected data. At the block level, the Census Bureau released 235 tables regarding both households and residents of the households.

A comprehensive reconstruction of both households and respondents by an adversary would involve reconstructing attribute values for every resident and using this information to assign them to specific households. The adversary can verify the accuracy of their reconstruction by simply verifying that the reconstructed data is consistent with all 235 tables released to the public. It is important to note that an adversary can reconstruct only the protected data from which the tables were created and never the original data (since this data is never used). Hence, the vulnerable individuals are still protected.

The Census Bureau attempted to reconstruct the original individual-level responses to the 2010 Census using only the published tabular data. There was no attempt to assign persons to households, and the reconstruction did not include the Relationship

attribute. Thus, in each block, only (Sex, Age Group, Race, Ethnicity) are reconstructed (Hawes 2022). For a significant majority of the data, Ruggles (2018) pointed out that this reconstruction amounts to nothing more than looking up a set of tables, with absolutely no need for any optimization algorithms. Abowd et al. (2023, page 19) provide an extensive description of an optimization model using integer linear programming to calculate the number of Asian Hispanic Females aged 22 on a particular block. We now illustrate how this question can be answered using the information from published tables and simple arithmetic.

For a significant majority of the blocks, such reconstruction can be performed using only tables P12A-I, P8, and P9. Tables P8 and P9 provide, at the block level, the count of all individuals and the count of non-Hispanic individuals, respectively. Tables P12A–P12G provide the count of all individuals by sex and age-group for seven race categories (White, Black/African American, American Indian or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Other races, and Two or more races, respectively). Table P12H provides, by sex and age, the count of Hispanic individuals in the block. Finally, table P12I provides, by sex and age, the count of non-Hispanic White individuals. All these tables are available at <https://data.census.gov/>. Muralidhar (2022) showed that to the CB attempt to reconstruct individual year-of-age produced was little better than random assignment. Perhaps in recognition of the difficulty of reconstructing single years of age from the published census tables, in recent publications the Census Bureau as focused on reconstructing grouped ages (e.g. 0–4, 5–9).

Table 1 provides the information from tables P8 and P9 (Part A), and data extracted from tables P12A-I for females in age-group 22–24 (Part B), from Block 3000, Tract 72, in New York County, New York. Table 1 (Part C) also provides the completed block counts computed using Parts A & B as follows:

- (1) From P12I, the count of White, non-Hispanic females in age-group 22–24 is readily available as 43. Subtracting this count from all White females in age-group 22–24 (P12A) provides the count of White, Hispanic females.
- (2) From P8 and P9, the count of Asian, Hispanic females in the block is 0. Hence, all 5 Asian females in this age-group must be non-Hispanic.
- (3) From table P12H, there are 2 Hispanic females in the 22–24 age-group of which one is White. Hence, the remaining Hispanic female must belong to the “Other” race category.

Now that the table has been constructed, reconstruction is simply a matter of creating a list of these 50 respondents. All this is performed using simple arithmetic without any need for any additional computation.

It is important to note that Table 1 is not an isolated example. *Every White individual can be reconstructed* (over 74% of the population). In 88% of the occupied blocks (5,462,028 of 6,207,027), *everyone in the block can be reconstructed*. For over 90% of the population, *every age group belonging to a (Block, Race) can be reconstructed*. Examination of individual age-groups (as shown in Table 1) increases the proportion of the reconstructed population.

CB has repeatedly claimed that reconstruction was possible only because of improvements in technology and algorithms (Abowd 2021). Our illustration above shows that this claim is demonstrably false. This study has demonstrated that for over 90% of the

**Table 1.** Reconstruction of Females, Age group (22–24), Block 3000, Tract 72, in New York County, New York (NH = Not Hispanic, H = Hispanic). Values in bold in Part C are computed from the information in Parts A and B.

Part A: Count of Individuals by Race (Entire Block)				Part B: Count of Female, Age Group (22–24) (By Race)			Part C: Complete reconstruction of race and ethnicity of Female, Age group (22–24)		
All (P8)	NH (P9)	H (P8–P9)		All	NH	H	All	NH	H
474	447	27	Total	50 (P12)		2 (P12H)	50	<b>48</b>	2
377	355	22	White	44 (P12A)	43 (P12I)		44	43	<b>1</b>
10	9	1	Black	0 (P12B)			0	<b>0</b>	<b>0</b>
1	1	0	AIAN	0 (P12C)			0	<b>0</b>	<b>0</b>
80	80	0	Asian	5 (P12D)			5	<b>5</b>	<b>0</b>
0	0	0	NHPI	0 (P12E)			0	<b>0</b>	<b>0</b>
6	2	4	Other	1 (P12F)			1	<b>0</b>	<b>1</b>
9	8	1	Two or more races	0 (P12G)			0	<b>0</b>	<b>0</b>

population, reconstruction is little more than looking up a set of tables, just as Ruggles (2018) had predicted.

It is important to note that CB researchers have recently claimed that implementing additional features “is straightforward” and provide a formulation for the relationship variable (Abowd et al. 2023). This is misleading since formulating a problem does not mean that an unique solution will be found. Unlike (Sex, Age Group, Race) where there is three-way cross tabulation by block (tables P12A-I), there are no such cross tabulations for the relationship variable. Without cross tabulations, it is highly likely that the lack of constraints will result in non-unique solutions in most cases. CB have *never* published any results of reconstruction which *includes the relationship variable*.

## 2.2 Putative Reidentification

Reconstruction as defined by the CB does not necessarily pose a confidentiality risk. To demonstrate a disclosure threat, the Census Bureau matched their reconstructed data to an external commercial data source that includes people's identities.

In each block, even though the reconstruction involved four attributes (Sex, Age, Race, Ethnicity), matching with the external source was performed using only two attributes (Sex, Age). As suggested by the CB's Research & Methodology Group (RMG) (our emphasis), "such reidentification studies are performed by looking for *unique combinations of variables in the microdata* that are thought to be identifying, looking for externally available data sets that contain the same variables, and then linking data records in the two data sets using the linkage variables." (McKenna 2019) But the CB putative reidentification performs matching on *all* the reconstructed records, not just those with unique combinations of characteristics (Abowd 2021). There is a significant problem with this approach.

Consider a block which consists of 10 females in the (22–24) age-group all of whom are non-Hispanic White. The CB reidentification used only two attributes (Sex, Age) as linkage variables. As a result, these people are indistinguishable from one another, and *any individual can be assigned any identity from the external source data*. Because of this, all non-unique (random) matches will be dismissed as unprovable. This is precisely the reason that RMG considers only records that are uniquely identifiable based on the linkage variables. But by "looping through all the records", the new procedure used by CB contradicts RMG recommendations and treats these matches as putative identification.

At the national level, only 17.4% of the people are uniquely identifiable by (Sex, Age Group). *This represents a strict upper bound on the proportion of people who can be putatively reidentified using (Sex, Age)*. By ignoring this criterion, CB claims putative identification at a remarkable 97%. This implies that *close to 80% of the putative reidentifications are random and unprovable*.

## 2.3 Confirmed Reidentification

Hawes (2022) claims that the identity of approximately 78% of the randomly identified people during putative identification are "confirmed". The CB considers cases confirmed if the putative matches (where the reconstructed data match the commercial data on age, sex, and block) also match on their Protected Identification Key (PIK). The PIK is a unique identifier the CB assigns to both the original census data and the commercial data based on name, address, age, and sex. (Abowd 2021, Appendix B, para 18, page 7) According to Wagner and Layne (2014) PIK "is an anonymous identifier as unique as a SSN".

The steps in the confirmation of reidentification are described in Abowd 2021 (Appendix B, para 21, page 8). We describe these steps (including the putative reidentification steps for completeness):

- (1) Putative reidentification: Match (Age, Sex) from the reconstructed data to the external commercial data.
  - a. If a unique match is found, append PIK from source data to the reconstructed data.

- b. If there are no unique matches, *randomly append* PIK from source data to the reconstructed data.
- (2) Confirmation: Compare (PIK, Age, Sex, Race, Ethnicity) in the reconstructed data with original Census data. If they match, it confirms identity.

The problem with using the PIK to confirm whether the reconstructed data can be tied to real identities is that the reconstructed data do not have a PIK: the only attributes in the reconstructed data are age, sex, race, ethnicity, and block. The PIK in the putative reidentified data is appended from the commercial data.

We return to the example of the 10 (Female, 22–24, White, non-Hispanic) individuals who were *randomly assigned* a (Name, Address, PIK) during putative reidentification. Using the identity confirmation procedure, *the identity of all 10 individuals is confirmed*. But this confirmation is a fallacy. Comparison of (Age, Sex, Race, Ethnicity) for these individuals is the same and hence does not confirm identity. *PIK is the only variable that uniquely confirms identity*. But PIK is the identifier. This is a textbook example of circular logic fallacy.

Modifying this approach slightly takes it to its absurd but logical conclusion. Assume that the adversary only creates a list of individuals in a block but does not reconstruct any of the variables using the tabular data. PIK is assigned to each person in this block in the external commercial data. Using CB procedure, the identity of every individual is confirmed, since PIK's match. *All you need to confirm the identity is knowledge of their identity!* As Ruggles (2024) points out, this amounts to nothing more than the fact that every person in this block has a PIK, not identity. If CB had followed RMG's recommended procedure, it is unlikely that the identity of *any* of the individuals would have been confirmed. Since CB did not follow the correct procedure, the confirmed reidentification results are vastly exaggerated.

### 3 Policies and Their Impact on Data Usefulness

The policies adopted by the Census Bureau impose constraints on the implementation of the disclosure limitation method. For decennial 2010 Census data, we know that there were two specific legal requirements that were in place. First, Block level population counts for total population and voting-age population (aged 18+) must be maintained (Dajani et al. 2017). Second, the data was also required to maintain consistency between person and housing tables. In 2010, these two requirements played a crucial role in the way the disclosure prevention method (data swapping) was implemented. Although CB considered alternative levels of data swapping, no relaxation of the primary requirements was ever considered.

DP-based disclosure prevention methodology for the decennial 2020 Census *does not satisfy either requirement*. Although initially CB promised that block level voting and non-voting age counts will be preserved (Dajani et al. 2017), they later changed the policy in which these counts are preserved only at the state level (Abowd et al. 2020). This is three levels of geography higher than the block level (with block-group, tract, and county being the intermediate levels). Recently, the Census Bureau also announced that the tabular data released from the 2020 Decennial Census will *not maintain consistency between person and household tables*. This implies that the population count for a given

block, block-group, tract, and county level will be different based on person tables compared to the housing tables. It is estimated that over 10% of the data may contain such inconsistencies (Menger 2023). Some of these inconsistencies include:

- (1) Blocks completely under water.
- (2) Blocks with only children and no adults.
- (3) Blocks with individuals but no households.
- (4) Blocks with households but no individuals.

Researchers have also identified that block level data regarding minorities are quite inaccurate (Kenny et al. 2021, 2024). The Census Bureau has also advised *against using block level results* since they are not accurate. In summary, compared to the 2010 tabular data release, the quality of the DPP output is poor.

## 4 What About Privacy?

One of the key features in the implementation of differential privacy is the selection of the privacy parameter  $\varepsilon$ . The smaller the value of  $\varepsilon$ , the greater the privacy protection and vice versa. The value  $e^\varepsilon$  represents the privacy loss.

CB considered two alternate procedures:

- (1) Bottom-up approach – where Laplace noise is added to every cell in the combination of (Block  $\times$  Sex  $\times$  Age  $\times$  Race  $\times$  Ethnicity) which results in 161,109,592,812 cells. This procedure was deemed to add too much noise to the data and hence not considered.
- (2) Top-down approach - where concentrated differential privacy is employed to reduce the amount of noise added (Abowd and Hawes 2023). This procedure does not provide strict  $\varepsilon$ -DP but only a relaxed version  $(\varepsilon, \delta)$ -DP. It also involves extensive post-processing to ensure that some basic requirements (non-negativity, integer values, etc.) are met.

Originally, CB chose the value of  $\varepsilon$  to be 4.2. Using this specification, CB generated a demonstration data set protected by the top-down procedure using 2010 data. This data was released to the public to allow users to assess the accuracy of DPP compared to the originally released data. The public reaction to the demonstration data was quite negative due to poor accuracy. As a result, CB increased the value of  $\varepsilon$  to 10.61 and released a second demonstration data set, which was also received poorly. CB increased  $\varepsilon$  to 19.41 and released a third demonstration data set. Finally, CB settled on a value of  $\varepsilon = 39.91$  for individual demographic data.

One of the problems with focusing on the value of  $\varepsilon$  is that, by itself, it provides no meaningful assessment of the privacy risk. It represents the log of the odds ratio (ratio of two probabilities), the probability of a negative outcome (disclosure) to the probability of a positive outcome (non-disclosure). From the very beginning, the developers of differential privacy have been insistent that this ratio (the value of  $\varepsilon$ ) should be small. According to Dwork (2011), “We tend to think  $\varepsilon$  of as, say, 0.01, 0.1, or in some cases,  $\ln 2$  or  $\ln 3$ .” We can assess the impact of this by constructing the privacy loss for different values of  $\varepsilon$  (Table 2).

