INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

BUYER BEWARE: UNDERSTANDING THE TRADE-OFF BETWEEN UTILITY AND RISK IN CART BASED MODELS USING SIMULATION DATA

Berlin,
7-8. Oktober, 2024

Jonathan Latner, PhD
Dr. Marcel Neunhoeffer
Prof. Dr. Jörg Drechsler

SECTION 1: INTRODUCTION

BACKGROUND

- We don’t know.
- We think this idea is interesting.
- We think others should know about it.
- But we are still developing it.
- Thoughts and suggestions would be helpful.

OVERVIEW

- It is well established that there is a trade-off between utility and privacy when generating synthetic data
- Utility and privacy in CART based synthesizers is high (Little et al., 2022; Danker and Ibrahim, 2021)
- Therefore, CART models are less sensitive to this trade-off than other SDGs (i.e. higher utility, lower risk)
- Using simulation data (Reiter et al., 2014), we show that synthetic data from CART models are disclosive
- The problem:
  - Disclosive in ways that are not observable with common privacy metrics
  - It is possible to increase protection (by reducing utility)

SECTION 2: SET UP THE ATTACK

Let me introduce a problem illustrating that CART is disclosive

THE ORIGINAL DATA: SIMULATE DATA WITH A UNIQUE RECORD

Borrowing from Reiter et al. (2014), we set the first 999 observations to be a random sample from a multinomial distribution for all combinations of $var1(0,1)$, $var2(0,1)$, $var3(0,1)$, $var4(0,1)$ except $(var1 = 1, var2 = 1, var3 = 1, var4 = 1)$, which we set to be the 1000^{th} observation.

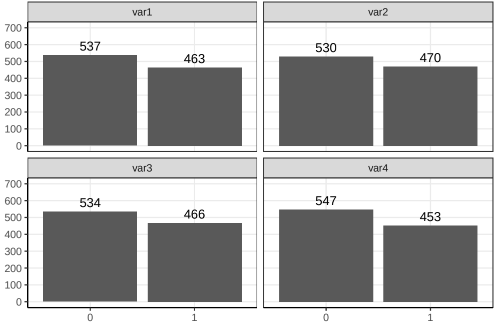


Figure 1: Frequency

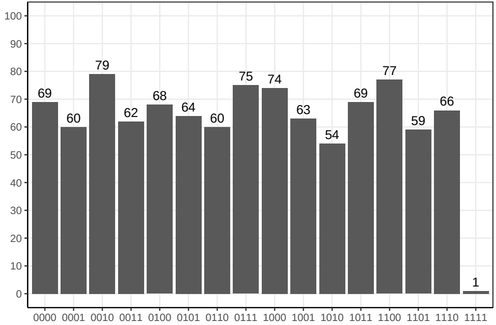


Figure 2: Histogram

GENERATE SYNTHETIC DATA WITH CART (SYNTHPOP)

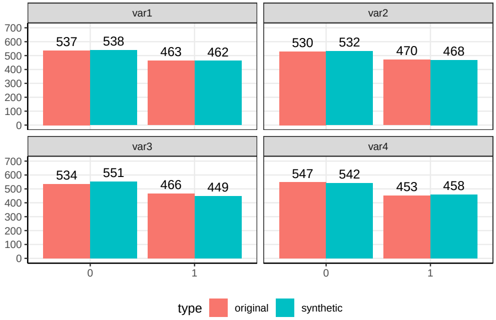


Figure 3: Frequency

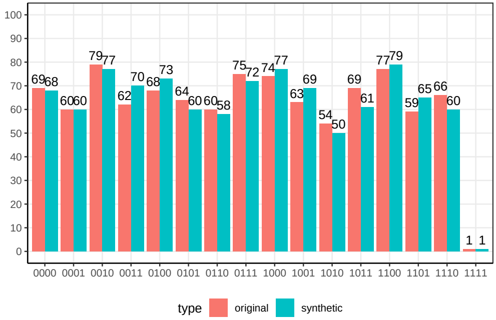
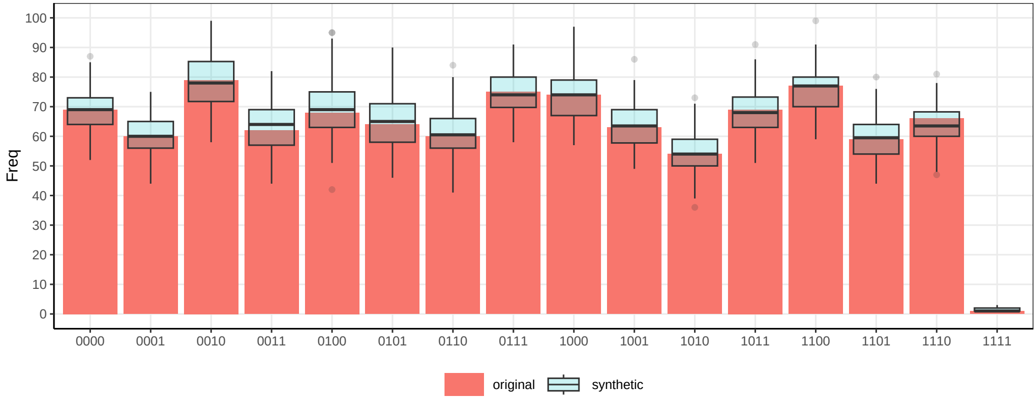


