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

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 AND OVERVIEW

OVERVIEW

- Background:
  - Synthetic data are increasingly used to share data while preserving privacy.
  - Numerous synthetic data generators (SDGs) using variety of methods
  - CART-based SDGs: high statistical utility with high privacy protection (Little et al., 2025)
- Research question:
  - If that is true, how would we know?
  - Do common privacy measures capture disclosure risk in synthetic data generated by CART models?
- Evaluate 3 privacy measures:
  1. Identity disclosure risk
  2. Attribute disclosure risk
  3. Bayesian estimation of disclosure risk
- 2 Data:
  1. Simulated dataset (Reiter et al., 2014 design: 1,000 obs., 4 binary vars., unique case).
  2. Public survey data: Social Diagnosis 2011 (SD2011).
- Contributions:
  - We show that CART-based models may produce synthetic data that sacrifices privacy protection for statistical utility.
  - Commonly used disclosure risk measures may not capture disclosure risk.
  - We propose some solutions for measuring disclosure risk (Bayesian).
  - More generally, users interested in measuring privacy risk should be aware of the challenges we describe here.

SECTION 2: GENERATE SIMULATED DATA (ORIGINAL AND SYNTHETIC)

ORIGINAL DATA SET: SIMULATED DATA

- Borrowing from Reiter et al. (2014), we create a data set with $n = 1000$ and 4 dichotomous, categorical variables.
- The first 999 observations are a random sample from 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 ($var1 = 1, var2 = 1, var3 = 1, var4 = 1$).
- The value of the simulated data is that we know there is a unique record because we created it.

SYNTHETIC DATA SET

- Generate 1 synthetic data set from a CART-based SDG using the Synthpop package in R
  - We use the default settings and hyperparameter values and set a seed=1237.
- As a sensitivity test, we create 10 synthetic data sets from the original simulated data.

COMPARE ORIGINAL AND 1 SYNTHETIC DATA COPY (SEED = 1237)

Figure 1: Frequency

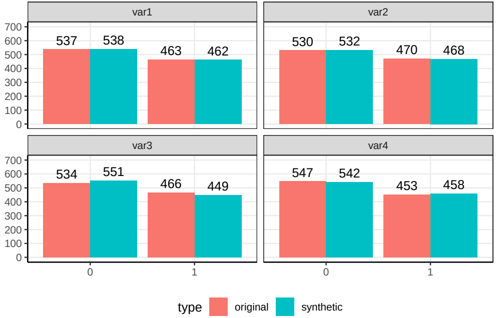
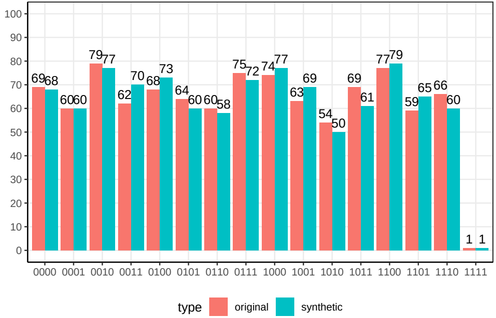
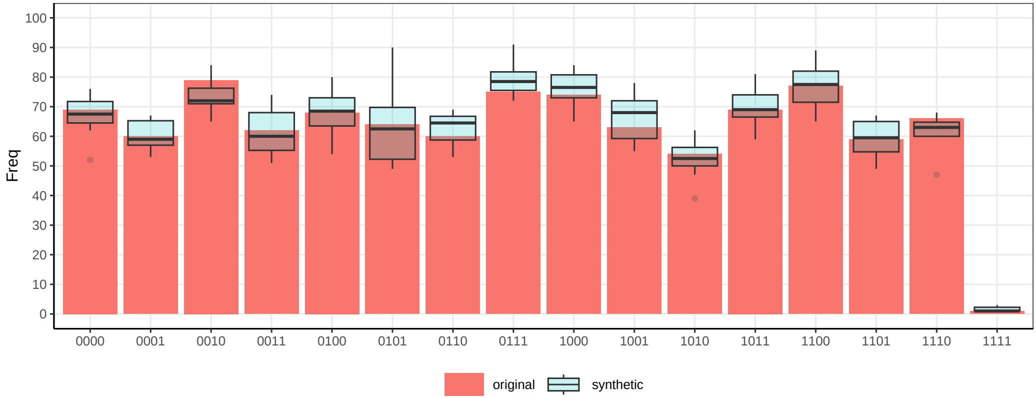


Figure 2: Histogram



COMPARE HISTOGRAM X 10 SYNTHETIC DATASETS

Figure 3: Multiple synthetic data sets does not reduce privacy risk



SUMMARY

- The problem: Synthetic data from CART models are disclosive in this simulation
- Next section: Can an attacker identify the disclosure?