**Table 2.** Privacy loss for different values of  $\varepsilon$ 

$\varepsilon$	0.01	0.1	4.2	10.6	19.61	39.91
Privacy Loss	1.01	1.11	67	32860	325215956	215125935915741000

So how does  $\varepsilon$  of 39.91 compare to the suggested values? When  $\varepsilon = 0.1$ , the odds of negative outcome to positive outcome are 1.11:1; with  $\varepsilon = 39.91$ , the odds are 215125935915741000:1.

Referring to the use of the value of  $\varepsilon = 14$  by Apple in one of their applications, McSherry, one of the co-inventors of differential privacy, is quoted as saying: “Anything much bigger than one is not a very reassuring guarantee. Using an epsilon value of 14 per day strikes me as relatively pointless” (Greenberg 2017). The selection  $\varepsilon = 39.91$  is anything but reassuring.

## 5 Conclusions

The decision to change the disclosure avoidance methodology from data swapping in 2010 to DPP in 2020 was predicated upon the fact that the disclosure risk from reconstruction and reidentification of the 2010 tabular data release was unacceptable. Considering that the original data was first checked for vulnerabilities before data swapping was applied to protect the vulnerable individuals, it is very surprising that this did not reveal the potential for reidentification. CB has consistently offered the explanation that “While the Census Bureau’s confidentiality methodologies for the 2000 and 2010 censuses were considered sufficient at the time, advances in technology in the years since have reduced the confidentiality protection provided by data swapping.” (Abowd 2021; Keller and Abowd 2023). In this paper, we have shown that this statement is provably false. Using only a few tables and simple arithmetic, we have shown that we can reconstruct (Sex, Age Group, Race, Ethnicity) for over 90% of the population. Given that the publicly released tables included a crosstabulation of age by sex by race and ethnicity for each census block, it seems improbable that CB was unaware of this simple reconstruction at the time.

This then raises the question as to why CB chose to release this data if it was vulnerable to reconstruction and subsequent reidentification. One plausible explanation is that CB concluded that the risk of reidentification, as defined by CB’s own RMG, showed risk of identity disclosure to be very small. Even CB initially acknowledged that the risk of reidentification to be very small (Abowd 2018). Subsequently however, the risk of confirmed reidentification increased to as high as 78% (Hawes 2022). That massive increase in reidentification risk can be attributed exclusively to the change in the procedure used for putative and confirmed reidentification. This new procedure, which contradicts the RMG reidentification procedure, results in random reidentifications being classified as confirmed reidentifications, thereby considerably inflating the reidentification risk.

Finally, we believe that the decision by CB to adopt DPP was premature. The decision was announced in September 2017. At this time, it is unclear whether CB had thoroughly investigated the impact of implementing DPP. CB did not release the first demonstration

data product until October 2019, almost two years after the announcement of the decision to adopt DPP. Given the magnitude of the change, it might have been prudent for CB to have waited for their announcement until after they had received feedback from the users. It might also have been prudent for CB to consider alternate courses of action. Absent any options, the only course of action was to escalate their commitment to DPP implementation. The result of this escalation is that the output from the 2020 decennial census is inaccurate, inconsistent, and “differential privacy delivers privacy mostly in name.” (Dwork et al. 2011).

## References

- Abowd, J.M.: Research data centers, reproducible science, and confidentiality protection. In: The Role of the 21<sup>st</sup> Century Statistical Agency, Bureau Presentation to Summer Dem Sem Sponsored by the Wisconsin Federal Statistical RDC, 5 June 2017 (2017)
- Abowd, J.M.: Staring down the database reconstruction theorem. In: Joint Statistical Meetings, Vancouver, BC, Canada, 30 July 2018 (2018)
- Abowd, J.M.: Declaration of John M. Abowd, State of Alabama et al. v United States Department of Commerce et al., Case No. 3:21-CV-211-RAH-ECM-KCN (2021)
- Abowd, J.M., et al.: The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau (supersedes the 2017 version) (2020). <https://www.census.gov/library/working-papers/2020/adrm/CED-WP-2020-009.html>
- Abowd, J.M., et al.: The 2010 Census Confidentiality Protections Failed, Here’s How and Why (2023). <https://arxiv.org/pdf/2312.11283>
- Brockner, J.: The escalation of commitment to a failing course of action: toward theoretical progress. Acad. Manag. J. **17**(1), 39–61 (1992)
- Dajani, A.N., et al.: The modernization of statistical disclosure limitation at the U.S. Census Bureau (2017). <https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf>
- Dwork, C.: A firm foundation for private data analysis. Comm. ACM **54**(1), 86–95 (2011)
- Francis, P.: A note on the misinterpretation of the US. In: Domingo-Ferrer, J., Laurent, M. (eds.) Proceedings of Privacy in Statistical Databases - PSD 2022 (2022)
- Garfinkel, S.L.: Modernizing disclosure avoidance: report on the 2020 disclosure avoidance subsystem as implemented for the 2018 end-to-end test (continued). In: 2017 Census Scientific Advisory Committee Fall Meeting, Suitland, MD, 15 September 2017 (2017)
- Greenberg, A.: How one of apple’s key privacy safeguards falls short. Wired (2017). <https://www.wired.com/story/apple-differential-privacy-shortcomings/>
- Hawes, M.: Reconstruction and re-identification of the demographic and housing characteristics file (DHC). Presentation to the Census Scientific Advisory Committee (2022). <https://www2.census.gov/about/partners/cac/sac/meetings/2022-09/presentation-reconstruction-and-re-identification-of-dhc-file.pdf>
- Keller, S.A., Abowd, J.M.: Database reconstruction does compromise confidentiality. Proc. Natl. Acad. Sci. **120**(12), e2300976120 (2023)
- Kenny, C.T., McCartan, C., Kuriwaki, S., Simko, T., Imai, K.: Evaluating bias and noise induced by the U.S. Census Bureau’s privacy protection methods. Sci. Adv. **10**, eadl2524 (2024)
- Kenny, C.T., Kuriwaki, S., McCartan, C., Rosenman, E.T.R., Simko, T., Imai, K.: The use of differential privacy for census data and its impact on redistricting: the case of the 2020 U.S. census. Sci. Adv. **7**, eabk3283 (2021)

- McKenna, L.: U.S. Census Bureau Reidentification Studies. Working Paper Number CED-WP-2019-008 (2019). <https://www.census.gov/library/working-papers/2019/adrm/CED-WP-2019-008.html>
- Menger, E.: Transparency Matters (2023). <https://appliedgeographic.com/2023/03/transparency-matters/>
- Muralidhar, K.: A re-examination of the census bureau reconstruction and reidentification attack. In: Domingo-Ferrer, J., Laurent, M. (eds.) Proceedings of Privacy in Statistical Databases - PSD 2022 (2022)
- Muralidhar, K., Domingo-Ferrer, J.: Database reconstruction is not so easy and is different from reidentification. *J. Off. Stat.* **39**(3), 391–398 (2023a)
- Muralidhar, K., Domingo-Ferrer, J.: A rejoinder to Garfinkel (2023) – legacy statistical disclosure limitation techniques for protecting 2020 decennial US Census: still a viable option. *J. Off. Stat.* **39**(3), 411–420 (2023b)
- Ruggles, S.: Implications of Differential Privacy for Census Bureau Data and Scientific Research. Working Paper No. 2018-6, Minnesota Population Center (2018)
- Ruggles, S.: When privacy protection goes wrong: how and why the 2020 census confidentiality program failed. *J. Econ. Perspect.* **38**(2), (2024)
- Wagner, D., Layne, M.: The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software. Working Paper Number CARRA WP 2014-01 (2014). <https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-01.html>



# Relational Or Single: A Comparative Analysis of Data Synthesis Approaches for Privacy and Utility on a Use Case from Statistical Office

Manel Slokom<sup>1,2(✉)</sup>, Shruti Agrawal<sup>1</sup>, Nynke C. Krol<sup>1</sup>, and Peter-Paul de Wolf<sup>1</sup>

<sup>1</sup> Statistics Netherlands, The Hague, The Netherlands

{s.agrawal,nc.krol,pp.dewolf}@cbs.nl

<sup>2</sup> Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

m.slokom@cwi.nl

**Abstract.** This paper presents a case study focused on synthesizing relational datasets within Official Statistics for software and technology testing purposes. Specifically, the focus is on generating synthetic data for testing and validating software code. Our study conducts a comprehensive comparative analysis of various synthesis approaches tailored for a multi-table relational database featuring a one-to-one relationship versus a single table. We leverage state-of-the-art single and multi-table synthesis methods to evaluate their potential to maintain the analytical validity of the data, ensure data utility, and mitigate risks associated with disclosure. The evaluation of analytical validity includes assessing how well synthetic data replicates the structure and characteristics of real datasets. First, we compare synthesis methods based on their ability to maintain constraints and conditional dependencies found in real data. Second, we evaluate the utility of synthetic data by training linear regression models on both real and synthetic datasets. Lastly, we measure the privacy risks associated with synthetic data by conducting attribute inference attacks to measure the disclosure risk of sensitive attributes. Our experimental results indicate that the single-table data synthesis method demonstrates superior performance in terms of analytical validity, utility, and privacy preservation compared to the multi-table synthesis method. However, we find promise in the premise of multi-table data synthesis in protecting against attribute disclosure, albeit calling for future exploration to improve the utility of the data.

**Keywords:** Relational data · data synthesis · inference · constraints · Single vs Multi

---

The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

## 1 Introduction

In recent years, the need for synthetic data has gained a lot of attention, particularly in official statistics. Synthetic data offers a viable solution to privacy concerns, allowing organizations to share and utilize data without compromising sensitive information. Various use cases for synthetic data include data release, testing, education, data augmentation, and bias mitigation [7]. For example, synthetic data can be publicly released to enhance transparency, increase collaboration among parties, or for educational purposes. In this paper, we investigate the use of synthetic data for testing technologies/algorithms. We collaborate with the Social Security department at Statistics Netherlands, exploring how synthetic data can validate and test their implementations effectively.

The dataset under investigation comprises two tables linked by a one-to-one relationship. This structure presents unique challenges in generating synthetic data that maintains the integrity of both inter-table and intra-table relationships. We aim to generate synthetic data while preserving these connections, ensuring the synthesized data remains faithful to the real dataset's structure. The generated synthetic data has to adhere to the constraints provided by the software engineers of social security. Several approaches have been proposed in the literature for generating single synthetic data, such as data distortion by probability distribution [14], synthetic data by multiple imputation [19], and synthetic data by Latin Hypercube Sampling [2]. In [5], the authors proposed an empirical evaluation of different machine learning algorithms, e.g., classification and regression trees (CART), bagging, random forests, and Support Vector Machines for generating synthetic data. For multi-table data synthesis, the authors in [18] proposed the Conditional Parameter Aggregation method for synthesizing relational data, emphasizing the need to account for the influence of child tables on parent tables. In [13] the authors present Incremental Relational Generator (IRG), which uses GANs to synthetically generate interrelated tables. In our study, we compare state-of-the-art single and multi-table data synthesis approaches. We use two open public toolkits: SynthPop<sup>1</sup>, and the Synthetic Data Vault (SDV)<sup>2</sup>. As we generate synthetic data, it is crucial to evaluate its utility and assess disclosure risks. For utility measure evaluation, we compare the performance of several regression models trained on real and synthetic datasets and tested on real data. Regarding disclosure risk, we focus on measuring the potential of the different synthetic datasets to protect against attribute disclosure, ensuring that sensitive information remains protected.

Our main research question examines *how can we generate (relational vs. single) synthetic data that protects users' private data while maintaining the utility of the data for testing technologies/algorithms purposes?* In essence, we aim to answer the following research questions through this study:

- *SubRQ1:* How can we create relational synthetic data? To what extent can we generate synthetic data by combining data from sources as a single table?

---

<sup>1</sup> <https://synthpop.org.uk/>.

<sup>2</sup> <https://sdv.dev/>.

- *SubRQ2*: Which method achieves the best synthesis quality on the grounds of analytical validity, utility, and privacy risks?
- *SubRQ3*: What are the risks of disclosure from different synthesis approaches?

## 2 Background and Related Work

In this section, we provide a brief overview of existing techniques for synthetic data generation and measuring the disclosure risk.

### 2.1 Synthetic Data Generation

Synthetic data have been around for quite some time in the world of Statistical Disclosure Control (SDC). However, in recent years a lot of renewed interest in synthetic data has developed. Partly because of new computational possibilities but just as well in view of new regulations like the European General Data Protection Regulation (GDPR, [8]). Synthetic data are available in two flavors: fully synthetic and partially synthetic. In the current paper, we will focus on fully synthetic data: all attributes of all records are synthesized based on the real data [4,6].