Figure 4: Histogram

COMPARE HISTOGRAM X 100 SYNTHETIC DATASETS

Figure 5: Multiple synthetic data sets does not reduce privacy risk



SETTING UP THE ATTACK

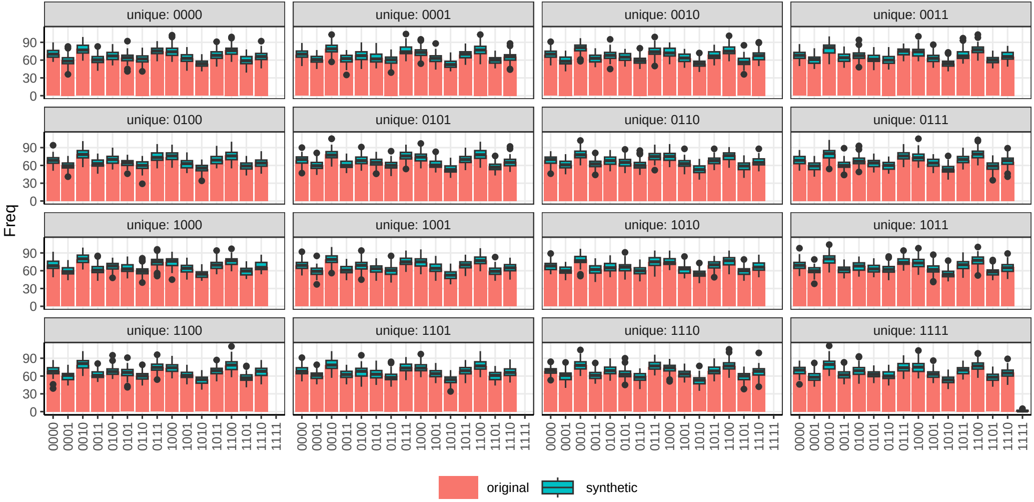
- Its a game between two entities.
  - The statistical agency has the data and wants to release it in a privacy preserving way.
  - The attacker wants to identify someone in the data (either membership or attribute inference).
- We assume the attacker has the following knowledge
  - Knows the SDG model type (i.e. sequential CART).
  - Knowledge of everyone in the data except one.
  - The 16 possible combinations that the last one could be.
- The question: What can the attacker learn from a released synthetic data set about that individual they do not have knowledge of?

DESCRIBING THE ATTACK

- The attacker sees the synthetic data
- The attacker runs the same synthetic data model (SDG) for all of the 16 different possibilities.
- Then they want to update their beliefs about what the last record could be
- The synthetic data contains that 16th category only in the case where the last observation is the unique one.

ILLUSTRATING THE ATTACK

Figure 6: Histogram of 16 worlds x 100 synthetic datasets



SUMMARY

- The problem: CART can be disclosive
- The reason: If a record is not in the original data, then it can never be in the synthetic data. Therefore, a record can only be in the synthetic data if it is also in the original data.
- If you knew every record but one, and you knew every possible combination of unique values, then you could identify the missing record.
- What do the privacy metrics in Synthpop tells us?

COMPARING PRIVACY MEASURES (SET.SEED = 1237, I.E. UNIQUE = 1)

```
> print(t1, plot = FALSE, to.print = "ident")
Disclosure risk for 1000 records in the original data

Identity disclosure measures
from keys: var1 var2 var3
For original  ( UiO )  0 %
For synthetic ( repU ) 0 %.
> print(t1, plot = FALSE, to.print = "attrib")

Table of attribute disclosure measures for var1 var2 var3
Original measure is  Dorig and synthetic measure is DiSCO
Variables Ordered by synthetic disclosure measure

       attrib.orig attrib.syn check1 Npairs check2
1 var4           0          0             0
```

```
> replicated.uniques (sds, df_ods)
    var1 var2 var3 var4
973    1    1    1    1
Uniques and replicated uniques for  1  synthesised data set(s)
 from keys:  var1 var2 var3 var4

Uniques in  original data:
 1 from  1000 records ( 0.1 %)
Uniques in synthetic data:
 1 from  1000 records ( 0.1% )

Replicated uniques:
 1
as a % of uniques in synthetic  100%
as a % of original records (repU) 0.1%
```

