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 (1000th) observation is ($var1 = 1, var2 = 1, var3 = 1, var4 = 1$).

GENERATE ORIGINAL DATA USING A SIMULATION
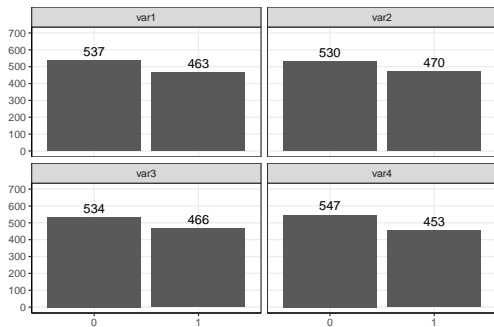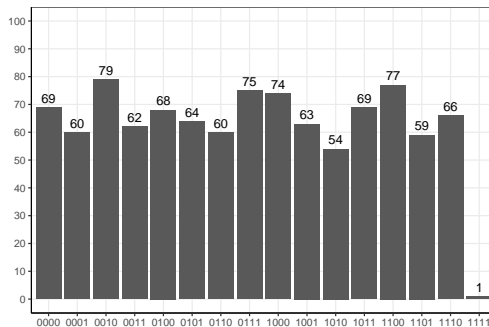
Figure 1: Frequency

Figure 2: Histogram

GENERATE SYNTHETIC DATA WITH CART (SYNTHPOP)
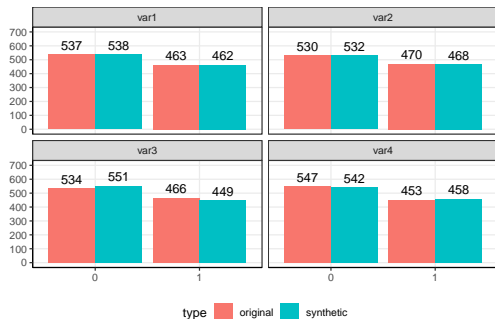
Figure 3: Frequency

Figure 4: Histogram
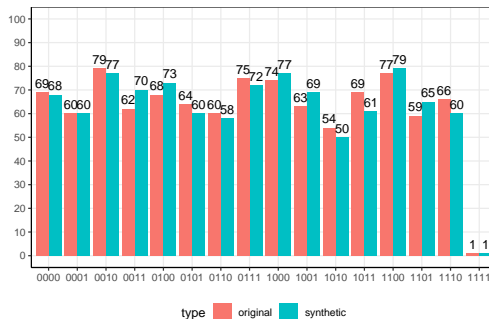
COMPARE HISTOGRAM X 10 SYNTHETIC DATASETS

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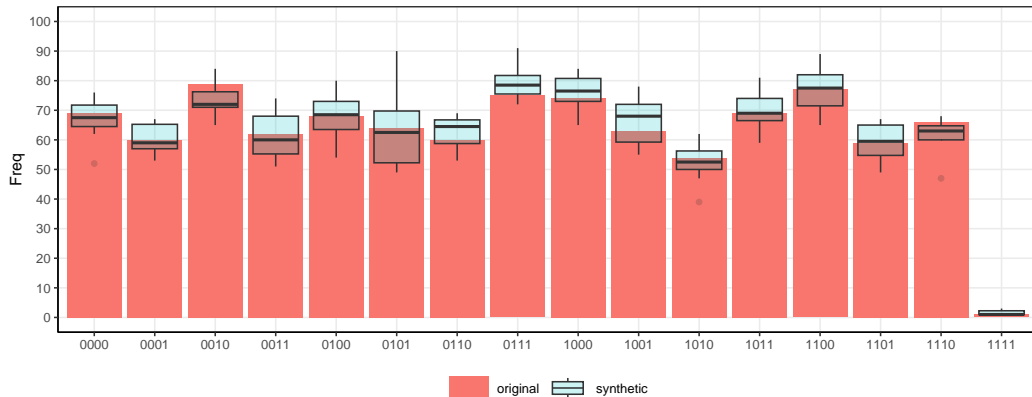
SUMMARY

- 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

released synthetic data   synthetic data from attack

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

The literature on privacy measures for synthetic data is well-developed (Wagner and Eckhoff, 2018).

