Buyer Beware: Understanding the trade-off between utility and risk in CART based models using simulation data

Jonathan Latner$^{1[0000-0002-1825-0097]}$, Marcel Neunhoeffer$^{1,2[0000-0002-9137-5785]}$, and Jörg Drechsler$^{1,2,3[0009-0009-5790-3394]}$

[1] Institute for Employment Research, Nuremberg, Germany {jonathan.latner, marcel.neunhoeffer,joerg.drechsler}@iab.de
[2] Ludwig-Maximilians-Universität, Munich, Germany
[3] University of Maryland, College Park, USA

Abstract. This paper evaluates identity risk and attribute risk in synthetic data generated by CART-based models, using both a controlled simulated dataset and publicly available data. We find that common privacy metrics may fail to detect disclosure risks and, in some cases, misrepresent actual privacy threats. Additionally, CART-based models, while maintaining high statistical utility, may compromise privacy protection. Our findings highlight challenges in measuring privacy risk in synthetic data and suggest improvements for more accurate risk assessment.

Keywords: Synthetic data · Privacy · CART · synthpop

1 Introduction

The generation of synthetic data has gained prominence as a means to share data while preserving privacy. There are numerous methods and tools, both for generating synthetic data and measuring privacy. On the one hand, there is a general perception that generating synthetic data are easy [3], and to a certain degree this is true. On the other hand, it is not always clear if the resulting synthetic data are in fact providing privacy protection. Our question of interest asks to what extent do commonly used privacy measures accurately capture disclosure risk in synthetic data generated by CART-based models?

In this paper, we evaluate standard and easily available privacy measures to estimate risk from synthetic data generated by CART based models. Specifically, we use two commonly understood measures of privacy: identity risk and attribute risk [8]. And we use CART because previous research indicates that CART-based synthetic data generators (SDGs) generate synthetic data with both high statistical utility and relatively low privacy risks compared to other methods [4,2,1]. While the privacy measures and CART are implemented using the Synthpop package [6] in R, our findings are more generally applicable to both risk measures and CART models.

We use two types of data. One is a simulated dataset with 1.000 observations and four binary variables [9]. Crucially, the last or 1.000^{th} observation is a unique combination of the four binary variables. In so doing, we create data with an observation we want to protect using synthetic data generated from a CART model. We also use publicly available data from Social Diagnosis 2011 (SD2011) to extend our analysis beyond a simulated setting and into a real-world setting.

Our approach follows a structured framework. Using the simulated data and drawing on concepts from the attack model in differential privacy, we estimate risk measures for an attacker with full knowledge of the synthetic data generator, all synthetic observations, and all original observations except the disclosive record. First, we test whether an attacker can identify the disclosive record. They can. Next, we test whether the privacy measures detect a known disclosure risk in the simulated data. They do not. Finally, we extend the analysis using the SD2011 data to assess whether the measurement problem is specific to the simulated case or part of a broader challenge. The results suggest the latter.

We make three contributions. Common privacy metrics both may not capture disclosure risk in synthetic data generated from simulated data, and also may misstate privacy risks in real data. Further, relatedly, and in contrast to previous research, CART-based models may produce synthetic data that sacrifices privacy protection for statistical utility. Finally, we propose some solutions for measuring disclosure risk. More generally, users interested in measuring privacy risk should be aware of the challenges we describe here.

2 Privacy measures

The literature on privacy measures for synthetic data is well-developed [10]. One reason why there are so many measures of privacy is because there is no one agreed upon understanding of either what defines risk nor how one should measure it. We use two commonly understood measures of privacy implemented by the Synthpop package in R (version 1.9-0) [8]: identity risk and attribute risk.

Identity risk measures the ability to identify individuals in the data from a set of known characteristics, i.e. 'keys'. The maximum number of keys are one less than the total number of variables in the data. Here, the keys are the first 3 binary variables (q), but this choice is arbitrary as all variables are binary.

The following steps are used to calculate identity risk. For a given set of keys (q), the intruder will look for the unique records in the original data (UiO) that are also unique in the synthetic data (UiS). $repU$ (replicated uniques) is the measure of identity risk and defined by equation 1:

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

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. $s_{.q} = 1$ 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 risk measures the ability to identify a previously unknown characteristic of an individual. In this approach, an attacker has access to synthetic data, knows one or more identifiers in the original data (i.e. q or keys, as in above), and wants to infer a sensitive attribute (t). Attribute risk or $DiSCO$ (Disclosive in Synthetic Correct in Original) is the subset of records in the original data for which the keys (q) in the synthetic data are 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.

