INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency
AnigeD
Finanziert von der Europäischen Union
NextGenerationEU
Bundesministerium für Bildung und Forschung

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

UNECE Expert Meeting on Statistical Data Collection 2025,
Barcelona,
15-17. October, 2025

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

SECTION 1: INTRODUCTION

RESEARCH QUESTION AND APPROACH

- Research Question:
  - Do common privacy measures accurately capture disclosure risk in synthetic data generated by CART models?
- Context:
  - Synthetic data increasingly used to share data while preserving privacy.
  - CART-based SDGs: high statistical utility, relatively low privacy risk.
- Privacy Measures Evaluated:
  - Identity disclosure risk
  - Attribute disclosure risk
- Data:
  - Simulated dataset (Reiter design: 1,000 obs., 4 binary vars., unique case).
  - Public survey data: Social Diagnosis 2011 (SD2011).
- Contribution:
  - Assess validity of empirical disclosure risk measures and their implications for evaluating synthetic data generators.

SECTION 2: GENERATE THE ORIGINAL AND SYNTHETIC DATA

GENERATE ORIGINAL 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 SYNTHETIC DATA

- Synthpop

COMPARE ORIGINAL AND SYNTHETIC DATA

Figure 1: Frequency

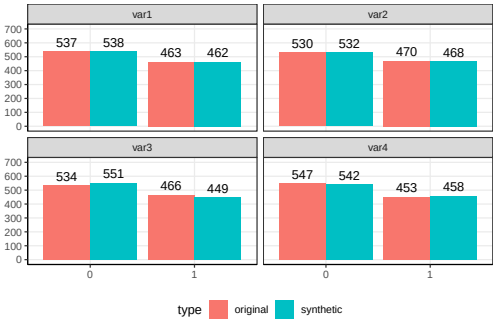
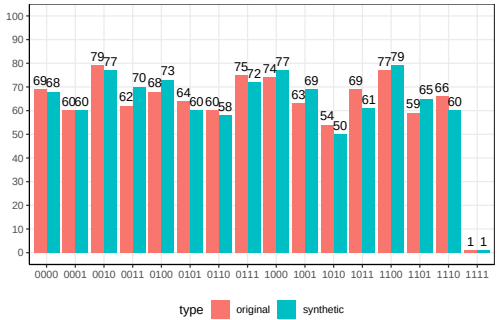
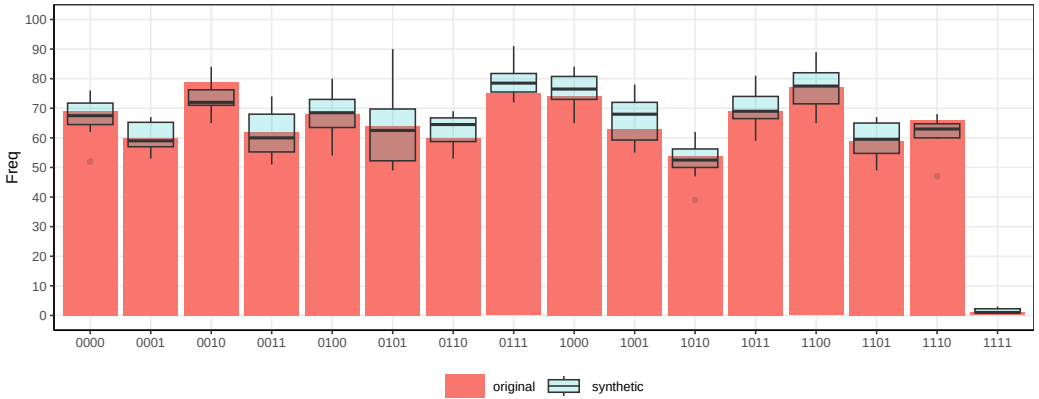


Figure 2: Histogram



COMPARE HISTOGRAM X 10 SYNTHETIC DATASETS

Figure 3: 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?