Common privacy measures - Synthpop (Raab et al., 2024)

- 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)

We will show that attribute disclosure risk only indicates risk when there is no risk

IDENTITY DISCLOSURE

repU (replicated uniques) are unique records in the original data that are also unique in synthetic data and is the measure of identity risk. Formally, *repU* is defined by equation 1:

$$repU = 100 \sum (s_{.q} | d_{.q} = 1 \wedge s_{.q} = 1)/N_d \qquad (1)$$

where $d_{.q}$ is the count of records in the original data with the keys corresponding to a given value of q and $s_{.q}$ is the equivalent count for the synthetic data.

In a given value of q, $s_{.q} | d_{.q} = 1$ is a unique record in the original data conditional on also existing in the synthetic data.

AND

$s_{.q} = 1$ is the unique record in the synthetic data.

This is summed over unique values of q and divided by the total number of records in the data (N_d) and multiplied by 100 to transform the count into a percentage.

ATTRIBUTE DISCLOSURE

DiSCO (Disclosive in Synthetic Correct in Original) is the subset of records in the original data for which the keys (q) in the synthetic data is disclosive. q is disclosive if all records in the synthetic data with the same q have a constant target (t), i.e. no variation in t, as defined by the following equation 2:

$$DiSCO = 100 \sum^{q} \sum^{t} (d_{tq}|ps_{tq} = 1)/N_d \tag{2}$$

where $d_{tq}|s_{tq} = 1$ indicates whether the synthetic data matches the original data for the combination of t and q given the condition that the synthetic data for the combination of t and q is disclosive (i.e., target t is uniquely determined by the keys q).

This is summed over unique values of t and unique values of q and divided by the total number of records in the data (N_d) and multiplied by 100 to transform the count into a percentage.

COMPARING DISCLOSURE RISK MEASURES

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

data	identity	attribute
Original	0.00	0.00
Synthetic	0.00	0.00

Table 2: x 10 synthetic data sets

data	identity	attribute
Original	0.00	0.00
Synthetic	0.00	1.32

UNDERLYING INFORMATION

Table 3: Frequency statistics

Combine	Original	Synthetic Data									
	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: Disclosure risk measures from 10 synthetic data sets

data	identity	attribute
Original	0.00	0.00
Synthetic 1	0.00	0.00
Synthetic 2	0.00	6.60
Synthetic 3	0.00	0.00
Synthetic 4	0.00	0.00
Synthetic 5	0.00	0.00
Synthetic 6	0.00	0.00
Synthetic 7	0.00	0.00
Synthetic 8	0.00	6.60
Synthetic 9	0.00	0.00
Synthetic 10	0.00	0.00
Average	0.00	1.32

SUMMARY

- Using common privacy measures, CART generates synthetic data with 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
- Further, *DiSCO* only captures attribute risk when there is no attribute risk (i.e. no unique observation)

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 noise with differential privacy (ϵ-DP)

    – CTREE vs. CART (variables as factors - bug, not a feature)

GENERATE SYNTHETIC DATA WITH CART (MODIFIED PARAMETERS)