3 Data

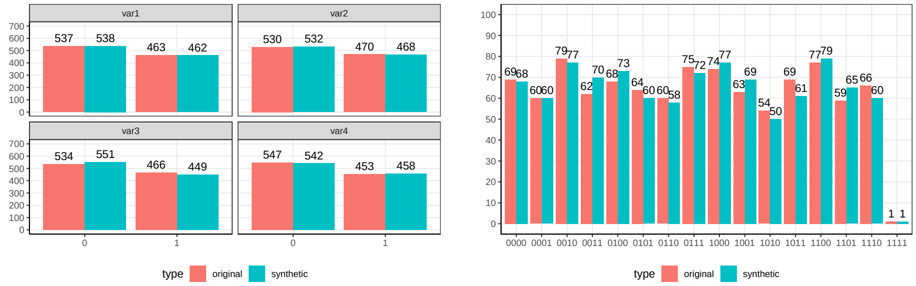
To examine our question of interest, we use two types of data: simulated and real data. Following Reiter et al. [9], we simulate one data set with 1.000 observations and four binary categorical variables. This is our 'original' data set. The first 999 records are sampled from a multinomial distribution for all combinations of var1(0,1), var2(0,1), var3(0,1), var4(0,1), except the last 1000^{th} record is a unique combination (var1 = 1, var2 = 1, var3 = 1, var4 = 1). Next, we generate one synthetic data set from a CART-based SDG using the Synthpop package in R with default parameters (seed=1237). As a sensitivity test, we create 10 synthetic data sets from the original data. The value of the simulated data is that we know there is a disclosive record because we created it.

At the same time, an obvious critique is that the simulated data set is not representative of real world data. We agree with this critique, even though the value of the simulated data is to illustrate a sort of 'bound' on disclosure risks. Therefore, we use a second data set from Social Diagnosis 2011 (SD2011), publicly available from the Synthpop package in R to examine our question of interest using real world data.

4 Results from simulated data

We begin with the simulated data that contains four categorical variables and 1.000 observations, one of which is unique or disclosive. Figure 1a shows the frequency distribution within each of the four variables and figure 1b the frequency histogram across all four variables. In the original data, there is one observation with a combination (1,1,1,1) that is not visible if we look at the distribution within each of the variables.

Fig. 1: Compare original and synthetic data



(a) Frequency

(b) Histogram

On the one hand, the synthetic data generated by the CART model have high levels of utility because they almost perfectly match the frequency of values in the original data not only within the four variables but also across all four variables. On the other hand, the synthetic data perfectly replicates the single disclosive record.

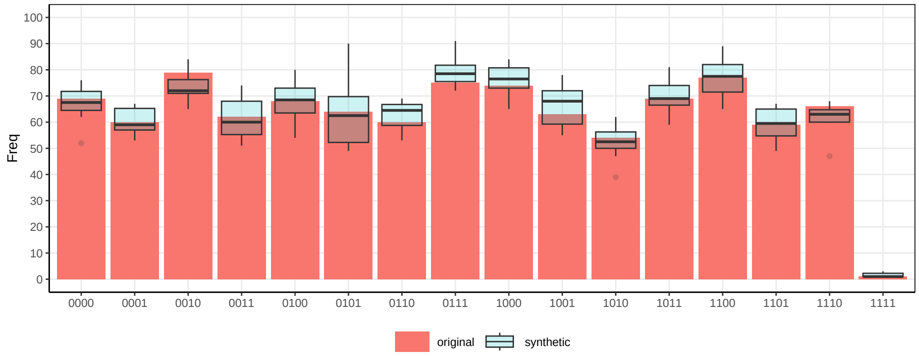
The disclosive record does not disappear if we generate multiple synthetic data copies (10), as shown in figure 2. The frequency of the disclosive record ranges from 0 (2 data sets), 1 (5 data sets), 2 (1 data sets), and 3 (2 data sets).[4] Regardless of whether one, five, or ten synthetic data sets were released, it would be clear which record was the disclosive record. As a result, synthetic data from CART models do not protect the unique observation in our simulated data set.