*Single-table Synthesis.* Several approaches are available for generating synthetic single data, including multiple imputations [19], Latin Hypercube Sampling [2], machine learning approaches like classification and regression trees (CART), bagging, random forests, and Support Vector Machines [5]. The authors showed that data synthesis using CART results in synthetic data that provides reliable predictions and low disclosure risks. CART, being a non-parametric method, helps in handling mixed data types and effectively captures complex relationships between attributes [5]. Other approaches involve generative models like General Adversarial Networks, especially CTGAN, and Tabular Variational Autoencoders (TVAE) [25], and TableGAN [17].

*Multi-table Synthesis.* for relational data, in [18], the authors introduced Conditional Parameter Aggregation (CPA). CPA addresses the challenge of maintaining relationships between tables in a relational database. It operates by iterating through each record in a parent table and performing a conditional lookup to gather data from all child tables that reference it. In [13] the authors present Incremental Relational Generator (IRG), which uses GANs to synthetically generate a table-by-table synthetic relational database. In [10], the authors propose FakeDB, a general framework to generate synthetic data that preserves a wide variety of semantic integrity constraints as well as a broad set of statistical properties, across an entire relational database. In [9], the authors have conducted a comprehensive study for applying GAN to relational data synthesis.

## 2.2 Disclosure Risk Measures and Threat Model Formulation

Disclosure risk is defined as the risk that a user or an attacker can use the protected data to derive sensitive information on an individual among those in the real data [3]. Different types of disclosure are mentioned in the context of statistical disclosure control [12]: identity disclosure, attribute disclosure, and inferential disclosure. In the context of fully synthetic data, identity disclosure is often considered to be a non-threat. However, depending on the accuracy of the generating process, still a (very) small identification risk could remain. Moreover, attribute disclosure or inference disclosure is very well possible with synthetic data. Recent developments to estimate attribute disclosure in synthetic data include the so-called Correct Attribution Probability (CAP) [11, 15, 24]. CAP assumes that the attacker knows the values of a set of key attributes for an individual in the original data set, and aims to learn the respective value of a target sensitive attribute. From machine learning, in [22, 23], the authors discuss a use case of synthetic data related to releasing trained machine learning models. They investigate privacy risks associated with model inversion attribute inference attacks. In our view, it is still not clear how protective synthetic data are in terms of statistical disclosure. Indeed, in [21] it is stated that ‘*disclosure risk measures for synthetic data after its generation are still ad-hoc, and a more formal framework is needed for measuring the risk of attribute disclosure*’.

**Table 1.** The threat model that we address in our paper.

Component	Description
<i>Adversary: Objective</i>	To infer if a target individual has received assistance.
<i>Adversary: Resources</i>	The attacker has a pre-trained classifier or a subset of data to train one. The subset of data can also be the synthetic data.
<i>Vulnerability: Opportunity</i>	Possession of clean-text data and the ability to infer individual’s sensitive data
<i>Countermeasure</i>	Make access to real data unreliable

In our work, the measuring and mitigation of privacy risks of synthetic data are founded on the concept of the threat model. A threat model is a theoretical framework that defines what constitutes a privacy violation or breach, such as linking identity to a record, resulting in the leakage of sensitive information. In a widely recognized schema proposed by [20], a threat model comprises two key components: the *adversary* and the *vulnerability*. First, *the adversary’s objective* defines what the adversary seeks to accomplish, with potential goals including re-identification attacks, inference attacks, or membership inference attacks. Second, *the adversary’s resources* define what the adversary can do, encompassing different levels of knowledge and resources. We focus on black-box scenarios, where the adversary has limited knowledge of the system. Next, *the*

*vulnerability-opportunity* determines what an adversary is willing to do. Finally, *the vulnerability-countermeasure* component suggests potential solutions to protect against specific attacks. In Table 1, we provide details about the threat model that we aim to address in this paper.

### 3 The Applied Synthesis Approaches

There are different ways to create fully synthetic data. Based on the number of tables to be synthesized, we have investigated single-table and multi-table synthesis approaches. Our work explores different synthesis methods that accommodate various types of data structures. Since the relationship between the tables in the current scope is one-to-one (1-1), it is possible to merge them and synthesize them as a single table. Alternatively, the two tables can be synthesized independently using a multi-table synthesis method.

Data synthesis is a two-step process [6]. The first step consists of training a model using the real data to learn the joint distributions. The second step involves generating synthetic values for each attribute in turn, using the estimated model for the conditional distribution of that attribute, and using as input the synthetic values already produced for the previous attributes.

#### 3.1 Single Data Synthesis Approach

In this section, we describe the single table synthesis approaches we use in our experiments. We select two state-of-the-art approaches: (1) CART is a fully conditional specification (FCS) method, (2) TVAE is a generative model.

*Synthetic Data Generation Using CART.* CART takes as a parameter the matrix of predictors to model the data and sample the synthesized records. The first column to be synthesized is sampled from the distribution in the real data. The sequence of the synthesis of columns and the predictors of each column are important hyperparameters. After fitting the decision tree to a specific set of inputs, a synthesized value is generated by randomly selecting an item from the leaf node where the input parameters fall. This approach maintains reasonable analytical validity while ensuring that exact replicas of real data are not produced. However, it's crucial to tune the hyperparameters of the tree to prevent overfitting, which helps mitigate privacy risks. In our experiments, we use the Synthpop package in R that offers a sequential synthesis approach [16]. Synthpop provides sampling methods based on linear models or decision tree-based models. We use CART in our experiments as it has shown to perform the best in the literature [5].

*Synthetic Data Generation Using TVAE.* This is a neural network based on encoder-decoder architecture adapted for tabular data. This synthesizer uses the variational-autoencoder architecture to learn a model from real data and create synthetic data [25]. The encoder is a neural network that outputs the

parameters of the normal distribution of the latent space parameters. The latent space parameters from the normal distribution are further fed into the decoder. Thus, TVAE completely hides the input parameters from being passed into the synthetic data. We note that even though TVAE underperforms CTGAN, it is chosen for this study since it takes time of the order of 1/10 that by CTGAN [25]. TVAE is implemented in the SDV Python package with the default network architecture.<sup>3</sup>

### 3.2 Relational Data Synthesis Approach

In a relational database, tables are often interconnected, with one table referencing records in another. To effectively synthesize relational datasets while preserving their complex dependencies, we use the Hierarchical Modeling Algorithm (HMA).<sup>4</sup> HMA recursively models the relationships across all tables in a dataset, ensuring that the generative process respects the hierarchical and relational structure inherent in the data. This approach involves training individual models for each table, conditioned on the context provided by their related tables. By capturing how fields in different tables interrelate, HMA constructs a comprehensive representation of the entire dataset. During the data generation phase, HMA sequentially generates synthetic data for each table, maintaining the learned dependencies and ensuring consistency across the dataset. This method allows for the creation of realistic and coherent synthetic data, suitable for various downstream applications such as testing and analysis, while protecting the privacy of the real data.

## 4 Experimental Setup

In this section, we describe our data, as well as the privacy, and utility measures.

### 4.1 Data Set

Our data consists of two tables. The first table, *GBA*, contains records for 27 million unique individuals in the Netherlands. Each record corresponds to a unique individual and includes basic information such as an ID (*RINPER-SOON*), country of birth (*GBAGEBOORTELAND*), year of birth (*GBAGE-BOORTEJAAR*), gender (*GBAGESLACHT*), and the year of birth of the person’s parents (*GBAGEBOORTEJAARMOEDER* and *GBAGEBOORTEJAAR-VADER*). Note that there are missing values for some records, specifically 38% missing values for the mother’s birth year and 41% for the father’s birth year.

The second table, *Bij*, contains records for 0.47 million unique individuals who have received some form of social benefit. Each record provides information about whether the individual received one of three kinds of benefits (*bijstand*,

---

<sup>3</sup> <https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/tvaesynthesizer>.

<sup>4</sup> <https://docs.sdv.dev/sdv/multi-table-data/modeling/synthesizers/hmasynthesizer>.

*ioaw, ioaz*), an ID (*RINPERSOON*), dates when the benefits first started (*aanvangbijstand, aanvangioaw, aanvangioaz*), and the start and end dates for the benefits in the corresponding year (*Aanvbjstndpersoon, Eindbjstndpersoon*).

For the purpose of single-table synthesis, these two datasets were inner joined on the ID. Due to the complexity and time needed to run the experiments, especially for GAN, we randomly selected 50K individuals. The 50K records are used to generate both single and multi-table synthetic data. The relationships between the tables in our dataset are depicted in Fig. 4 (Appendix Sect. A). For multi-table synthesis, the tables were joined later on for the purposes of analytical, utility, and privacy assessments.

## 4.2 Measuring Utility

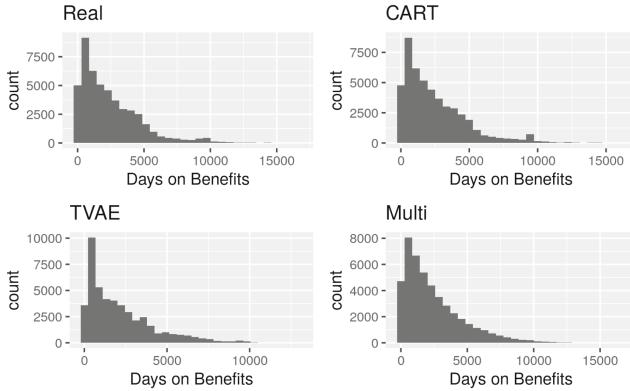
As discussed earlier, the main purpose of using synthetic data is to test technologies/algorithms. The first step in this evaluation is to validate the utility of synthetic data compared to real data. In our case study, this involves maintaining the same data structure, uni- and bi-variate distributions, and respecting the constraints (cf. Sect. 5). In this section, we examine the effectiveness of a linear regression model in predicting the duration (measured in days) individuals receive benefits. We adopt the  $\mathcal{TSTR}$  (train on synthetic and test on real) and  $\mathcal{TRTR}$  (train on real and test on real) strategies. For a fair comparison, linear regression models are trained on real and synthetic datasets, respectively, and then tested on exclusive real test data randomly sampled from the entire population, totaling approximately 2000 records.

**Outcome Attribute and the Other Attributes.** With the dataset available in official statistics, it is of interest to determine which factors influence the duration of time a person receives social benefits. This outcome attribute *Days.benefits* (number of days people are on benefits) is derived from available dates in the dataset. Figure 1 illustrates the distribution of the outcome attribute in both real and synthetic datasets. The distribution generated by the CART method shows the closest resemblance to the real data. The Multi method follows, while the distribution produced by the TVAE method appears noticeably distinct compared to the real data distribution.

To predict the number of days on benefits, we utilize attributes such as year of birth (*GBAGEBOORTEJAAR*), gender (*GBAGESLACHT*), and the presence of other benefits (*ioaw*). These attributes are used to assess model performance across all datasets. It is important to note that our selection of real data included only individuals receiving benefits at a specific date. Some attributes contained significant missing values or were not relevant for modeling our outcome.

We opt for a simple linear regression model to illustrate the disparities between real and synthetic data. This model is transparent and widely used in statistical research. We depict our formula in Eq. 1:

$$\text{Days.Benefits}_i = \alpha + \beta_1 \text{GBAGEBOORTEJAAR}_i + \beta_2 \text{GBAGESLACHT}_i + \beta_3 \text{ioaw}_i \quad (1)$$



**Fig. 1.** The distribution of the number of days people are on benefits for the real data (top left) and synthetic data generated using the CART, TVAE, and Multi-table.

We use the Root Mean Squared Error (RMSE) as a performance indicator for the regression model. Lower values indicate a better fit. More results are present in the Appendix (Sect. A.2)

### 4.3 Measuring Attribute Inference Attack

Following our threat model (cf. Sect. 2.2), we evaluate attribute inference attacks using three machine learning algorithms. In this section, we describe the subset of data available to the attacker, the inference attack models, and the metrics used to measure the success of an attack.

**Subset of Data.** In our experiments, we assume the attacker has access to a subset of data or a pre-trained model. This subset, possibly obtained through scraping or as an internal actor. The subset of data includes information on gender, birth date, parents' birth dates, and country of origin. Additionally, the attacker has a dataset of 10K target individuals for whom they aim to infer whether they received assistance (Bijstand). The binary outcome attribute “bijstand” (1 for received assistance, 0 otherwise) is notably unbalanced. The attacker uses this subset of data to train machine learning classifiers, which are then applied to the target data for inference. We also explore different sizes of attacker training data to determine the minimum number of records needed for a successful attack.

**Machine Learning Models.** *Naive Bayes (NB)* is a simple yet powerful probabilistic machine learning model based on Bayes’ theorem. Second, *Decision Tree (DT)* is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. Third, *XGBoost (GBC)* is a powerful and efficient implementation of the gradient boosting framework [1].

