



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 Neunhoeffler
Prof. Dr. Jörg Drechsler



SECTION 1: INTRODUCTION

BACKGROUND

- This is work in progress.
- We think this idea is interesting.
- We think others should know about it.
- But we are still developing it.
- Comments and critiques 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
- **The solution:** It is possible to increase protection (by reducing utility)
- **The question:** Why would we reduce utility if we did not know there was a problem?

SECTION 2: SET UP THE ATTACK

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).
- The question: What can the attacker learn from a released synthetic data set about an individual they do not have knowledge of?

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 1000th observation.

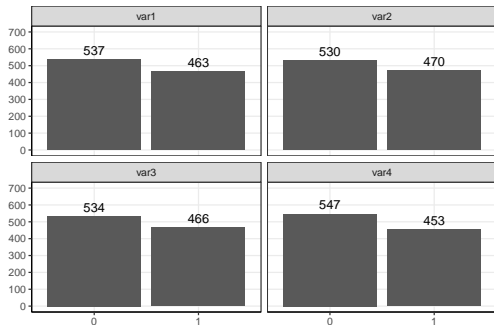


Figure 1: Frequency

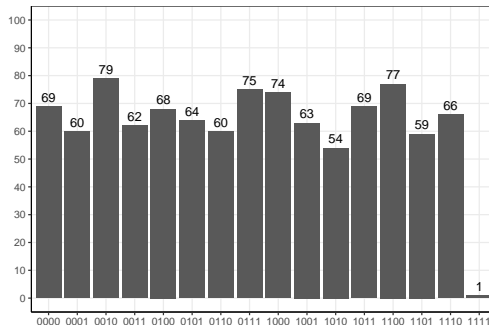


Figure 2: Histogram

GENERATE SYNTHETIC DATA WITH CART (SYNTHPOP)

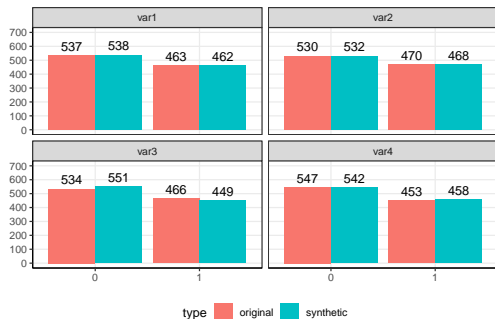


Figure 3: Frequency

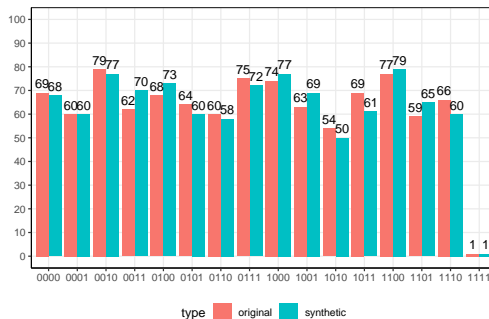
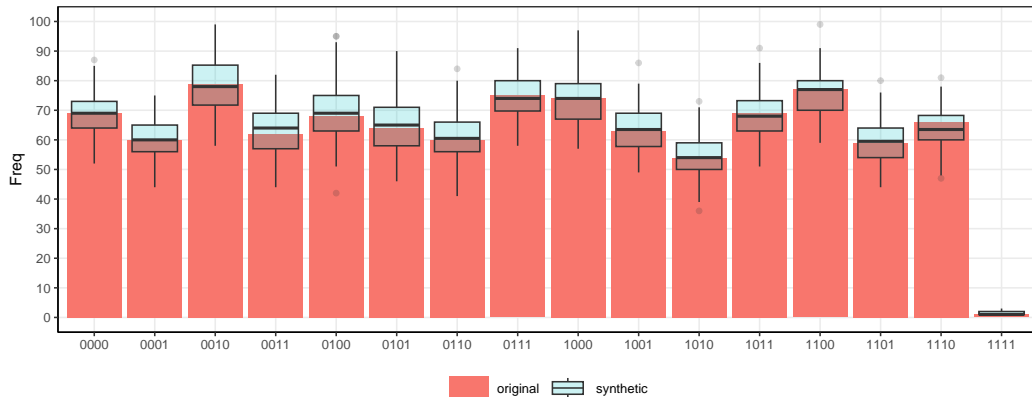