4.1 The attack

One way to think about disclosure risk is to imagine a game between two entities. On one side, there is a statistical agency who has the data and wants to release it in a privacy preserving way. On the other side, there is an attacker who wants to identify someone in the data (either membership or attribute inference). The

[4] For reference, the frequency of the disclosive record would be similar if we created 100 synthetic data sets, ranging from 0 (41 data sets), 1 (38 data sets), 2 (14 data sets), and 3 (7 data sets).

Fig. 2: Frequency



question is what can the attacker learn from a released synthetic data set about an individual they do not have knowledge of?

In this scenario, we assume a ‘strong’ attacker similar to the attack model in differential privacy (DP). In so doing, we assume that the attacker knows the SDG used to generate the synthetic data. In our case, this is sequential CART. They know all observations except the last one. In addition, given the nature of the data, they know all 16 possible combinations that the last record could be. In this attack, the attacker sees the synthetic data and then runs the same CART-based SDG for each of the 16 different possibilities, sequentially. Then, they update their beliefs about what the last record could be.

Figure 3 illustrates the results of the attack. In the top left cell, the attacker guesses that the last record in the original data is 0,0,0,0. They then generate 10 synthetic data sets using a CART-based SDG and compare the histogram to the released synthetic data, as shown in figure 2.

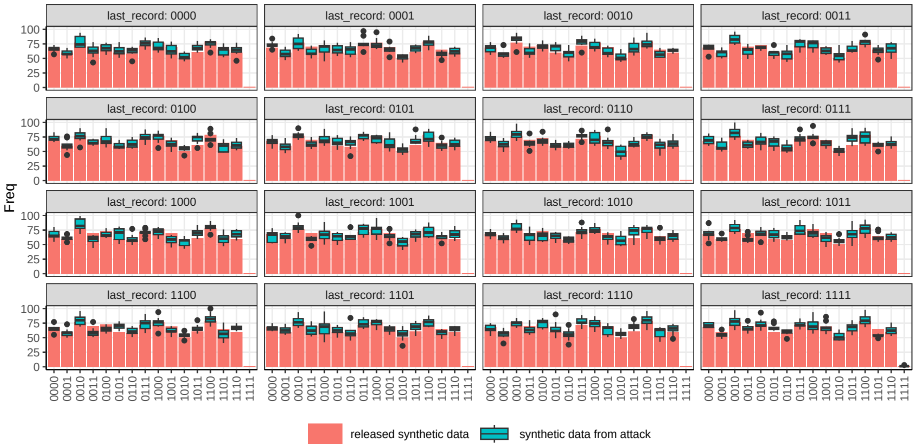
If the attacker guesses that the last record is 0,0,0,0, then they are not able to replicate the single unique record in the synthetic data. Next, they update their beliefs about the last record and guess that the last record in the original data is 0,0,0,1. They then repeat the process as described above. This is shown in the top row, second column from the left. Like their first guess, they cannot replicate the released synthetic data.

The attacker then repeats this process for all 16 possible combinations of the last record. Finally, if they guess that the last record is 1,1,1,1, then they are able to replicate the released synthetic data, as shown in the bottom, right cell. The result is a successful attack with confirmation that the guess about the values of the last, unique observation is correct.

4.2 Measuring privacy risk

The larger concern is whether we are able to measure this disclosure. Results from our simulated data show that CART produces synthetic data that is dis-

Fig. 3: Histogram of 16 worlds x 10 synthetic datasets



closive because it replicates unique records from the original data set without adding sufficient noise. Other research also indicates that CART-based SDGs can produce synthetic data with high levels of utility because they reproduce a high proportion of the original data [5]. The question is can we measure this observable risk?

In table 1, the columns display identity and attribute risk measures (the columns) in the original and synthetic data (the rows). For reference, we replicated table 1 with 10 synthetic copies from figure 2, as shown in table A.2 in the Appendix. Results are qualitatively similar.

Table 1: Disclosure risk measures

data	identity	attribute
Original	0.00	0.00
Synthetic	0.00	0.00

The identity risk measures are 0 for both the original and synthetic data. This is correct because we know that there are multiple combinations of $var1 = (0, 1)$, $var2 = (0, 1)$, $var3 = (0, 1)$. Therefore, the attribute risk correctly identifies that there is zero risk of identity disclosure because there is no unique combination of observations with the three keys.

However, the attribute risk measures are also 0 for both the original and synthetic data. This is a problem because we know that when $q = 111$, there is a unique record if $t = 1$.