To evaluate the success of our attribute inference attacks, we use: The *F1-score macro-average* measures a test’s accuracy by considering both precision

and recall. It is the harmonic mean of precision and recall, with an F1 score ranging from 0 (worst) to 1 (perfect precision and recall). The macro-average approach calculates the F1-score independently for each class and then averages them, treating all classes equally. The *MCC* takes into account true and false positives and negatives, providing a balanced measure even if the classes are of very different sizes. It returns a value between  $-1$  and  $+1$ , where  $1$  indicates perfect prediction,  $0$  indicates random prediction, and  $-1$  indicates total disagreement between prediction and observation. We repeat the attribute inference attack experiments ten times and report the average and standard deviation.

## 5 Analytical Validity

Analytical validity of synthetic data is crucial in the evaluation process, ensuring its suitability for software testing. This involves maintaining identical data structures and types across all three synthetic datasets compared to the real data. Additionally, we verify that our synthetic datasets adhere to the rules and constraints provided by the social security team for our testing purposes. There are 6 conditional constraints as listed in Table 2, which specify the expected behavior of numerical date columns when the corresponding binary column is either 1 or 0. In the real data, these constraints are observed for records where the binary column is either False or True. The CART synthetic data fully satisfies these constraints. However, the TVAE and Multi synthesis models fail to learn these constraints. For instance, only 4.1% of records in the real data where *bijstand* = False meet constraint 1, whereas only 4.0% of records in the TVAE and Multi synthetic data adhere to this constraint, compared to 100% compliance in the real data and CART synthesized data. Similar discrepancies are observed for constraints 2 through 6 as well (further details are in Sect. A.1).

**Table 2.** Constraint check on the real vs. CART, TVAE, and Multi-Table Synthesized Data (%).

Constraint	Real	CART	TVAE	Multi	Priori
1. No Date for <i>Bijstand</i> when <i>Bijstand</i> = 0	100	100	4.0	4.0	4.1
2. No Date for <i>Ioaw</i> when <i>Ioaw</i> = 0	100	100	96.1	96.2	96.1
3. No Date for <i>Ioaz</i> when <i>Ioaz</i> = 0	100	100	99.5	99.5	99.5
4. Date for <i>Bijstand</i> when <i>Bijstand</i> = 1	100	100	96.1	95.7	95.9
5. Date for <i>Ioaw</i> when <i>Ioaw</i> = 1	100	100	3.3	4.1	3.8
6. Date for <i>Ioaz</i> when <i>Ioaz</i> = 1	100	100	0.0	0.8	0.4

## 6 Utility Measures

In this section, the results from our analysis on the utility of the synthetic datasets will be presented. In Table 3, the coefficients (betas) and their standard errors from the linear regression model trained on real data and synthetic data. Large differences can be observed, with the linear model trained on synthetic data. Notably, the linear model trained on synthetic data by the CART approach closely resembles the coefficients observed in the model trained on real data, indicating better alignment in predictive performance compared to the other synthetic data generation methods.

**Table 3.** Coefficients from a linear model and their respective standard errors.

	Real		CART		TVAE		Multi	
	Coef	Std Err	Coef	Std Err	Coef	Std Err	Coef	Std Err
(Intercept)	98286.67	1256.77	97910.80	1311.37	62993.60	1195.37	2908.35	1351.86
GBAGEBOORTEJAAR	-48.71	0.64	-48.45	0.67	-31.44	0.61	-0.29	0.69
GBAGESLACHT	332.89	19.66	153.49	20.32	1328.67	24.50	209.12	20.94
ioaw	-1332.74	231.36	-1.36	238.11	-596.09	103.39	3.89	54.22

In Table 4, the Root Mean Squared Error (RMSE) for models trained on real data and synthetic data are presented. Again, we see that the RMSE for the model trained on CART data is most comparable to the RMSE for the model trained on real data, suggesting that the CART synthetic data approach provides predictions closest to those derived from real data. Next, the Multi approach performs second best, followed by TVAE.

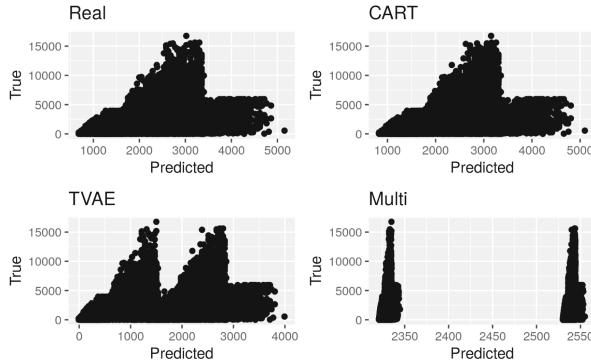
**Table 4.** RMSE for models trained on real data and synthetic datasets.

	Real	CART	TVAE	Multi
RMSE	2129.42	2133.43	2291.83	2249.53

In Fig. 2, we show the predicted values of the number of days a person is on benefits against the true number of days they are on benefits. In a perfect prediction, all values would fall on a diagonal line. The predictive value of our model trained on real data can be improved, but the CART model resembles its results. On the other hand, the predicted values from the model trained on synthetic data generated by the TVAE method show a different pattern. For the Multi-method, the pattern in predicted values versus true values diverges even further from the pattern when real data is used.

## 7 Attribute Disclosure Risk

In this section, we provide our results of the attribute inference attack. In Table 5, we provide our results on attribute inference attack. The results show that mod-



**Fig. 2.** Utility measure: True values of the outcome attribute in our test set, the number of days people are on benefits, and the predicted values from linear models trained on real data (top left), and on synthetic data generated by CART, TVAE, and Multi.

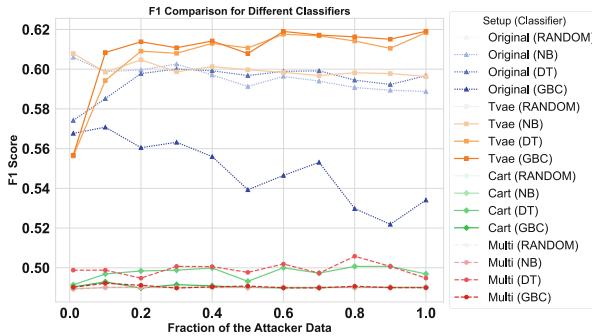
els trained on TVAE data achieved the highest scores across all metrics, outperforming the random classifier. This indicates a higher risk of sensitive information leakage when using TVAE-synthesized data. In contrast, models trained on CART and Multi data sets have lower scores than the random classifier, suggesting that these two approaches are more effective at protecting sensitive information, thereby providing better privacy. The performance of models trained on CART and Multi data sets demonstrates their potential for reducing data leakage and improving data privacy.

Figure 3 provides a comparison of the F1-scores macro-average for different machine learning classifiers (Random, NB, DT, and GBC) across different fractions of attacker data. We show the performance on different attacker data sizes. These figures compare the performance of attribute inference attack models trained on synthetic data (Multi, TVAE, CART) to that of a model trained on real data.

We observe that across all conditions, models trained on TVAE synthetic data have the highest F1 scores surpassing those of models trained on real data. This confirms that TVAE does not help to protect against attribute disclosure. However, looking at the performance of the models trained on CART and Multi, we see that the F1 scores are around 0.5 and below. This demonstrates that Multi and CART offer higher protection against attribute disclosure compared to that of TVAE.

**Table 5.** Results of the attribute inference attack are measured in terms of F1 macro and MCC. We compare the performance of a random classifier to DT, NB, and XGBoost. We compare the performance of models trained on different training data: Real, TVAE, CART, and Multi. Gray-highlighted scores indicate classifier performance lower than real data.  $\pm$  denotes the standard deviation over ten runs of the experiments. Note that the test set is the same real target individuals.

Classifiers		Random		DT		NB		XGBoost	
Data Sets		F1	MCC	F1	MCC	F1	MCC	F1	MCC
<i>Real</i>	0.4901	0.00	0.5967	0.1943	0.5888	0.2821	0.5341	0.1252	
	$\pm$ 0.000	$\pm$ 0.000	$\pm$ 0.0087	$\pm$ 0.0174	$\pm$ 0.0024	$\pm$ 0.0041	$\pm$ 0.0185	$\pm$ 0.0402	
<i>TVAE</i>	0.4901	0.00	<b>0.6168</b>	<b>0.2440</b>	<b>0.5963</b>	<b>0.2614</b>	<b>0.6169</b>	<b>0.2384</b>	
	$\pm$ 0.000	$\pm$ 0.000	$\pm$ 0.0009	$\pm$ 0.0026	$\pm$ 0.0018	$\pm$ 0.0015	$\pm$ 0.0089	$\pm$ 0.0174	
<i>CART</i>	0.4901	0.00	0.4969	-0.0060	0.4901	0.00	0.4901	-0.0012	
	$\pm$ 0.000	$\pm$ 0.000	$\pm$ 0.0015	$\pm$ 0.0030	$\pm$ 0.0000	$\pm$ 0.0000	$\pm$ 0.0000	$\pm$ 0.0011	
<i>Multi</i>	0.4901	0.00	0.4948	-0.0097	0.4901	0.00	0.4900	-0.0045	
	$\pm$ 0.000	$\pm$ 0.000	$\pm$ 0.0029	$\pm$ 0.0058	$\pm$ 0.0000	$\pm$ 0.0000	$\pm$ 0.0000	$\pm$ 0.0005	



**Fig. 3.** Attribute inference attack measured using F1 Scores Macro on DT, NB, XGBoost (GBC) for the different synthetic data Multi, TVAE, CART.

## 8 Conclusion and Future Work

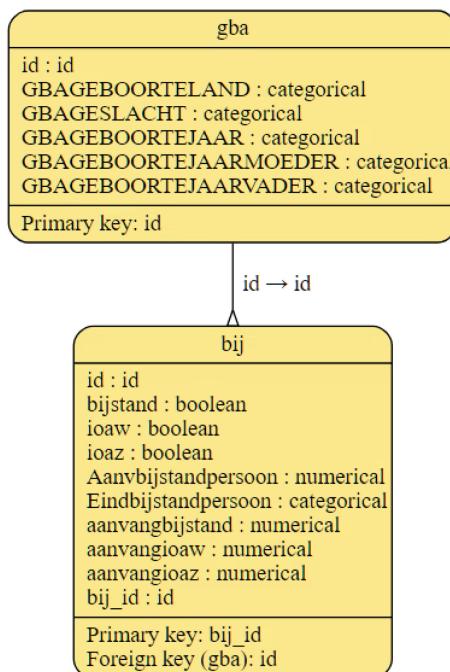
In this paper, we conducted a comparative analysis of different synthesis approaches, focusing on the specific challenge posed by a relational multi-table structure with a one-to-one relationship for testing technologies/algorithms. By juxtaposing single table synthesis against multi table synthesis, we aimed to discern the strengths and limitations of each method. Our approach involved merging the two tables into a single entity for single-table synthesis, facilitating a direct comparison with the multi-table synthesis technique.

Through extensive experimentation, we evaluated the efficacy of various synthesis methods in respecting constraints/rules, ensuring analytical validity, maintaining the utility of the data, and mitigating risks associated with attribute disclosure. Our findings revealed that among the single synthesis approaches, CART emerged as the most effective solution for generating synthetic data within our

particular use case. CART demonstrated superior performance in preserving the integrity of the synthesized data while meeting the constraints imposed by the analytical framework. On the other hand, the multi-table synthesis method demonstrated promise in capturing the intricate inter- and intra-relationships inherent in the data structure. While it proved effective in protecting against attribute disclosure, comparable to the performance of CART, its utility effectiveness faced limitations. This suggests that while the multi-table approach holds potential, further refinement and optimization are necessary to fully exploit the relational structure embedded within the data. Future research should focus on improving the utility of the multi-table synthesis method to ensure its practical applicability across diverse analytical frameworks and use cases.

**Acknowledgments.** We would like to thank the SOZ team at Statistics Netherlands for providing us with the data and knowledge of the rules. This work was partly supported by the AI, Media, and Democracy Lab, NWA.1332.20.009.

## A Appendix



**Fig. 4.** The metadata of our relational data. We have two tables *gba* and *bij* that are connected through  $id \rightarrow id$ . The primary id of table *gba* is a foreign key in table *bij*.

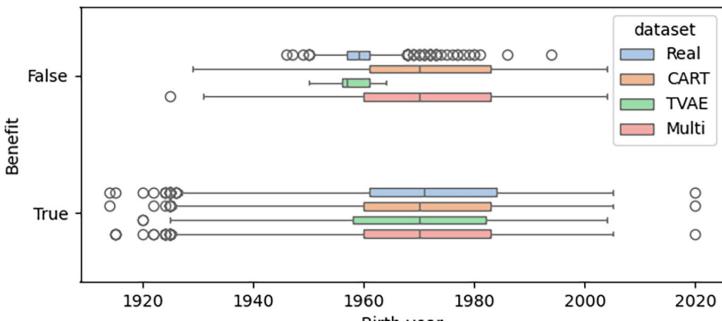
### A.1 Analytical Validity

Table 6 shows the counts or percentages of categories captured in the categorical or boolean columns synthetic datasets compared to the real and.

