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

AnigeD

Finanziert von der Europäischen Union
NextGenerationEU

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

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

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

SECTION 1: GENERATE THE ORIGINAL AND SYNTHETIC DATA

- Borrowing from Reiter et al. (2014), we create a data set with $n = 1000$ and 4 dichotomous, categorical variables.
- 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 the last one
- The last (1000^{th}) observation is $(var1 = 1, var2 = 1, var3 = 1, var4 = 1)$.

GENERATE ORIGINAL DATA USING A SIMULATION

Figure 1: Frequency

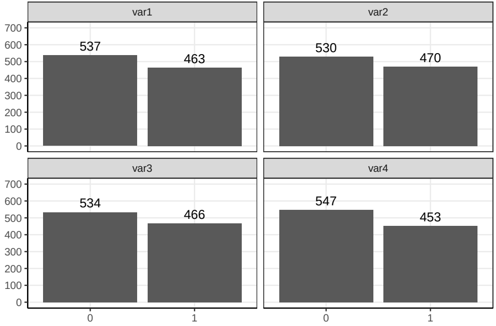
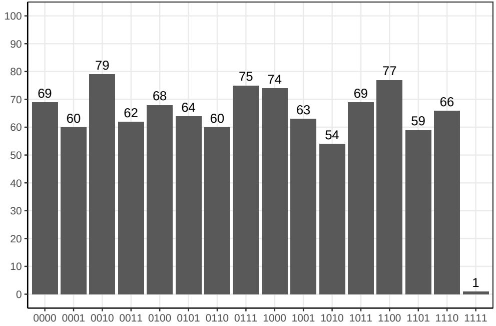


Figure 2: Histogram



GENERATE SYNTHETIC DATA WITH CART (SYNTHPOP)

Figure 3: Frequency

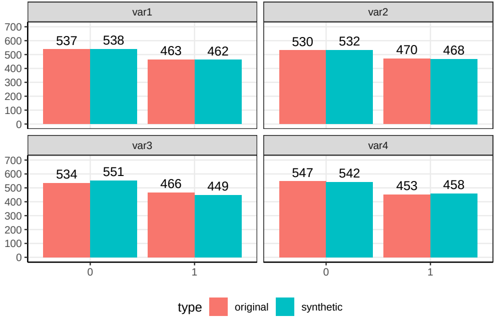
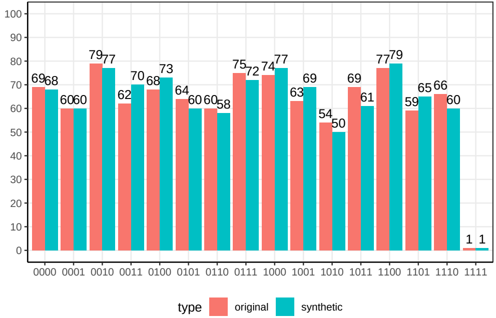
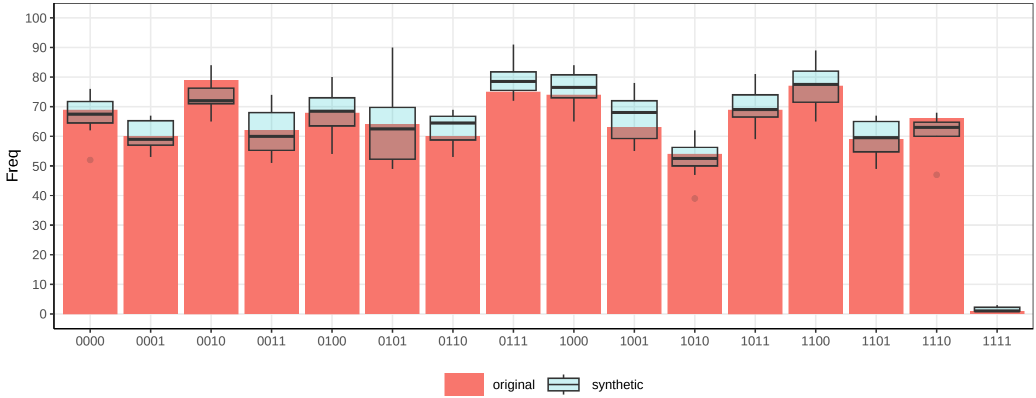


Figure 4: Histogram



COMPARE HISTOGRAM X 10 SYNTHETIC DATASETS

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



SUMMARY

- The problem (in our data): Synthetic data from CART models are disclosive
- The reason:
  - A record can only be in the synthetic data if it is also in the original data (in this simulated data).
  - Or the opposite: if a record is not in the original data, then it can never be in the synthetic data.
- Next section: Can an attacker identify the disclosure?