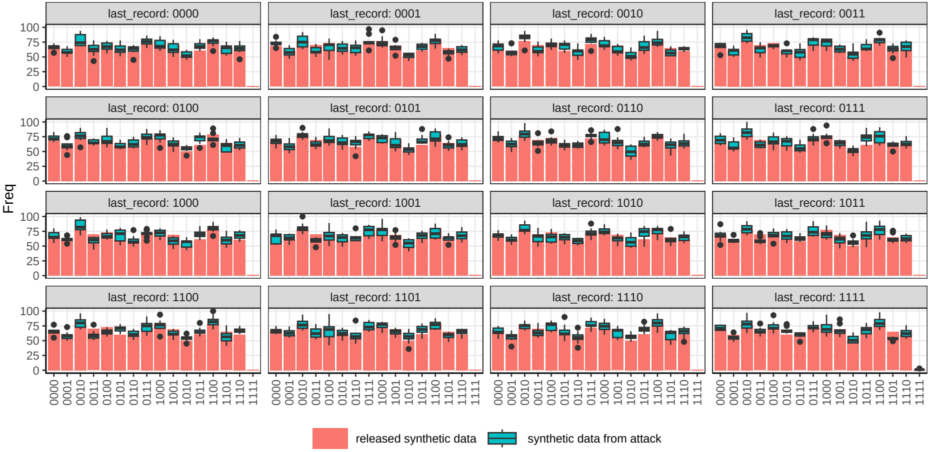
SECTION 3: THE ATTACK

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 original 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
- Next section: Can we measure this disclosure risk?

SECTION 4: MEASURING DISCLOSURE RISK

THREE DISCLOSURE RISK MEASURES

- 2x Common disclosure risk measures reflect the current state of the art (Raab et al., 2025)
  - Identity risk (*repU*): the ability to identify individuals in the data from a set of known characteristics or ‘keys’ (q).
    - $q = var1(0, 1), var2(0, 1), var3(0, 1)$
    - This should be 0 because these three attributes yield $2^3 = 8$ possible combinations, none of which are unique in the dataset
  - Attribute risk (*DiSCO*): the ability to find out from the keys (q) something, not previously known or ‘target’ (t)
    - $t = var4(0, 1)$
    - This should be > 0 because when $q = 111$, there is a unique record if $t = 1$.
- 1x Alternative disclosure risk measure
  - Bayesian approach (Reiter et al., 2014)
    - If posterior probability is close to the prior (e.g., uniform distribution), little or no new information is revealed.
    - If posterior probability is substantially larger, the intruder has learned something about the last or unique record.
  - In our data this should be > 0, i.e. positive

RESULTS DISCLOSURE RISK MEASURES

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

Data	Unique	Identity Risk (*repU*)	Attribute Risk (*DiSCO*)	Bayesian Estimate of Risk
Original	1	0.00	0.00	1.00
Synthetic	1	0.00	0.00	1.00

Table 2: x 10 synthetic data sets

Data	Unique	Identity Risk (*repU*)	Attribute Risk (*DiSCO*)	Bayesian Estimate of Risk
Original	1	0.00	0.00	1.00
Synthetic 1	1	0.00	0.00	1.00
Synthetic 2	0	0.00	6.60	0.02
Synthetic 3	1	0.00	0.00	1.00
Synthetic 4	3	0.00	0.00	1.00
Synthetic 5	2	0.00	0.00	1.00
Synthetic 6	1	0.00	0.00	1.00
Synthetic 7	3	0.00	0.00	1.00
Synthetic 8	0	0.00	6.60	0.03
Synthetic 9	1	0.00	0.00	1.00
Synthetic 10	1	0.00	0.00	1.00
Average	-	0.00	1.32	-

SUMMARY

- According to 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)
- Only Bayesian approach captures disclosure risk and uncertainty about risk

SECTION 5: IS THIS SCENARIO REALISTIC?

REAL WORLD DATA (SD2011)

- Replicate the approach the authors of Synthpop (Raab, 2024; Raab et al., 2024)
- Original are from Social Diagnosis 2011 (SD2011).
- Measure disclosure risk
  - 4 keys (*q*): `sex` `age` `region` `placesize`.
  - 1 target (*t*): `depress`
- Generate 5 synthetic copies with a ‘bad’ synthesizer
  - We set *t* as constant for all observations in all 5 synthetic data sets.
  - Therefore, we know risk declined (because we reduced it).
- Do common disclosure risk measures (*repU*, *DiSCO*) capture this decline?
  - Why not Bayesian approach? High-dimensional, real data is too computationally complex. Only good for low-dimensional data

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).
- This is good, but our example illustrates a more general problem: *DiSCO* may mismeasure risk by indicating it is rising, when it declined

SECTION 6: CONCLUSION

SUMMARY

- CART-based synthetic data generators may provide high levels of utility without reducing privacy risk
- Common privacy metrics may fail to detect or even misstate disclosure risk.
- Bayesian approach can be a good solution, but only in low-dimensional data
- 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