**Table 6.** Measurement of representation of underrepresented categories captured in the real, the CART, TVAE and multi-table synthetic datasets.

Column	Real	CART	TVAE	Multi
Country (unique codes)	218	205	61	200
bijstand = 0	4.1%	4.0%	3.4%	4.1 %
ioaw = 1	3.8%	3.8%	3.9%	4.0%
ioaz = 1	0.4%	0.4%	0.006%	0.5%

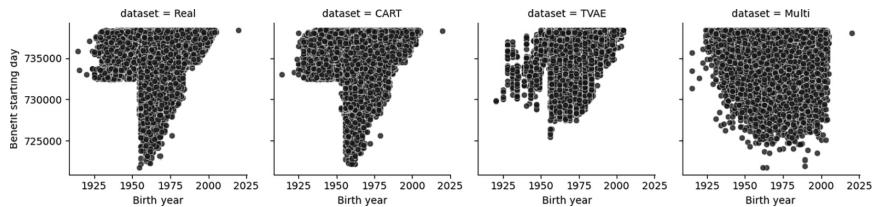
*Bivariate Distribution.* The distribution between column pairs in the real data is compared with those in the synthetic datasets. CART based synthesis outperforms TVAE and Multi model for all but the case of benefit (True or False) vs birth year, as shown in Fig. 5 and Fig. 6.



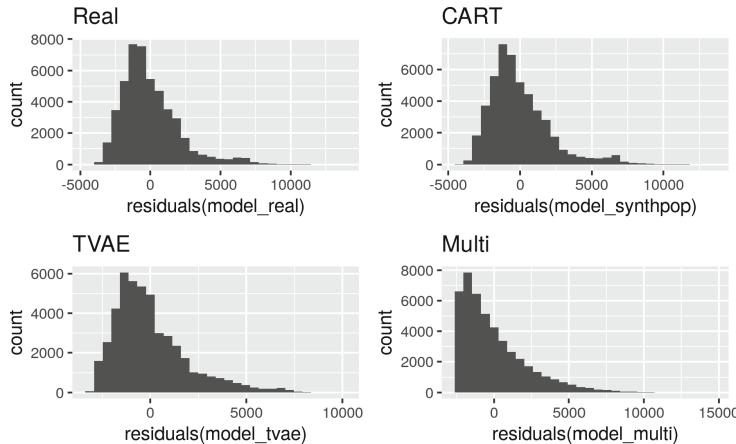
**Fig. 5.** Distribution of birth year in different datasets who received or did not receive the benefit. People from only a specific range of birth years did not receive the benefit in the real data. This characteristic has been very well captured by the TVAE synthetic data.

### A.2 Utility

**Residuals.** To check the assumptions of our models, we look at the residuals (errors) of the models trained on real and synthetic data. Residuals are visualized in Fig. 7. Although some skewness is present in all plots, the model trained on synthetic data generated by the Multi-method does seem to violate the assumptions of a linear model most.



**Fig. 6.** Starting date of benefit vs birth year. This characteristic has been very well captured in the CART synthetic data.



**Fig. 7.** Residuals of the linear models trained on real data and synthetic data generated using CART, TVAE, and Multi methods.

## References

1. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
2. Dandekar, R.A., Cohen, M., Kirkendall, N.: Sensitive micro data protection using Latin hypercube sampling technique. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 117–125. Springer, Heidelberg (2002). [https://doi.org/10.1007/3-540-47804-3\\_9](https://doi.org/10.1007/3-540-47804-3_9)
3. Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical data protection. J. Comput. Appl. Math. **164–165**, 285–293 (2004). proceedings of the 10th International Congress on Computational and Applied Mathematics
4. Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. Trans. Data Priv. **1**(3), 105–130 (2008)
5. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. Comput. Stat. Data Anal. **55**(12), 3232–3243 (2011)

6. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, vol. 201. Springer, New York (2011)
7. Nations Economic Commission for Europe, U., et al.: Synthetic data for official statistics: a starter guide (2023)
8. European Parliament and Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). OJ 2016 L 119, pp. 1–88 (2016)
9. Fan, J., Chen, J., Liu, T., Shen, Y., Li, G., Du, X.: Relational data synthesis using generative adversarial networks: a design space exploration. Proc. VLDB Endow. **13**(12), 1962–1975 (2020)
10. Gao, C., Jajodia, S., Pugliese, A., Subrahmanian, V.: FakeDB: generating fake synthetic databases. IEEE Trans. Dependable Secure Comput. 1–12 (2024)
11. Hittmeir, M., Mayer, R., Ekelhart, A.: A baseline for attribute disclosure risk in synthetic data. In: Proceedings of the 10th ACM Conference on Data and Application Security and Privacy, pp. 133–143 (2020)
12. Hundepool, A., et al.: Statistical Disclosure Control. Wiley, Hoboken (2012)
13. Li, J., Tay, Y.: IRG: generating synthetic relational databases using GANs. arXiv preprint [arXiv:2312.15187](https://arxiv.org/abs/2312.15187) (2023)
14. Liew, C.K., Choi, U.J., Liew, C.J.: A data distortion by probability distribution. ACM Trans. Database Syst. **10**(3), 395–411 (1985)
15. Elliot, M.: Final report on the disclosure risk associated with synthetic data produced by the SYLLS team (2014). <http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/>. Accessed 13 Oct 2023
16. Nowok, B., Raab, G.M., Dibben, C.: synthpop: bespoke creation of synthetic data in R. J. Stat. Softw. **74**(11), 1–26 (2016)
17. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on Generative Adversarial Networks. In: Proceedings of the 44th International Conference on Very Large Data Bases (VLDB Endowment), vol. 11, no. 10, pp. 1071–1083 (2018)
18. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: IEEE International Conference on Data Science and Advanced Analytics, pp. 399–410 (2016)
19. Rubin, D.B.: Discussion statistical disclosure limitation. J. Official Stat. **9**(2), 461–468 (1993)
20. Salter, C., Saydjari, O.S., Schneier, B., Wallner, J.: Toward a secure system engineering methodology. In: Proceedings of the Workshop on New Security Paradigms, pp. 2–10. NSPW (1998)
21. Schlomo, N.: How to measure disclosure risk in microdata? Surv. Stat. **86**, 13–21 (2022)
22. Slokom, M., de Wolf, P.P., Larson, M.: When machine learning models leak: an exploration of synthetic training data. In: Domingo-Ferrer, J., Laurent, M. (eds.) Proceedings of the International Conference on Privacy in Statistical Databases: Corrected and updated version on arXiv at:<https://arxiv.org/abs/2310.08775> (2022)
23. Slokom, M., de Wolf, P.P., Larson, M.: Exploring privacy-preserving techniques on synthetic data as a defense against model inversion attacks. In: Athanasopoulos, E., Mennink, B. (eds.) ISC 2023. LNCS, vol. 14411, pp. 3–23. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-49187-0\\_1](https://doi.org/10.1007/978-3-031-49187-0_1)

24. Taub, J., Elliot, M., Pampaka, M., Smith, D.: Differential correct attribution probability for synthetic data: an exploration. In: Domingo-Ferrer, J., Montes, F. (eds.) PSD 2018. LNCS, vol. 11126, pp. 122–137. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-99771-1\\_9](https://doi.org/10.1007/978-3-319-99771-1_9)
25. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alche Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 7335–7345 (2019)



# A Case Study Exploring Data Synthesis Strategies on Tabular vs. Aggregated Data Sources for Official Statistics

Mohamed Aghaddar<sup>1(✉)</sup>, Liu Nuo Su<sup>1(✉)</sup>, Manel Slokom<sup>1,2(✉)</sup>,  
Lucas Barnhoorn<sup>1(✉)</sup>, and Peter-Paul de Wolf<sup>1(✉)</sup>

<sup>1</sup> Statistics Netherlands, The Hague, The Netherlands

m.aghaddar@cbs.nl, ln.su@cbs.nl, lucasbarnhoorn@gmail.com, pp.dewolf@cbs.nl

<sup>2</sup> Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

m.slokom@cwi.nl

**Abstract.** In this paper, we investigate different approaches for generating synthetic microdata from open-source aggregated data. Specifically, we focus on macro-to-micro data synthesis. We explore the potential of the Gaussian copulas framework to estimate joint distributions from aggregated data. Our generated synthetic data is intended for educational and software testing use cases. We propose three scenarios to achieve realistic and high-quality synthetic microdata: (1) zero knowledge, (2) internal knowledge, and (3) external knowledge. The three scenarios involve different knowledge of the underlying properties of the real microdata, i.e., standard deviation, and covariate. Our evaluation includes matching tests to evaluate the privacy of the synthetic datasets. Our results indicate that macro-to-micro synthesis achieves better privacy preservation compared to other methods, demonstrating both the potential and challenges of synthetic data generation in maintaining data privacy while providing useful data for analysis.

**Keywords:** Synthetic data · Aggregated data · Reconstruction attack · Macro-to-Micro · Micro-to-Micro · post-processing

## 1 Introduction

The generation of synthetic microdata from aggregated data sources is an emerging area of research that has yet to be explored. In this paper, we investigate a proof-of-concept approach evaluating the possibility of generating synthetic microdata from aggregated data sources. Our case study explores the use of synthetic data within Official Statistics for education and testing technologies.

Existing approaches in the literature include using Synthetic Reconstruction (SR) [15] and Combinatorial Optimization (CO) [20]. In [11], the authors

---

The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

proposed SYNC which estimates joint distributions from aggregated data using Gaussian copulas. The authors in [1] proposed the GenSyn framework, which extends the capabilities of synthetic data generation. It uses a combination of univariate and multivariate frequency tables from a target geographical location and auxiliary locations to generate synthetic microdata.

Our work builds upon the SYNC work [11]. The process of synthetic data generation via Gaussian Copulas (SYNC) offers a simple yet effective method for creating microdata that retains the dependencies between variables. We extend SYNC by employing different levels of knowledge about the real microdata. We propose three distinct scenarios for this purpose: **scenario 1: no Knowledge** - Synthetic data generation is performed without any prior information about the real data. **Scenario 2: internal knowledge** - This scenario leverages internal data, using the number of records and the covariance matrix derived from the real dataset to guide the generation process. **Scenario 3: external knowledge** - External resources, specifically ChatGPT 4.0, are used to estimate both the number of iterations and the covariance matrix. This scenario evaluates the accuracy of external insights in approximating the real microdata.

Our experiments compare the performance of macro-to-micro synthetic data to those of micro-to-micro synthetic data. Our results indicate that while micro-to-micro synthetic data generally offers better predictive performance, macro-to-micro approaches provide a viable alternative, particularly for testing and educational purposes. We also assess the privacy risks associated with these synthetic datasets, using matching techniques to evaluate the potential for identity disclosure.

## 2 Background and Related Work

In this section, we provide a brief overview of existing work on synthetic data generation. Then, we describe disclosure risk.

### 2.1 Synthetic Data Generation

Synthetic data generation involves developing a model that represents the real data, from which synthetic data are generated. These methods are designed to preserve specific statistical properties and relationships among variables that are present in the real data [5,8,19]. This section provides an overview of related work that has shaped current practices in synthetic data generation.

*Micro to Micro Synthesis.* Several techniques have been proposed in the literature for generating synthetic data, such as data distortion via probability distribution [12], synthetic data by multiple imputation [17], and synthetic data by Latin Hypercube Sampling [4]. An empirical assessment of different machine learning algorithms for synthetic data generation highlighted the effectiveness of non-parametric methods like Classification and Regression Trees (CART) found

in [7]. The authors showed that CART not only manages mixed data types effectively but also preserves complex inter-variable relationships [7].

Recently, more sophisticated approaches have been proposed in the literature to generate synthetic data [2, 16, 21]. In [21], the authors propose the state-of-the-art generative model called *CTGAN*. CTGAN addresses challenges that were not taken into account in previous GAN models such as mixed data types, non-Gaussian distributions, and sparse data scenarios. CTGAN’s conditional generator adjusts to various data variables dynamically, making it highly effective for diverse datasets. These advancements in GAN technology not only enhance the quality and utility of synthetic datasets but also open new avenues for research in privacy-preserving data synthesis.