COMPARING PRIVACY MEASURES (SET.SEED = 1240, I.E. UNIQUE = 3)

```
> print(t1, plot = FALSE, to.print = "ident")
Disclosure risk for 1000 records in the original data

Identity disclosure measures
from keys: var1 var2 var3
For original  ( UiO )  0 %
For synthetic ( repU ) 0 %.
> print(t1, plot = FALSE, to.print = "attrib")

Table of attribute disclosure measures for var1 var2 var3
Original measure is  Dorig and synthetic measure is DiSCO
Variables Ordered by synthetic disclosure measure

       attrib.orig attrib.syn check1 Npairs check2
1 var4           0          0             0
```

```
> replicated.uniques (sds, df_ods)
Uniques and replicated uniques for  1  synthesised data set(s)
 from keys:  var1 var2 var3 var4

Uniques in  original data:
 1 from  1000 records ( 0.1 %)
Uniques in synthetic data:
 0 from  1000 records ( 0% )

Replicated uniques:
 0
as a % of uniques in synthetic  NaN%
as a % of original records (repU) 0%
```

SUMMARY

- Using common privacy and utility measures, CART generates synthetic data with both high utility and low risk
- However (and this is the point):
  - We know there is a problem (because we created it)
  - We know that common measures do not capture the problem
- We are also not alone in identifying this problem (Manrique-Vallier and Hu, 2018)

SECTION 3: SOLUTION

- The good news: we know how to solve the problem
- The bad news: we don’t know how to identify the problem

THE GOOD NEWS: SOLUTIONS

- cp = 0.05 (default = $1e^{-8}$): prevent large trees (i.e. overfitting)
- minbucket = 75 (default = 5): the minimum number of observations in any terminal node
- Other options also exist

COMPARE HISTOGRAM X 100 SYNTHETIC DATASETS

Figure 7: CART (default)

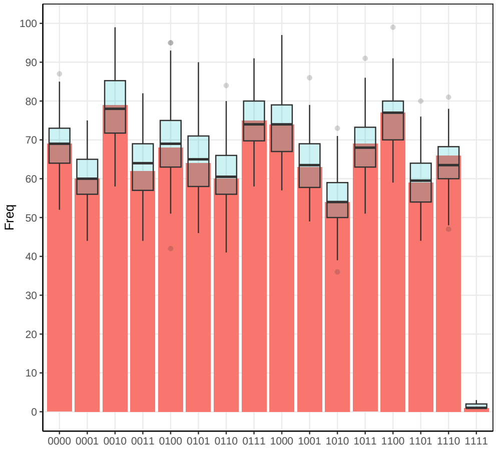
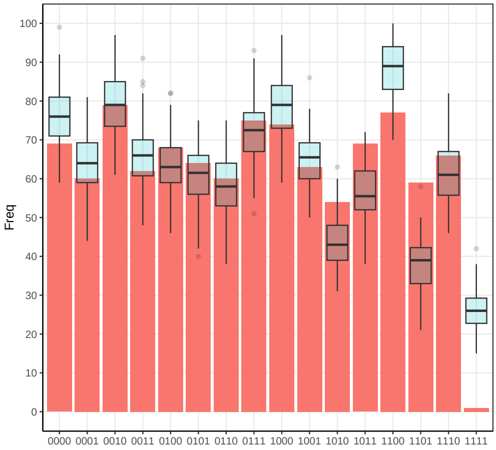


Figure 8: CART (modified)



THE BAD NEWS

- Specifically, we don’t know how to identify the privacy risk
- Generally, we have to know a problem exists before we would do something about it

SECTION 4: CONCLUSION

CONCLUSION

- It has long been understood that there is a trade-off between utility and risk
- Previous research indicated that CART models were less sensitive to this trade-off than other SDGs
- Using a simulated data set, we show that CART are sensitive to this trade-off
- The good news: It is possible to reduce risk in CART with parameters
- The bad news:
  - We must sacrifice utility
  - Common privacy metrics do not capture risk in our simulated data
- Question: If you did not know there was a problem, why would you sacrifice utility?

CONTACT

Jonathan Latner

jonathan.latner@iab.de

Reproducible code:

- Github: https://github.com/jonlatner/KEM_GAN/tree/main/latner/projects/simulation