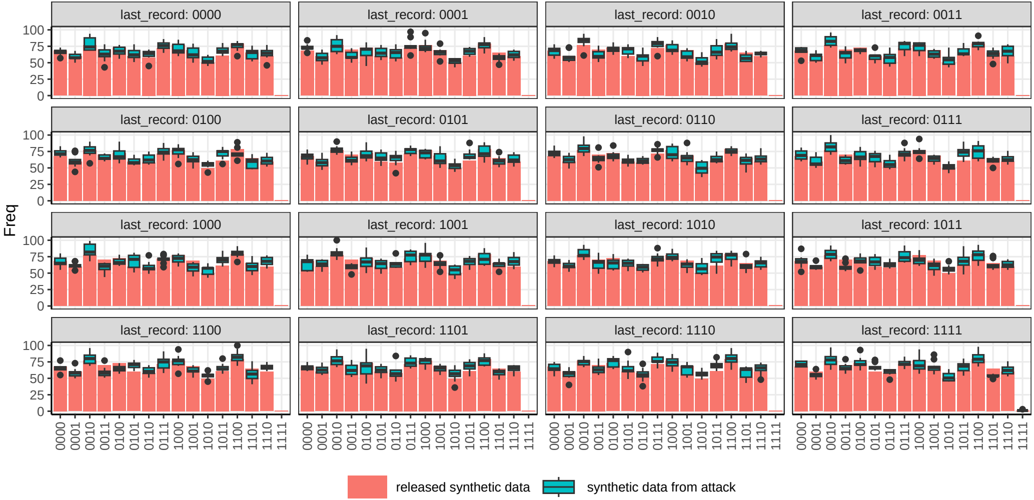
SECTION 2: THE ATTACK

DESCRIBING THE ATTACK

- We assume a ‘strong’ attacker similar to the attack model in differential privacy (DP).
- An attacker has the following knowledge
  - Knows the SDG model type (i.e. sequential CART).
  - Knowledge of all observations in the data except the last 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 WITH CART (DEFAULT PARAMETERS)

Figure 6: Histogram of 16 worlds x 100 synthetic datasets



SUMMARY

- In our attack with our assumptions, the attacker can easily identify the last record
- The reason (to repeat):
  - A record can only be in the synthetic data if it is also in the original data (in this simulated data).
  - Or the opposite: if a record is not in the original data, then it can never be in the synthetic data.
- Next section: Can we measure this disclosure?

SECTION 3: MEASURING PRIVACY

COMMON PRIVACY MEASURES - SYNTHPOP (RAAB ET AL., 2024)

- Replicated uniques (#)
- Identity disclosure (%): the ability to identify individuals in the data from a set of known characteristics or ‘keys’ (*q*).
- Attribute disclosure (%): the ability to find out from the keys something, not previously known or ‘target’ (*t*)

ATTRIBUTE DISCLOSURE

Disclosive in Synthetic: is the percent of records in SD where the keys (*q*) identify a unique target (*t*). In our case, when there is no unique record in the SD, this equals the percent of records with 1110 in SD.

$$D_{syn} = 100 \sum^{q} \sum^{t} (s_{tq} \mid ps_{tq} = 1)/N_s \qquad (1)$$

Disclosive in Synthetic: is the percent of all records in OD where *q* in SD is disclosive (i.e., *t* values for *q* are constant in SD). In our case, this is the percent records with 1111 or 1110 in the OD (i.e. 67%) when there is no unique record in SD.