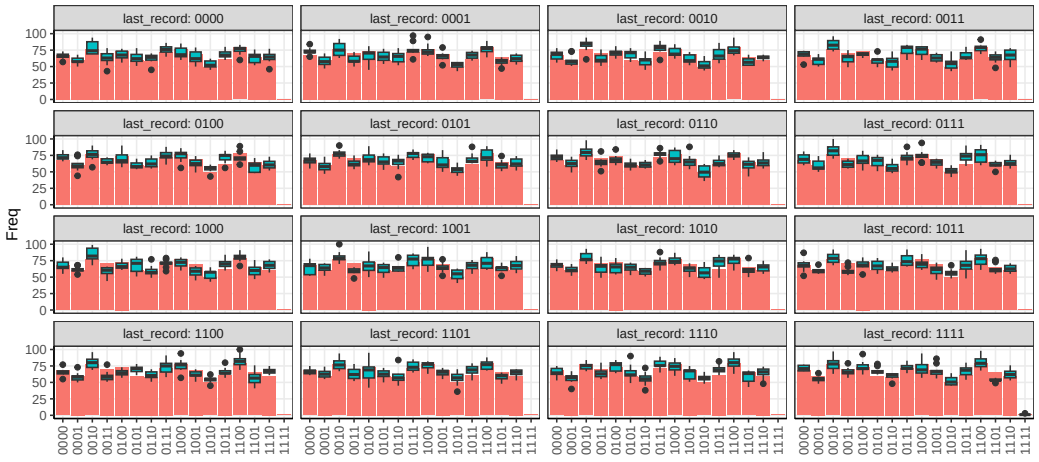
SECTION 3: THE ATTACK

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

HISTOGRAM OF 16 WORLDS X 10 SYNTHETIC DATASETS

Figure 4



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 4: MEASURING DISCLOSURE RISK

DISCLOSURE RISK MEASURES

- The literature on privacy measures for synthetic data is well-developed (Wagner and Eckhoff, 2018).
- Common privacy measures - Synthpop (Raab et al., 2025)
  - Identity risk (*repU*): the ability to identify individuals in the data from a set of known characteristics or 'keys' (*q*).
  - Attribute risk (*DiSCO*): the ability to find out from the keys something, not previously known or 'target' (*t*)
- Less common measures (Reiter et al., 2014)
  - Bayesian risk: the probability that an intruder assigns to the true record after seeing the synthetic data.
    - If this probability is close to the prior, little information is revealed.
    - If it is much higher, the intruder has gained information.
  - Less common because it is complicated to calculate, especially for more complex real-world data sets as the possible combination of values grows exponentially with the number of columns and possible values per column.

RESULTS DISCLOSURE RISK MEASURES

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

Data	Identity Risk (*repU*)	Attribute Risk (*DiSCO*)	Bayesian Estimate of Risk
Original	0.00	0.00	1.00
Synthetic	0.00	0.00	1.00

Table 2: x 10 synthetic data sets

Data	Identity Risk (*repU*)	Attribute Risk (*DiSCO*)	Bayesian Estimate of Risk
Original	0.00	0.00	1.00
Synthetic 1	0.00	0.00	1.00
Synthetic 2	0.00	6.60	0.02
Synthetic 3	0.00	0.00	1.00
Synthetic 4	0.00	0.00	1.00
Synthetic 5	0.00	0.00	1.00
Synthetic 6	0.00	0.00	1.00
Synthetic 7	0.00	0.00	1.00
Synthetic 8	0.00	6.60	0.03
Synthetic 9	0.00	0.00	1.00
Synthetic 10	0.00	0.00	1.00
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 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.

RESULTS

Table 3: Risk measures for `depress` from keys: `sex`, `age`, `region`, `placesize` (SD2011)

	Identity risk (*repU*)		Attribute risk (*DiSCO*)	
Data	Raab et al., 2024	Modified	Raab et al., 2024	Modified
Original data	48.38	48.38	53.30	53.30
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 for all records in the synthetic data

SUMMARY

- When we create synthetic data to reduce attribute disclosure risk, *DiSCO* measure increases
- The package authors are aware of the problem
  - 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`).
- We agree, but our example illustrates that the disclosure measure increases, when it should decrease.
- The key point is that we show that *DiSCO* mismeasures risk using real world data

SECTION 6: CONCLUSION

SUMMARY

- Common privacy metrics may fail to detect or even misstate disclosure risk.
- CART-based synthetic data generators reproduce original data with high utility, but offer little protection for disclosive records under default settings.
- Adjusting parameters (adding noise) can reduce disclosure risk, but at the cost of lower utility.
- Key takeaway: users must understand both how SDGs generate data and how privacy measures operate—there is no one-size-fits-all solution.

THANK YOU

Jonathan Latner: jonathan.latner@iab.de

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