Figure 4: Histogram

COMPARE HISTOGRAM X 100 SYNTHETIC DATASETS

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

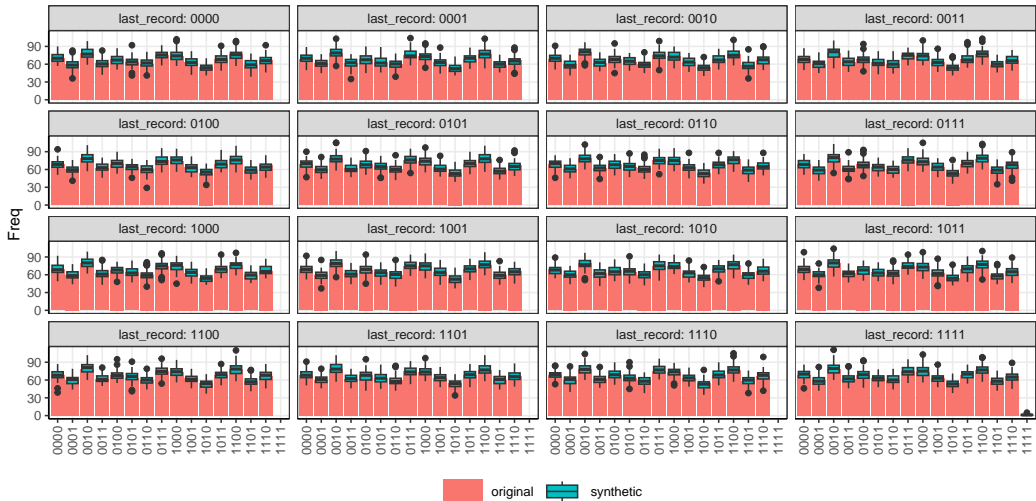


DESCRIBING THE ATTACK

- 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 attacker sees the synthetic data
- The attacker runs the same synthetic data model (SDG) for all of the 16 different possibilities.
- Then they update their beliefs about what the last record could be

ILLUSTRATING THE ATTACK

Figure 6: Histogram of 16 worlds x 100 synthetic datasets



SUMMARY

- The problem: CART can be disclosive
- The reason (in our data):
 - A record can only be in the synthetic data if it is also in the original data.
 - Or the opposite: if a record is not in the original data, then it can never be in the synthetic data.

SECTION 3: MEASURING PRIVACY

PRIVACY MEASURES

- Synthpop (Raab et al., 2024)
 - Identity disclosure: the ability to identify individuals in the data from a set of known characteristics (i.e. 'keys').
 - Attribute disclosure: the ability to find out from the keys something, not previously known
 - Replicated uniques

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

```
1 > print(t1, plot = FALSE, to.print = "ident")
2 Disclosure risk for 1000 records in the original data
3
4 Identity disclosure measures
5 from keys: var1 var2 var3
6 For original ( Ui0 ) 0 %
7 For synthetic ( repU ) 0 %.
8 > print(t1, plot = FALSE, to.print = "attrib")
9
10 Table of attribute disclosure measures for var1 var2 var3
11 Original measure is Dorig and synthetic measure is DiSCO
12 Variables Ordered by synthetic disclosure measure
13
14      attrib.orig attrib.syn check1 Npairs check2
15 1 var4          0          0          0
```

```
1 > replicated.uniques (sds, df_ods)
2      var1 var2 var3 var4
3 973      1      1      1
4 Uniques and replicated uniques for 1 synthesised data set(s)
5 from keys: var1 var2 var3 var4
6
7 Uniques in original data:
8 1 from 1000 records ( 0.1 %)
9 Uniques in synthetic data:
10 1 from 1000 records ( 0.1% )
11
12 Replicated uniques:
13 1
14 as a % of uniques in synthetic 100%
15 as a % of original records (repU) 0.1%
```