*Macro to Micro Synthesis.* Data synthesis is a promising method for creating detailed microdata while preserving privacy. It is still in its infantile status and has not yet gained widespread recognition. Despite its potential applications in various domains, only a limited number of researchers are currently exploring this area. For instance, traditional methods like Synthetic Reconstruction (SR) [15] and Combinatorial Optimization (CO) [20] have been foundational in this field. SR methods, such as Iterative Proportional Fitting (IPF) [3], use sample data to estimate joint distributions and generate synthetic populations by fitting to marginal constraints. CO methods, including techniques like Simulated Annealing [10] and Genetic Algorithms [18], iteratively select the best combination of individuals to minimize a fitness function. These methods, however, often require initial sample data, which may not always be available. More recently, the authors in [1] proposed the GenSyn framework, which extends the capabilities of synthetic data generation. It uses a combination of univariate and multivariate frequency tables from a target geographical location and auxiliary locations to generate synthetic microdata. In [11], the authors proposed the SYNC algorithm. SYNC uses a Gaussian copula to estimate the joint distribution of variables based on the covariance matrix derived from aggregated data. Our work builds upon the SYNC algorithm.

## 2.2 Disclosure Risk

The concept of disclosure risk pertains to the likelihood that an attacker can exploit the protected data to extract sensitive information about individual records included in the real dataset [6]. This risk is a critical concern and is often classified into several types [9]: *Identity disclosure* occurs when an attacker identifies a specific individual from a released record. *Attribute disclosure* involves the revelation of sensitive information about an individual, potentially leading to privacy breaches without necessarily identifying the individual directly. *Inferential disclosure* occurs when an attacker can deduce values of certain characteristics of an individual more precisely than if the data had not been released, enhancing their knowledge indirectly through analysis of the available data.

Our work is closely related to reconstruction attacks. Reconstruction attacks aim to recover real datasets from aggregate information or model outputs, pos-

ing severe privacy risks. In [13], the authors propose a detailed analysis of the Census Bureau’s reconstruction attack. The authors highlight how reconstructed microdata can be re-identified using external data sources. In [14], the authors showed that by using traditional statistical disclosure control (SDC) techniques, reconstruction can be averted.

### 3 Macro-to-Micro Synthetic Data Generation

We reproduce and follow up on the work in [11]. Synthetic data generation via Gaussian Copulas (SYNC for short) provides a simple yet effective method for generating synthetic microdata while preserving the dependencies between variables. The process begins by fitting the marginal distributions for each variable in the dataset. This involves identifying and modeling the best-fitting distribution (e.g., normal, exponential) for each variable. Once the marginal distributions are determined, the data is transformed into a uniform distribution using the cumulative distribution function (CDF) of each marginal distribution. Next, a Gaussian copula is fitted to the transformed data. A copula is a function that captures the dependency structure between variables independently of their marginal distributions. The Gaussian copula, in particular, uses the multivariate normal distribution to model these dependencies. After fitting the copula, synthetic data can be generated by sampling from the copula and then transforming the samples back to the real data space using the inverse of the marginal CDFs. This two-step transformation ensures that the synthetic data maintains the real dependencies and distributions.

Due to the skewed nature of the data, we explore scaling and compare it to non-scaled data. Scaling ensures all variables contribute equally and can improve synthetic data generation. For scaled data, we normalize the variable ranges to a specified interval, which is particularly useful for variables with different units and magnitudes. In contrast, non-scaled data retains original units and ranges, potentially introducing bias. Then, we extend the SYNC algorithm by employing different levels of knowledge about the real microdata. We propose three distinct scenarios for this purpose:

**Scenario 1: No Knowledge.** In this scenario, both the number of iterations and the covariance matrix are randomly sampled, with no knowledge of the real data. We generate synthetic microdata without any prior information about the underlying data structure. This approach yields a similar outcome to that of the SYNC algorithm.

**Scenario 2: Internal Knowledge.** In this scenario, we assume access to the real microdata, including the number of companies (records), variance, and covariance. The number of iterations is set to the number of records in the real dataset, and the covariance matrix is derived from the real microdata. This approach leverages internal data for higher fidelity in the synthetic data. While not efficient for very large datasets, it is suitable when high accuracy is prioritized over computational efficiency and for comparison.

**Scenario 3: External Knowledge.** In this scenario, external resources, specifically ChatGPT 4.0, are used to estimate both the number of iterations and the covariance matrix. The aim is to determine if an attacker could utilize external resources to reconstruct the real microdata and assess how closely an attacker can approximate the real microdata. To do so, we provide ChatGPT with the open-source aggregated dataset to make assumptions about the variance, covariance, and number of companies. We do not provide any microdata or metadata to ChatGPT. This simulates a scenario where an attacker, unable to access the real microdata, uses ChatGPT along with open-source aggregated data to attempt to reconstruct the real microdata. The use of ChatGPT raises broader questions about the potential of Large Language Models being trained on open data available on the internet.

To implement this process, the aggregated data, corresponding metadata, and table information were downloaded from Statline<sup>1</sup> and uploaded to ChatGPT. Initially, a summary of the data was requested from ChatGPT. Following this, ChatGPT was asked to make a selection of the variables to be used in the synthetic data generation, this is the same selection as described in Sect. 4.1. Extensive questions were then posed regarding these selected variables to gather detailed insights. Finally, ChatGPT was asked to provide estimations of the means, variances, and covariances of the variables. These estimates were then used in the SYNC algorithm to generate the synthetic microdata.

We compare the Macro-to-Micro synthetic datasets against different Micro-to-Micro synthetic datasets using open-source toolkits. For the CART-based implementation, we use Synthpop<sup>2</sup>, and for the CTGAN-based implementation, we use the SDV package<sup>3</sup>. We use the default synthpop parameters for the CART algorithm with no smoothing in the leaf nodes, while the CTGAN synthesizer is optimized using two hidden layers with 256 nodes each with 2000 epochs using batch sizes of 1000.

## 4 Experimental Setup

In this section, we provide the experimental setup necessary for our experiments. We describe the open-source aggregated data and the internal microdata. We then discuss the rules for analytical validity, the regression analysis for utility, and the measures used for disclosure risk.

### 4.1 Macro and Micro Datasets

In this study, two distinct data sources are used:

(1) *Open-source aggregated data.* The primary source is the publicly open-source aggregated data, available on the online database of Statistics Netherlands, Statline. We specifically focused on “Trade and Industry; Employment and Finance

---

<sup>1</sup> <https://opendata.cbs.nl/statline/#/CBS/nl/>.

<sup>2</sup> <https://synthpop.org.uk/>.

<sup>3</sup> <https://sdv.dev/>.

per Sector, NACE<sup>4</sup> 2008.” This dataset comprises aggregated employment, financial, and trade information for each sector based on the Statistics Netherlands’ Standard Industrial Classification of all Economic Activities 2008. The data is derived from one of the largest company surveys conducted by the Production Statistics (PS), which samples a variety of companies across the Netherlands. The PS survey provides information related to industries’ employment and financial status. For this research, a subset of the complete dataset is used. The subset is extracted using specific filters focusing on the year 2021 and the following NACE codes: NACE 58 for Publishing, NACE 62 for IT Support Activities, and NACE 63 for Information Service Activities.

(2) *The internal microdata.* The internal microdata includes detailed records about companies collected from a questionnaire. The questionnaire targets three sectors: Publishing, Support Activities in the Field of IT, and Information Service Activities. The microdata contains 1781 records and 108 variables. Among these, one variable is designated as the company identifier. The remaining 107 variables are mixed types, with one being categorical and the others numerical.

For simplicity and a fair comparison between macro-to-micro and micro-to-micro data synthesis approaches, we select eight variables to continue with revenue, other operating income, cost of goods sold, personnel and other operating expenses, depreciation, net financial income/expenses, and provisions allocations. The resulting data contains a total of 1781 records. We note that these eight variables are a summary of the questionnaire.

## 4.2 Rules for Analytical Validity

The dataset used in this work contains specific relationships between the variables in the data. For the synthesizer to be effective, the synthesized dataset must also contain the relationships in the real dataset. We identify five categories of relationship rules in the dataset, explained further in Appendix A.1.

We implement the *summation* and *equality* rules by post-processing the synthesized data. That is, if we have a summation rule, we only synthesize the elements of the sum and not the sum itself. We manually add the sum back into our synthesized dataset. For the equality rule, the same principle is applied by deleting one of the variables and adding it back after. For the remaining rules, we found that the process of synthesizing is unfit for further post-processing. We find many inter-dependencies exist between the variables within the rules, and adhering to the other rule groups would mean violating a number of the summation and equality rules. Furthermore, while the post-processing will achieve a perfect, zero violation of the rules, the real dataset does contain imperfections. We argue that, while the rules do exist in theory, human error is also prominent while filling out the survey. Hence the non-zero violation percentages in the real dataset. From a practical point of view, it is useful to evaluate the synthesized datasets using the relationship rules. We post-process the rules such that the

---

<sup>4</sup> The Nomenclature of Economic Activities (for short NACE).

rules with an equality constriction are satisfied, while the others are left to the synthesizer itself since they have a greater feasible solution space.

### 4.3 Regression Models

With the current literature, many approaches have been created to capture non-linear relationships between variables. This adds another challenge to the privacy of the variables in our dataset. In particular, training prediction models using the synthesized datasets can give valuable information on protected variables.

We research if the synthetic data can be used to decipher the relationship between variables and the revenue of a company. We select the variables that are mentioned in Sect. 4.1 and aim to predict the net revenue variable. We train the model using the CART and CTGAN full synthetic datasets, and test using a 30% subset of the real dataset. We employ a simple linear regression model defined as

$$y = \beta_0 + \beta_1 x + \epsilon, \quad (1)$$

with the parameters optimized by minimizing the squared error  $\epsilon^2$ . The parameter  $\beta_0$  represents the constant, while  $\beta_1$  is a vector of parameters for every variable. The obtained forecast values from the test set are evaluated using the MSE defined as  $\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2$ . The regression is implemented for the both micro-to-micro and macro-to-micro approach.

### 4.4 Measuring the Disclosure Risk

We evaluate the success of our macro-to-micro and micro-to-micro synthetic data using matching metrics. We aim to see if our synthetic data contains duplicated or sufficiently close real records. This is particularly important for macro-to-micro data synthesis because it will give us an idea of how closely we approximate the real microdata.

We use NewRowSynthesis as our matching measure.<sup>5</sup> The NewRowSynthesis metric evaluates the novelty of each record in a synthetic dataset by determining whether it exactly replicates a record from the real dataset. A perfect score of 1.0 (100%) indicates that all synthetic records are novel, with no matches to the real data, whereas a score of 0.0 (0%) signifies that all synthetic records are duplicates of the real data. For categorical and boolean data, the metric requires the exact matches, while for numerical and date time types, it uses a scaled value comparison. In our experiments, we set the tolerance threshold between [0.0, 1.0].

## 5 Analytical Validity

In this section, we evaluate the analytical validity of our micro-to-micro synthetic datasets before and after the application of post-processing. We look at the

---

<sup>5</sup> <https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/newrowsynthesis>.

extent to which the synthetic data respects or violates our predefined rules. Table 1 shows the violation percentage of the real dataset and the violation percentages of the synthesized datasets before and after post-processing. We observe that the real dataset does not adhere perfectly to the rules, as human error exists in the data. Furthermore, the summation and equality constraints have high violation percentages when post-processing is not applied, and zero violation when it is applied. We see low violation percentages for the non-negative rules and similar violation percentages for the inequality and if-rules between with and without post-processing. The CART and CTGAN synthesizers seem to have a hard time adjusting to the inequality and if-rules. In particular, for the CTGAN, the synthesizer violates about a third of all rules in those categories. We argue that the synthesizing process of the CART is more effective in capturing the patterns of the rules.

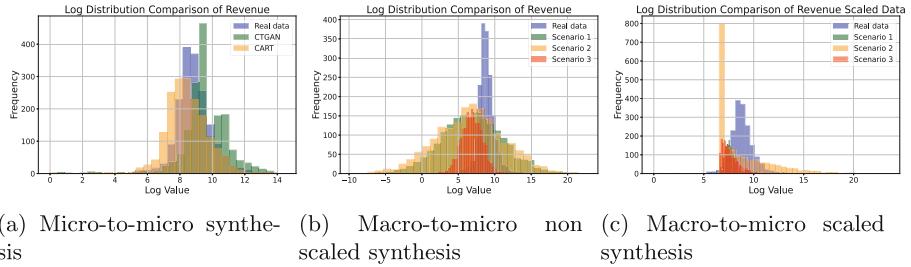
**Table 1.** Results of our validation of the violated rules before and after post-processing (%)

Rules	<i>real</i>	Before		After	
		CART	CTGAN	CART	CTGAN
Summation	0.018	<b>70.58</b>	<b>98.23</b>	<b>0.0</b>	<b>0.0</b>
Non negative	0.0017	0.001	0.82	0.21	0.97
Equality	0.0	<b>63.12</b>	<b>95.63</b>	<b>0.0</b>	<b>0.0</b>
Inequality	0.01	<b>4.80</b>	<b>25.37</b>	6.19	<b>35.7</b>
If... Then	0.46	6.12	<b>45.65</b>	8.77	<b>37.93</b>

Figure 1 shows the log distribution of the revenue variable in the different synthesized datasets and the different scenarios of the attacker. For Micro-to-Micro, the revenue distribution of the CART (Fig. 1(a)) is most similar to the real distribution in the data. CTGAN shows a slightly different distribution compared to the real data. For Macro-to-Micro, in Fig. 1(b), we observe that different scenarios result in different distributions of the variable. In particular, the scenario that contains a lot of information for the variable will result in a narrower spread for the revenue. For instance, Scenario 3, based on external knowledge from ChatGPT, seems to yield the smallest spread but is not centered on the real data. We note that in Fig. 1(c), the synthesized data is not scaled back, which is why the distributions look different from the real data distributions. More results related to the uni- and bi-variate distribution and KS-complement are provided in Sect. A.