$$DiS = 100 \sum^{q} \sum^{i=1,...,T} \sum^{j=1,...,T} (d_{iq} \mid ps_{jq} = 1)/N_d \qquad (2)$$

% Disclosive in Synthetic Correct in Original: percent of all records in OD where *q* in SD is disclosive and the disclosed *t* value matches the true *t* value in OD. In our case, this is the percent records with 1110 in the OD (i.e. 66%) when there is no unique record in SD.

$$DiSCO = 100 \sum^{q} \sum^{t} (d_{tq} \mid ps_{tq} = 1)/N_s \qquad (3)$$

COMPARING DISCLOSURE RISK MEASURES

Table 1: x 1 synthetic data set (seed = 1237)

data	identity	unique	attribute
Original	0.00	1.00	0.00
Synthetic	0.00	1.00	0.00

Table 2: x 10 synthetic data sets

data	identity	unique	attribute
Original	0.00	1	0.00
Synthetic	0.00	see table 3	1.32

UNDERLYING INFORMATION

Table 3: Frequency statistics

	Original	Synthetic Data									
Combine	0	1	2	3	4	5	6	7	8	9	10
0000	69	68	66	71	73	76	62	72	52	64	67
0001	60	60	53	57	56	58	60	67	67	57	67
0010	79	77	71	73	71	71	84	65	70	77	74
0011	62	70	51	56	68	63	55	74	57	68	52
0100	68	73	63	80	54	61	79	65	73	66	71
0101	64	60	77	49	66	52	90	52	53	65	71
0110	60	58	68	66	61	69	56	67	65	64	53
0111	75	72	91	86	81	80	77	82	77	75	72
1000	74	77	84	80	73	70	81	82	65	76	73
1001	63	69	66	57	68	73	56	68	75	78	55
1010	54	50	54	57	51	47	50	39	62	58	54
1011	69	61	59	77	71	66	69	75	69	68	81
1100	77	79	77	76	83	78	66	65	88	70	89
1101	59	65	52	54	57	66	67	59	65	49	60
1110	66	60	68	60	64	68	47	65	62	64	60
1111	1	1	0	1	3	2	1	3	0	1	1

Table 4: Attribute risk measures from 10 synthetic data sets

m	Dsyn	iS	DiS	DiSCO
1	0	100	0	0
2	6.8	100	6.7	6.6
3	0	100	0	0
4	0	100	0	0
5	0	100	0	0
6	0	100	0	0
7	0	100	0	0
8	6.2	100	6.7	6.6
9	0	100	0	0
10	0	100	0	0
Average	1.3	100	1.34	1.32

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
- We are also not alone in identifying this problem (Manrique-Vallier and Hu, 2018)

SECTION 4: SOLUTION

THE GOOD NEWS: SOLUTIONS

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

GENERATE SYNTHETIC DATA WITH CART (MODIFIED PARAMETERS)

Figure 7: minbucket

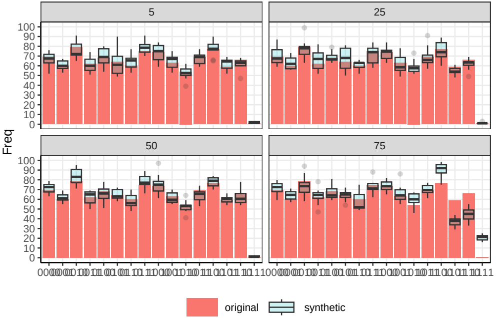
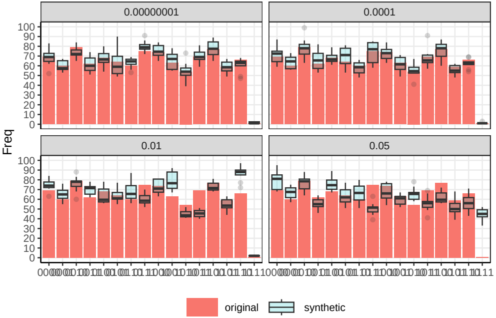
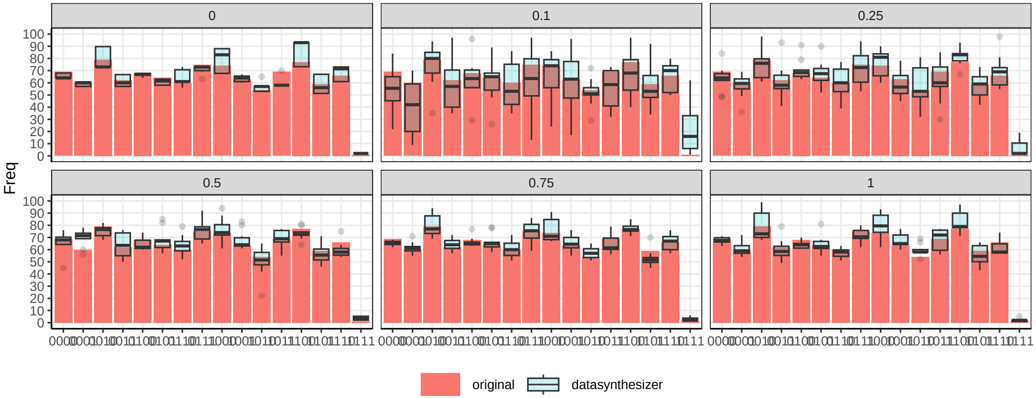