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

```
1 > print(t1, plot = FALSE, to.print = "ident")
2 Disclosure risk for 1000 records in the original data
3
4 Identity disclosure measures
5 from keys: var1 var2 var3
6 For original ( Ui0 ) 0 %
7 For synthetic ( repU ) 0 %.
8 > print(t1, plot = FALSE, to.print = "attrib")
9
10 Table of attribute disclosure measures for var1 var2 var3
11 Original measure is Dorig and synthetic measure is DiSCO
12 Variables Ordered by synthetic disclosure measure
13
14      attrib.orig attrib.syn check1 Npairs check2
15 1 var4           0           0         0
```

```
1 > replicated.uniques (sds, df_ods)
2 Uniques and replicated uniques for 1 synthesised data set(s)
3 from keys: var1 var2 var3 var4
4
5 Uniques in original data:
6 1 from 1000 records ( 0.1 %)
7 Uniques in synthetic data:
8 0 from 1000 records ( 0% )
9
10 Replicated uniques:
11 0
12 as a % of uniques in synthetic NaN%
13 as a % of original records (repU) 0%
```


SUMMARY

- Using common privacy measures, CART generates synthetic data with low risk
- 1 measure indicates there may be a problem, but all the other measures indicate there is no problem.
- 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

SECTION 4: SOLUTION

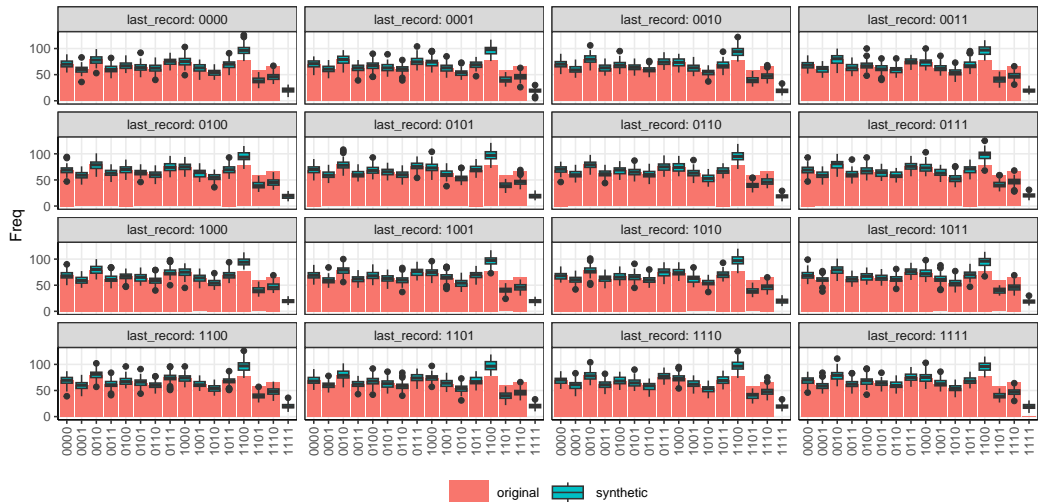
- The good news: we know how to solve the problem
- The bad news: we don't know how to identify the problem

THE GOOD NEWS: SOLUTIONS

- Reduce utility by preventing overfitting
 - minbucket = 75 (default = 5): increase the minimum number of observations in any terminal node
 - cp = 0.05 (default = $1e^{-8}$): decrease the size of the tree
 - Other options also exist

COMPARE HISTOGRAM X 100 SYNTHETIC DATASETS

Figure 7: Histogram of 16 worlds x 100 synthetic datasets



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 5: CONCLUSION

SUMMARY

- 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:
 - Common privacy metrics do not capture risk in our simulated data
 - We must sacrifice utility
- Question: If you did not know there was a problem, why would you sacrifice utility?

IS THE SCENARIO REALISTIC? IS THIS A PROBLEM?

- No, this is not a problem. It is not realistic to assume that
 - A synthetic dataset would be generated with only binary variables
 - A scenario where an intruder would know all observations except one
- Yes, this is a problem
 - Unique records
 - are always the records we need to protect most
 - It is well known that SDGs struggle to protect unique records
 - The simulation
 - We show that disclosure happened
 - We show that risk measures did not capture the disclosure

CONCLUSION

- We are not saying (more specifically)
 - That all synthetic data are disclosive
 - That CART-based SDGs are bad
- Instead, we are saying (more generally)
 - Do not assume that risk measures will identify all problems
 - This simulation offers a type of ‘bound’ on understanding disclosure risks

CONTACT

Jonathan Latner

`jonathan.latner@iab.de`

Reproducible code:

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