

Master's Thesis

---

# **Towards Data Privacy: Evaluating different synthetic data approaches with the data from Covid-19 Trends and Impact Surveys**

---

Department of Statistics  
Ludwig-Maximilians-Universität München

**Yue Xiong**

Munich, Month Day<sup>th</sup>, 2022



Submitted in partial fulfillment of the requirements for the degree of M. Sc.  
Supervised by Dr. Anna-Carolina Haensch

## **Abstract**

To be completed...

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Related Work</b>	<b>2</b>
2.1	Data Privacy . . . . .	2
<b>3</b>	<b>Synthetic Data and Differential Privacy</b>	<b>3</b>
<b>4</b>	<b>Evaluation of Utility and Risk Assessment of Remaining Disclosure</b>	<b>4</b>
<b>5</b>	<b>Evaluating Different Synthetic Datasets generated from CTIS</b>	<b>5</b>
<b>6</b>	<b>General Discussion</b>	<b>6</b>
<b>7</b>	<b>Conclusion</b>	<b>6</b>
<b>A</b>	<b>Appendix</b>	<b>V</b>
<b>B</b>	<b>Electronic appendix</b>	<b>VI</b>

# 1 Introduction

Bishop (2006) introduced this and that. Another statement that needs a reference, but the authors are not named directly (Bishop, 2006). Another statement where the reference is just one possible source (see, e.g., Bishop, 2006).

## 2 Related Work

In this chapter, general definition of data privacy and the necessity to maintain data privacy are described. Furthermore, in order to keep data privacy, this chapter also presents an overview of methods to achieve the goal of privacy preserving data analysis and publication, which have been adopted in several domains. Here, we want to emphasize the use of synthetic data and one of its most popular applications, i.e., differentially private synthetic data, that are able to prevent disclosure in the process of synthetic data generation.

### 2.1 Data Privacy

It is well-acknowledged that we have entered a data-driven world and data are often regarded as significant constituents for our society. At the same time, an open society can also learn from these data so as to develop feasible and practical policy guidelines (Evans and King, 2021). Especially during the outbreak of coronavirus disease 2019 (COVID-19), more and more increased concerns are raised that it is essential for a society to utilize such data, which are widely-spread in the population and analyzed with regard to various perspectives, to advance sophisticated planning and develop more concrete social welfare benefits for the citizens. Consequently, both seen from the public health perspective and the economy perspective, the on-going COVID-19 global pandemic serves as a rigid reminder that detailed data are urgently needed to assist in decision making, damage control scenarios. Regardless of prevalent consensus reached to leverage more microdata, the inappropriate use of such information can cause harm in data confidentiality and privacy as sometimes the attacks from an intruder may result in the leakage of an individual's sensitive information, e.g., identity, address and salary, etc.

On the premise of possible outcomes brought by the misuse of microdata, it is crucial that we encourage proper and legal use of the collected datasets.

### 3 Synthetic Data and Differential Privacy

Bishop (2006) introduced this and that. Another statement that needs a reference, but the authors are not named directly (Bishop, 2006). Another statement where the reference is just one possible source (see, e.g., Bishop, 2006).

## 4 Evaluation of Utility and Risk Assessment of Remaining Disclosure

Bishop (2006) introduced this and that. Another statement that needs a reference, but the authors are not named directly (Bishop, 2006). Another statement where the reference is just one possible source (see, e.g., Bishop, 2006).

## 5 Evaluating Different Synthetic Datasets generated from CTIS

Additional material goes here



## **6 General Discussion**

A concise summary of contents and results

## **7 Conclusion**

A concise summary of contents and results

# A Appendix

Additional material goes here

## **B Electronic appendix**

Data, code and figures are provided in electronic form.

## References

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer.
- Evans, G. and King, G. (2021). Statistically valid inferences from differentially private data releases, with application to the facebook urls dataset, *Political Analysis* pp. 1–21.

## Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Location, date

---

Name