## 6 Utility Measures

In this section, we investigate the potential uses of the synthetic datasets. We explore the use case where prediction models are trained using the synthetic data, and tested on real observations. Then, we report the MSE of the models.



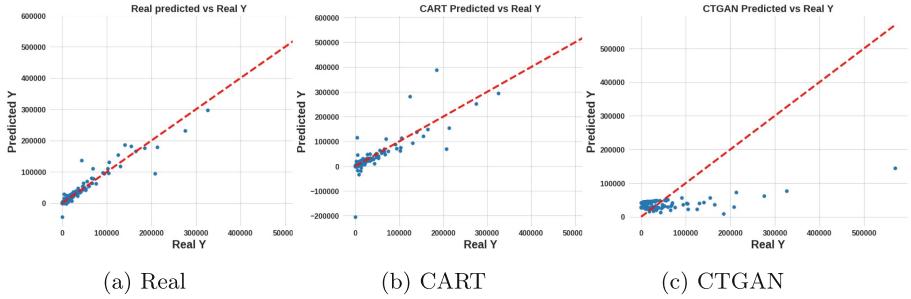
**Fig. 1.** Univariate log distributions of the variable Revenue

Although individual coefficients can be investigated, we note that the emphasis of the work is not on the effect of the variable on the revenue. More results are available in our Appendix Sect. A.3. Figures 2, 3, and Table 2 show the result of a linear regression model trained on real and synthetic datasets, respectively, and tested on real data.

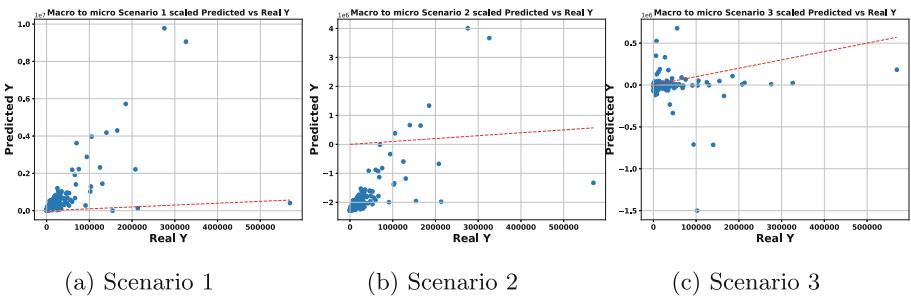
**Table 2.** Prediction error measured using MSE for Linear regression model trained on real data vs. Micro-to-Micro synthetic data vs. Macro-to-Micro synthetic data (Scenario 1, Scenario 2, Scenario 3).

Data		Synthesis	MSE
Real	None	<i>None</i>	$7.220 \times 10^7$
Micro-to-Micro	None	<i>CART</i>	$3.202 \times 10^8$
	None	<i>CTGAN</i>	$1.737 \times 10^9$
Macro-to-Micro	<b>Non Scaled</b>	<i>Scenario 1</i>	$2.606 \times 10^{15}$
		<i>Scenario 2</i>	$3.068 \times 10^{10}$
		<i>Scenario 3</i>	$3.576 \times 10^9$
	<b>Scaled</b>	<i>Scenario 1</i>	$6.04 \times 10^{11}$
		<i>Scenario 2</i>	$4.682 \times 10^{12}$
		<i>Scenario 3</i>	$1.09 \times 10^{10}$

*Micro-to-Micro.* The predictions from the model trained on the CART synthetic data seem to fit the observations more than the model trained on the CTGAN synthetic data. The predicted revenue from the CTGAN model fits a much smaller range, with a downward bias. Comparing this result to Fig. 2, predictions of the CART synthetic dataset resemble the result using the real data. In Table 2, we observe that a regression model using the real training data gets an MSE score of  $7.220 \times 10^7$ . When using micro-to-micro synthetic data generated by CART and CTGAN methods, the MSE scores increase substantially to  $3.202 \times 10^8$  and  $1.737 \times 10^9$  respectively, indicating lower predictive performance compared to the real data. This finding is in line with the results of the figures, with the CART having the lowest MSE value.



**Fig. 2.** *Micro-to-Micro*: Predicted revenue plotted against real revenue.



**Fig. 3.** *Macro-to-Micro*: Predicted revenue plotted against real revenue.

*Macro-to-Micro*. In Fig. 3, we show the performance of the predictions from the models trained on Scenario 1, Scenario 2, and Scenario 3 using Macro-to-Micro synthetic data.

First, in Scenario 1 (Fig. 3a), where no knowledge of the real data was used, the predicted values show a poor fit with the actual values, indicating discrepancies and a lack of accuracy in the synthetic data. This result highlights the challenges in generating high-quality synthetic data without any prior knowledge of the real data. Second, Scenario 2 (Fig. 3b), which leverages internal knowledge, shows an improvement in the fit between the predicted and actual values compared to Scenario 1. The predicted values follow the general trend of the actual values better, though there are still notable deviations. This suggests that having internal knowledge enhances the quality of the synthetic data but may not fully capture the underlying relationships present in the real data. Third, in Scenario 3 (Fig. 3c), where external knowledge was utilized, the predicted values exhibit a closer fit to the actual values compared to the previous two scenarios. The predictions show less variability and align more closely with the actual revenue. This indicates that integrating external insights, such as those obtained from ChatGPT, can substantially improve the accuracy of the synthetic data generation process.

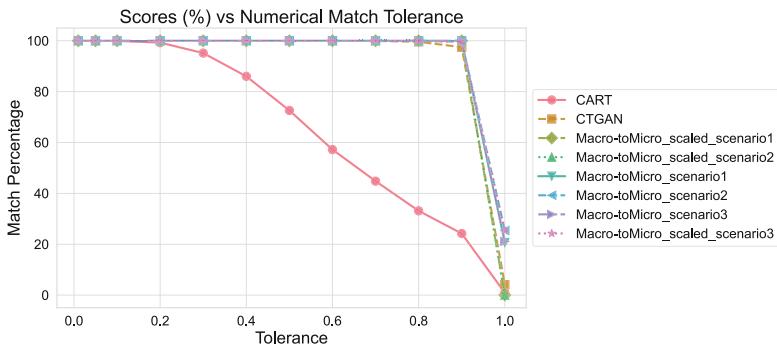
We also observe that the MSE scores vary widely for the three different scenarios (Scenario 1, Scenario 2, Scenario 3). For Scenario 1, the MSE is  $2.606 \times 10^{15}$ , for Scenario 2 it is  $3.068 \times 10^{10}$ , and for Scenario 3 it is  $3.576 \times 10^9$ . These MSE scores suggest a significant loss in predictive accuracy compared to the MSE from the real data. Additionally, comparing macro-to-micro synthetic data across different scenarios, there is variability in MSE scores, with Scenario 1 having the highest MSE and Scenario 3 having the lowest.

## 7 Privacy Measures

For micro-to-micro data synthesis, we investigate the privacy of the synthesized dataset through a matching approach. In particular, we see if any observations in the synthesized datasets can be matched with an observation in the real dataset.

Figure 4 depicts our results of applying the NewRowSynthesis metric to the macro-to-micro and micro-to-micro synthetic data generation methods. The tolerance threshold indicates the degree of tolerance allowed in the matching process. The match percentage with a score of 100% indicates that all records in the synthetic data are new and do not match any records in the real data.

For the micro-to-micro methods, CART and CTGAN, the match percentage generally decreases as the tolerance increases. This indicates that with higher tolerance, more records in the synthetic data are considered to match records in the real data. We observe that CART presents the highest risk of matching. We see that the match percentage starts decreasing from a tolerance of 0.3. For instance, at a tolerance of 0.6, the synthetic data generated using CART has a match percentage equal to 60%, which means that 40% of the records in the synthetic data can be copies of records in the real data. For the macro-to-micro scenarios, we show our three scenarios' results using scaled and non-scaled data. The match percentages for these scenarios vary as the tolerance increases. We observe that for a tolerance of 0.9, the match percentage achieves a score of



**Fig. 4.** Results of our matching using the NewRowSynthesis metric for Macro-to-Micro (Scenario 1, Scenario 2, Scenario 3) and Micro-to-Micro (CART, CTGAN). A score of 100% means that all records in the synthetic data are new.

around 20%. This indicates that the macro-to-micro synthetic data using the three scenarios can get slightly closer to the real microdata. Our results are not surprising, as we expect a higher disclosure from micro-to-micro data synthesis compared to macro-to-micro data synthesis.

## 8 Conclusion and Future Work

In this work, we explored the generation of macro-to-micro synthetic data using aggregated data sources, with a focus on applications in education and technology testing. Our study aimed to provide a suitable premise for these purposes. We built upon the work of SYNC [11] by proposing and implementing three distinct scenarios: zero knowledge, internal knowledge, and external knowledge, each offering unique approaches to data synthesis. Additionally, we introduced scaling as a means to further improve the synthesis process.

Through extensive experimentation, we compared the performance of the generated macro-to-micro synthetic data to that of micro-to-micro synthetic data. As anticipated, micro-to-micro synthetic data exhibited superior performance compared to macro-to-micro synthetic datasets. However, we identified challenges in ensuring that even micro-to-micro synthetic datasets adhere to provided constraints and rules. To address this issue, we introduced a post-processing approach to validate the analytical validity of the synthetic data, making it more suitable for testing purposes. Furthermore, our evaluation revealed that the CART method demonstrated the best utility performance, closely matching the results of regression models trained on real data. While macro-to-micro scenarios generally exhibited poorer predictive performance, our results hold promise for our intended use cases. Finally, we assessed the disclosure risk using matching techniques and observed that the CART method posed the highest risk of disclosure, as expected. Conversely, macro-to-micro synthetic datasets showed the best protection in terms of disclosure risk.

Looking ahead, several avenues for future research emerge from our findings. Firstly, we aim to explore additional methods and scenarios to further improve the utility and predictive accuracy of macro-to-micro synthetic data. This may involve refining existing synthesis techniques or incorporating novel approaches. Moreover, our work opens avenues for exploring other privacy threats, such as membership inference attacks and identity disclosure, especially in cases where the reconstructed synthetic data closely resembles real data. This highlights the potential risk associated with external knowledge, i.e., ChatGPT.

**Acknowledgments.** We are grateful to Guus van de Burgt for providing the data and invaluable insights, and to Arjen de Boer for his effective project management and extensive knowledge, both of which were crucial to the success of this project.

This work was partly supported by the AI, Media, and Democracy Lab, NWA.1332.20.009.

## A Appendix

### A.1 Dataset Rules

We explain the types of relationships present in the dataset and the rules the synthesizer must account for. We divide the rules into five categories: non-negative rules, summation rules, equality rules, inequality rules, and if-then rules. The *non-negative rule* imposes that the variable must be greater or equal to zero (e.g. number of workers and various expenses). For the *summation* rule, the sum of certain variables must equal another variable per definition. Think of the total expense variable being the sum of individual expenses, or the net revenue being a summation of sales and costs. The *equality* rule suggests that specific variables must be equal to another variable ( $=$ ), while the *inequality* rule imposes that some variables must be lesser or equal ( $\leq$ ) or greater or equal ( $\geq$ ) than other variables. This is the case of repeated variables in the survey (equality rules), and the variable of the number of workers being greater or equal to the full-time equivalent workers (inequality rule). The *if-then* rules impose the same rules as the previous ones, with the addition of requiring a certain condition to be satisfied (as some rules can only exist if a variable meets a certain value).

### A.2 Extra Analytical Validity Results

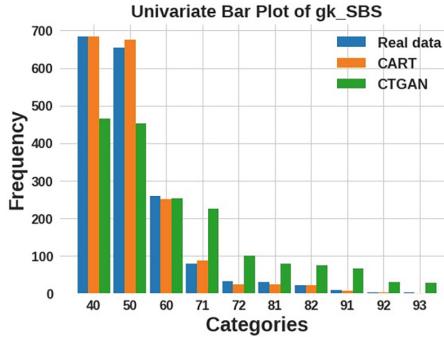
For micro-to-micro data synthesis, we measure the KS-Complement metric, which assesses the similarity between a real variable and its synthetic counterpart based on their marginal distributions. Our results are provided in Tables 3 and 4 (Fig. 5).