Figure 8: cp



OTHER OPTIONS: GENERATE NOISE WITH ϵ-DP

Figure 9: Datasynthesizer with DP



ILLUSTRATING THE ATTACK WITH CART (MODIFIED PARAMETERS)

Figure 10: mb = 75

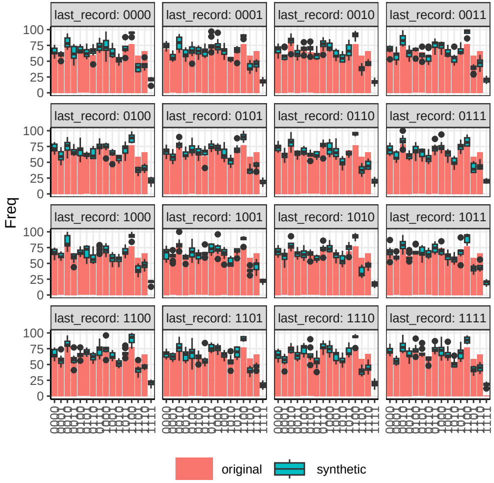
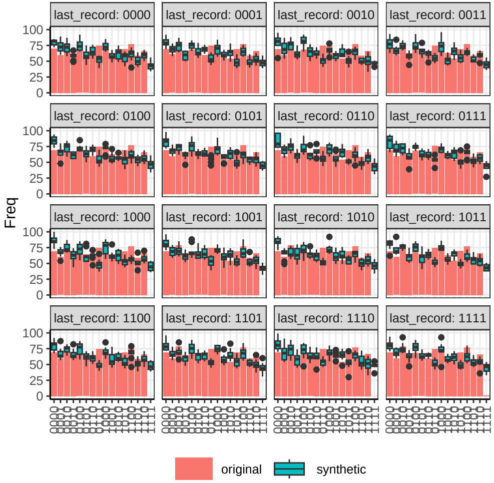


Figure 11: cp = 0.05



THE BAD NEWS

- We don’t know how to identify the privacy risk
- 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.
  - A ‘strong’ attacker is unrealistic.
    - Knows the SDG model type (i.e. sequential CART).
    - Knowledge of al observations in the data except the last one.
    - The 16 possible combinations that the last one could be.
- 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 while also providing utility
    - In this data, eliminating unique records does not solve the problem
  - The simulation
    - We show that a disclosure happened in this data
    - We show that these risk measures did not capture this disclosure

CONCLUSION

- We are not saying:
  - All synthetic data are disclosive
  - CART-based SDGs are disclosive
- We are saying:
  - Do not assume that all risk measures will identify all problems
  - This simulation offers a type of ‘bound’ on understanding disclosure risks

THANK YOU

Jonathan Latner: jonathan.latner@iab.de

Reproducible code: https://github.com/jonlatner/KEM_GAN/tree/main/latner/projects/simulation