Figure 7: minbucket
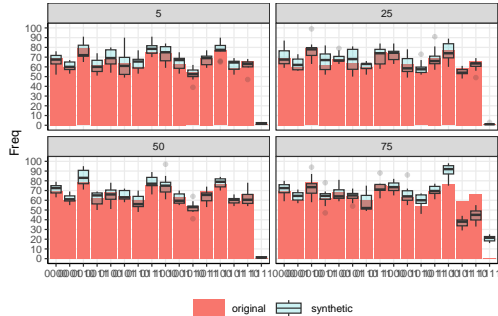
Figure 8: cp
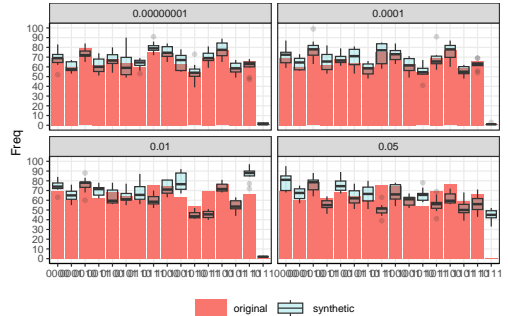
ILLUSTRATING THE ATTACK WITH CART (MODIFIED PARAMETERS)
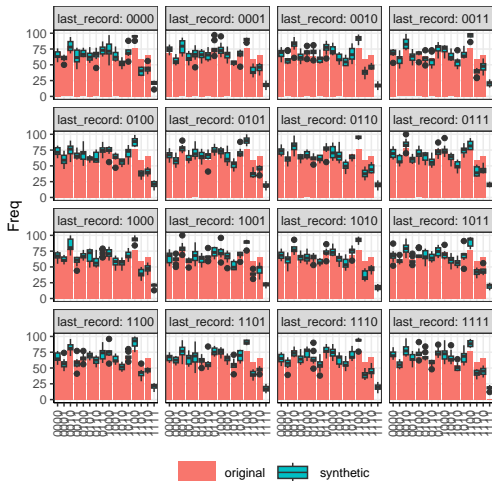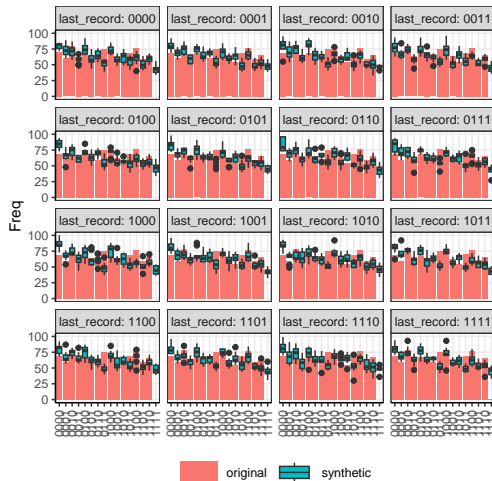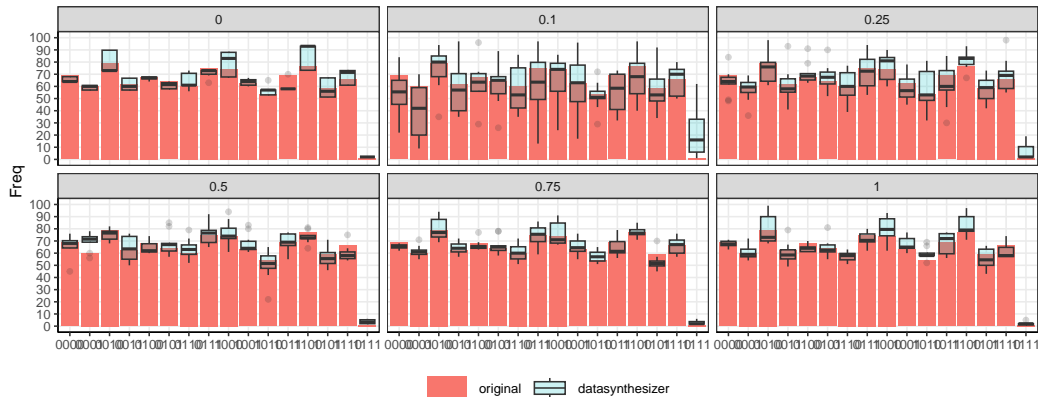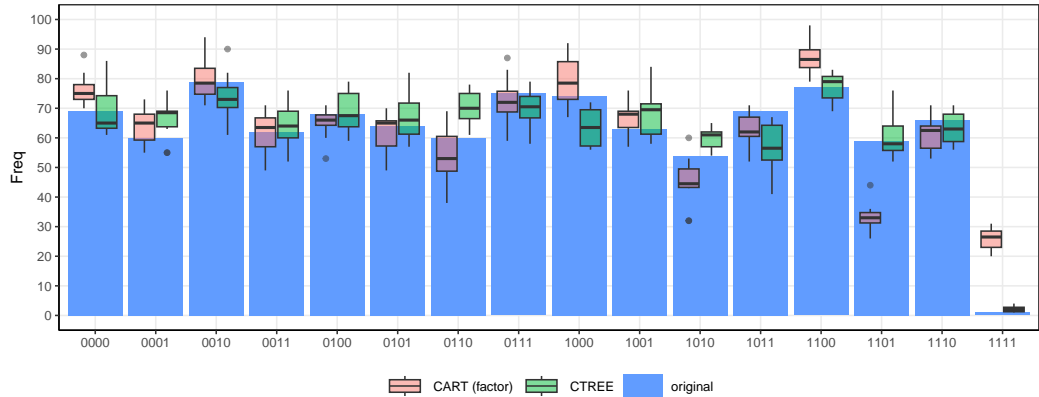
Figure 9: mb = 75