**Table 3.** KS-Complements macro-to-micro scaled data

Variable Name	KS-Complement Scenario 1 Scaled	KS-Complement Scenario 2 Scaled	KS-Complement Scenario 3 Scaled
<i>Personal Expenses</i>	0.0031	0.1412	0.0516
<i>Net Revenue</i>	<b>0.3422</b>	0.5585	0.3590
<i>Other Operating Income</i>	<b>0.4199</b>	<b>0.5150</b>	0.9033
<i>Costs of Goods Sold</i>	0.5862	0.8169	0.6457
<i>Other Operating Expenses</i>	0.3848	0.5690	0.4679
<i>Depreciation</i>	<b>0.5496</b>	<b>0.7717</b>	0.7950
<i>Net Financial Income/Expenses</i>	0.6757	0.7807	0.6772
<i>Provisions Allocations</i>	0.2837	<b>0.8820</b>	0.1591

**Table 4.** KS-Complements macro-to-micro

Variable Name	KS-Complement Scenario 1	KS-Complement Scenario 2	KS-Complement Scenario 3
<i>Personal Expenses</i>	0.0494	0.1541	0.0516
<i>Net Revenue</i>	<b>0.5036</b>	0.4843	0.3590
<i>Other Operating Income</i>	<b>0.8075</b>	<b>0.7977</b>	0.9033
<i>Costs of Goods Sold</i>	0.6941	0.6763	0.6457
<i>Other Operating Expenses</i>	0.5484	0.5165	0.4679
<i>Depreciation</i>	<b>0.7301</b>	<b>0.6401</b>	0.7950
<i>Net Financial Income/Expenses</i>	0.7403	0.6786	0.6772
<i>Provisions Allocations</i>	0.2068	<b>0.5156</b>	0.1591



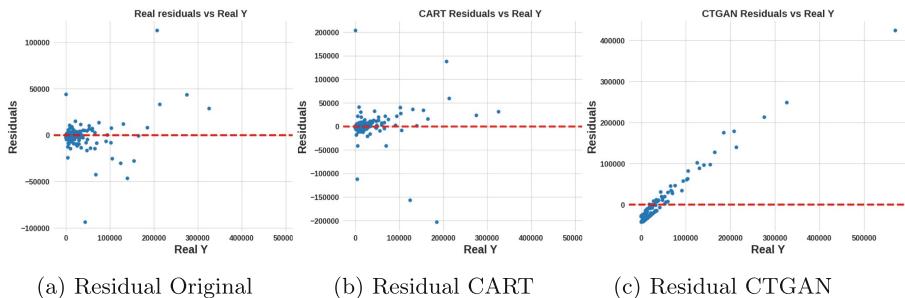
**Fig. 5.** Univariate distribution of the gk\_SBS variable

### A.3 Residuals from the Regression Models

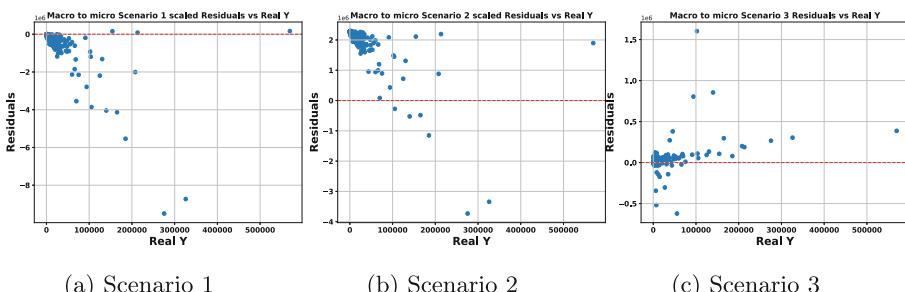
We analyze the performance of the synthetic data by examining the residuals of the linear regression model. Figure 6 shows the residual plots for the models trained on the real data, CART synthetic data, and CTGAN synthetic data. The residuals represent the differences between the observed and predicted values.

In the residual plot for the model trained on the real data (Fig. 6a), the residuals are evenly distributed around zero, indicating a good fit of the model. In contrast, the residual plot for the CART synthetic data (Fig. 6b) shows a similar distribution, though with a slightly larger variance, suggesting that the CART synthetic data maintains the relationships in the real data fairly well. The residual plot for the CTGAN synthetic data (Fig. 6c) reveals a different pattern. The residuals are more spread out and exhibit a clear upward trend, indicating that the CTGAN model has a downward bias and struggles to capture the variability in the real data. This pattern suggests that the CTGAN synthetic data may not be as effective in preserving the underlying relationships present in the real data. Overall, these residual plots highlight that the CART synthetic data provides a better approximation of the real data compared to the CTGAN synthetic data, aligning with the findings from the prediction accuracy analysis.

The residual plots for the synthetic data generated using the macro-to-micro approaches are shown in Fig. 7. In Scenario 1 (Fig. 7a), where no knowledge of the real data was used, the residuals are widely scattered and show a large deviation from zero. This indicates poor accuracy in the synthetic data, as the model fails to capture the true relationships present in the real data. Scenario 2 (Fig. 7b), which incorporates internal knowledge, shows an improvement in the residuals compared to Scenario 1. The residuals are closer to zero, suggesting that the synthetic data better approximates the real data, although there are still noticeable deviations. Scenario 3 (Fig. 7c), which leverages external knowledge, exhibits the most favorable residuals among the three scenarios. The residuals are tightly clustered around zero, indicating a higher accuracy in the synthetic data. This demonstrates that utilizing external insights significantly enhances the quality of the synthetic data generation process.



**Fig. 6.** Residual plot, with the prediction model trained on the real and micro-to-micro synthetic datasets



**Fig. 7.** Residual plot on the synthetic data generated using macro-to-micro scenarios with scaled data.

## References

1. Acharya, A., Sikdar, S., Das, S., Rangwala, H.: GenSyn: a multi-stage framework for generating synthetic microdata using macro data sources. In: IEEE International Conference on Big Data (Big Data), pp. 685–692 (2022)
2. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W.F., Sun, J.: Generating multi-label discrete patient records using Generative Adversarial Networks. In: Doshi-Velez, F., Fackler, J., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (eds.) Proceedings of the 2nd Machine Learning for Healthcare Conference, vol. 68, pp. 286–305 (2017)
3. Choupani, A.A., Mamdoohi, A.R.: Population synthesis using iterative proportional fitting (IPF): a review and future research. Transp. Res. Procedia **17**, 223–233 (2016)
4. Dandekar, R.A., Cohen, M., Kirkendall, N.: Sensitive micro data protection using Latin hypercube sampling technique. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 117–125. Springer, Heidelberg (2002). [https://doi.org/10.1007/3-540-47804-3\\_9](https://doi.org/10.1007/3-540-47804-3_9)
5. Domingo-Ferrer, J.: A survey of inference control methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining: Models and Algorithms, vol. 34, pp. 53–80. Springer, US (2008). [https://doi.org/10.1007/978-0-387-70992-5\\_3](https://doi.org/10.1007/978-0-387-70992-5_3)

6. Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical data protection. *J. Comput. Appl. Math.* **164–165**, 285–293 (2004). Proceedings of the 10th International Congress on Computational and Applied Mathematics
7. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. *Comput. Stat. Data Anal.* **55**(12), 3232–3243 (2011)
8. Garofalo, G., Slokom, M., Preuveneers, D., Joosen, W., Larson, M.: Machine learning meets data modification. In: Batina, L., Bäck, T., Buhan, I., Picek, S. (eds.) *Security and Artificial Intelligence. LNCS*, vol. 13049, pp. 130–155. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-98795-4\\_7](https://doi.org/10.1007/978-3-030-98795-4_7)
9. Hundepool, A., et al.: *Statistical Disclosure Control*. Wiley, New York (2012)
10. Kim, J., Lee, S.: A simulated annealing algorithm for the creation of synthetic population in activity-based travel demand model. *KSCE J. Civ. Eng.* **20**, 2513–2523 (2015)
11. Li, Z., Zhao, Y., Fu, J.: Sync: a copula based framework for generating synthetic data from aggregated sources (2020)
12. Liew, C.K., Choi, U.J., Liew, C.J.: A data distortion by probability distribution. *ACM Trans. Database Syst.* **10**(3), 395–411 (1985)
13. Muralidhar, K.: A re-examination of the Census Bureau reconstruction and re-identification attack. In: Domingo-Ferrer, J., Laurent, M. (eds.) *PSD 2022. LNCS*, vol. 13463, pp. 312–323. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-13945-1\\_22](https://doi.org/10.1007/978-3-031-13945-1_22)
14. Muralidhar, K., Domingo-Ferrer, J.: Database reconstruction is not so easy and is different from reidentification. *J. Off. Stat.* **39**(3), 381–398 (2023)
15. Murata, T., Harada, T.: Nation-wide synthetic reconstruction method. In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6 (2017)
16. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on Generative Adversarial Networks. In: *Proceedings of the 44th International Conference on Very Large Data Bases (VLDB Endowment)*, vol. 11, no. 10, pp. 1071–1083 (2018)
17. Rubin, D.B.: Discussion statistical disclosure limitation. *J. Off. Stat.* **9**(2), 461–468 (1993)
18. Thogarchety, P., Das, K.: Synthetic data generation using genetic algorithm. In: *2023 2nd International Conference for Innovation in Technology (INOCON)*, pp. 1–6 (2023)
19. Torra, V.: Privacy in data mining. In: Maimon, O., Rokach, L. (eds.) *Data Mining and Knowledge Discovery Handbook*, pp. 687–716. Springer, Boston (2009). [https://doi.org/10.1007/978-0-387-09823-4\\_35](https://doi.org/10.1007/978-0-387-09823-4_35)
20. Voas, D., Williamson, P.: An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *Int. J. Popul. Geogr.* **6**, 349–366 (2000)
21. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alche Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 32*, pp. 7335–7345 (2019)

# Author Index

## A

- Aghaddar, Mohamed 420  
Agrawal, Shruti 403  
Allmendinger, Richard 144  
Antunes, Luís 240

## B

- Barnhoorn, Lucas 420  
Blanco-Justicia, Alberto 213  
Brenzel, Hanna 310  
Brown, James Thomas 67

## C

- Carvalho, Tânia 240  
Clayton, Ellen W. 67  
Cohen, Aloni 3  
Cremer, Simon 297

## D

- D’Acquisto, Giuseppe 3  
de Wolf, Peter-Paul 403, 420  
Domingo-Ferrer, Josep 213  
Dove, Iain 102  
Drechsler, Jörg 115, 129, 178

## E

- Elliot, Mark 144

## F

- Fössing, Emma 178  
Francis, Brian 102  
Francis, Paul 161

## G

- Giessing, Sarah 225  
Green, Elizabeth 284

## I

- Iqbal, Mehtab 129

## J

- Jackson, James 102  
Jebreel, Najeeb 213  
Jehmlich, Lydia 297

## K

- Kantarcioglu, Murat 67  
Kikuchi, Hiroaki 329  
Kolb, Simon Xi Ning 225  
Krol, Nynke C. 403  
Kwatra, Saloni 344

## L

- Langsrud, Øyvind 87  
Latner, Jonathan 115  
Lenz, Rainer 297  
Little, Claire 144

## M

- Malin, Bradley A. 67  
Matheny, Michael 67  
Miletic, Marko 374  
Minami, Kazuhiro 35  
Mitra, Robin 102  
Moniz, Nuno 240  
Münnich, Ralf 310  
Muralidhar, Krishnamurty 213, 393

## N

- Naldi, Maurizio 3  
Navarro-Arribas, Guillermo 274  
Neuhoeffer, Marcel 115  
Nissim, Kobbi 3  
Novák, Jiří 194

## O

- Oganian, Anna 129  
Önen, Melek 358  
Oosugi, Hiroto 35

**P**

- Palm, Martin 310  
Parra-Arnau, Javier 358  
Perkonoja, Katarina 51

**R**

- Raab, Gillian M. 254  
Reimherr, Matthew 18  
Ritche, Felix 284  
Ruggles, Steven 393

**S**

- Sánchez, David 213  
Sariyar, Murat 374  
Slavkovic, Aleksandra 18  
Slokom, Manel 403, 420  
Su, Liu Nuo 420  
Sugiyama, Takumi 35

**T**

- Tang, Jui Andreas 225  
Templ, Matthias 194

Thees, Oscar 194

Toma, Ryotaro 329

Torra, Vicenç 274, 344

Tran, Tran 18

Trindade, Carolina 240

**U**

Ünsal, Ayşe 358

**V**

Virta, Joni 51

Vorobeychik, Yevgeniy 67

**W**

Weymeirsch, Jan 310

White, Paul 284

**Y**

Yamanaka, Io 35

**Z**

Zari, Oualid 358