Figure 10: cp = 0.05
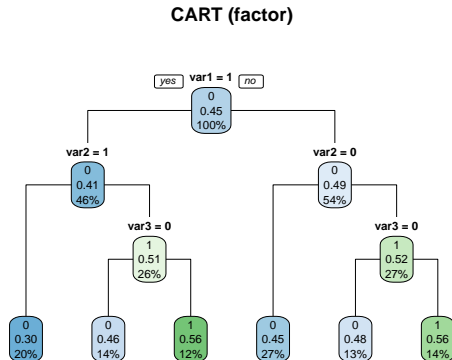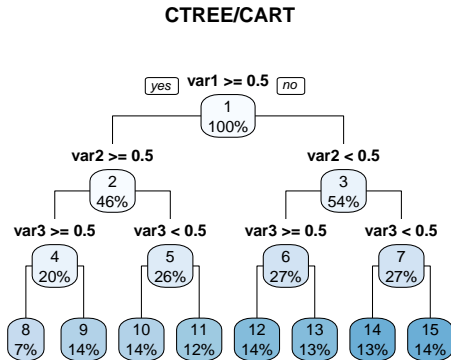
Figure 11: Datasynthesizer with DP

OTHER OPTIONS: CART (FACTOR) VS. CTREE

Figure 12

EXPLAINING DIFFERENCE BETWEEN CART (FACTOR) VS. CTREE

Figure 13

SUMMARY

- If we are willing to sacrifice utility, then we can reduce disclosure risk in CART-based SDG

SECTION 5: IS THIS SCENARIO REALISTIC?

REAL WORLD DATA (SD2011)

Following the authors of Synthpop (Raab, 2024; Raab et al., 2024), we rely on data from Social Diagnosis 2011 (SD2011).

In their paper, they generate 5 synthetic data sets to illustrate their method for measuring attribute disclosure by identifying values in the target variable `depress` from keys: `sex age region placesize`.

To illustrate why it is a problem to measure attribute disclosure as the set of records with constant t within q, we set t as constant for all observations in all 5 synthetic data sets. 0 was chosen because it is the most frequent value in the variable `depress` (22% of all records). By definition, this reduces attribute disclosure risk.

In their example, attribute risk is about 9%. However, when we modify `depress` so that it is constant (0), the risk *increased* to around 15%.

Therefore, even though we know risk declined (because we reduced it), *DiSCO* increases.

Table 5: SD2011

Table 6: Attribute disclosure measures for `depress` from keys: `sex age region placesize`

Data	Identity risk		Attribute risk	
	Original	Modified	Original	Modified
Original	0.00	0.00	0.00	0.00
Synthetic 1	14.82	14.82	8.96	14.74
Synthetic 2	14.20	14.20	9.90	14.82
Synthetic 3	15.16	15.16	10.46	14.94
Synthetic 4	14.12	14.12	9.68	14.50
Synthetic 5	14.30	14.30	8.88	14.66
Average	14.52	14.52	9.58	14.73

Note: Modified indicates that values of `depress`=0 in synthetic data

IDENTIFYING DISCLOSURE FROM 1-WAY

The package authors are aware that the *DiSCO* measure of attribute disclosure risk can indicate a high level of risk for a target variable where a high proportion of records have one level (Raab et al., 2024).

The package includes a flag to allow the user to identify values within a variable that explain most of the disclosures (`check_1way`).

The authors give an example where the target variable is `workab`, where 89% of the observations never worked abroad.

The authors suggest that this level of *t* for a group with the same *q* would not be disclosive.

We agree, but our example illustrates that the disclosure measure increases, when it should decrease.

SUMMARY

- When we create synthetic data to reduces attribute disclosure risk, *DiSCO* measure increases

- The key point is that we show that *DiSCO* mismeasures risk using real world data

SECTION 6: 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

  – *DiSCO* indicates risk when there is no risk (simulation)

  – *DiSCO* indicates risk increases when we know it decreased (SD2011)

THANK YOU

Jonathan Latner: `jonathan.latner@iab.de`

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