How can it be that there is no attribute disclosure risk when we know there is an attribute disclosure risk? The answer is that there is only an attribute disclosure risk when t is constant, when there is no variation within q. As a result, there is only an attribute disclosure risk when there are 0 copies of the unique record in the synthetic data.

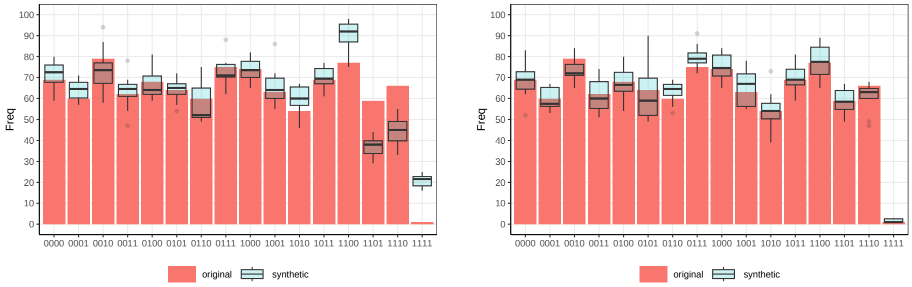
We can see this issue more clearly if we examine the frequency table from 10 synthetic data copies, as shown in figure 2. The underlying data and resulting privacy measures are shown in the appendix (table A.1 and A.2). If there is at least 1 unique record in the synthetic data, then there is no attribute risk because there are 2 values of t within q. At the same time, if a synthetic data set is released without the unique record, then there is an attribute disclosure risk because there is only 1 value of t within q.

Therefore, the attribute disclosure risk measure indicates a disclosure risk when we know there is no disclosure risk, and indicates there is no disclosure risk when we know there is a disclosure risk. The exercise illustrates a flaw in the measure of attribute disclosure risk using our simulated data. We know there is a disclosure problem (because we created it), but the problem is not identified by the measure.

4.3 Decreasing privacy risk

We can correct the problem of disclosure risk in synthetic data generated from CART-based SDGs. If we modify the parameters to avoid overfitting, then the disclosure problem disappears. There are multiple options to reduce overfitting in CART-based SDGs. We use two: increase the minimum number of observations per terminal node to 75 (default is 5) and increase the complexity parameter to 0.05 (default is 1e^{-8}), as shown in figure 4. We arrived at these values after doing sensitivity tests, as shown in figure A.1. Although values below these do not add enough noise to the data to reduce disclosure risks, these adjustments can reduce disclosure risks.

Fig. 4: Compare original and synthetic data



(a) Minimum bucket is 75

(b) Complexity parameter is 0.05

At the same time, if one wants to increase privacy, then one must sacrifice utility. Even with 10 synthetic data sets, figure 4 is less representative of the original data set than figure 2. Second, in an important way, we are voluntarily choosing to decrease utility in order to increase privacy, but remember that no disclosure risk measure indicates that there is a problem. It would be unusual to voluntarily reduce utility if there is no reason to suspect a disclosure risk, but it is important to know that options exist to increase privacy risk if there is a reason to suspect a disclosure risk.

5 A real world example

In the papers describing privacy measures implemented in Synthpop [7,8], the author(s) 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`. As described above, their preferred measure of attribute disclosure risk (DiSCO) is the set of records in the synthetic data with a constant target (t) for the a set of keys (q). In other words, there is no variation of t within q. In their example, attribute risk in the original data is 53%, which is reduced to about about 9% in the synthetic data (as shown in the table A.3 appendix).

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. However, according to the measure of attribute disclosure risk used by the package, the risk increased to around 15%.

We note that this problem is already understood and described by the package authors [8]. They provide a parameter where a user may check for a target where a high proportion of records have one level (`check_1way`). However, the problem is not the package, the problem is the definition.

The exercise illustrates a flaw in the measure of attribute disclosure risk using real world data. We modified the synthetic data by setting a uniform value for the target variable: `depress` $= 0$. If the measure correctly measured risk, then the risk measure should decline, but it rises. Therefore, the attribute risk measure indicates that disclosure risk increased when we know disclosure risk decreased (because we decreased it).

6 Conclusion

In this study, we use two data sets, one simulated data with a single disclosive record and one real-world data, to demonstrate three ideas. First, not only do common privacy metrics not detect the disclosure risks we know exist in the simulated data, but they can also misstate the disclosure risk in real-world data. Second, relatedly, CART-based synthetic data generators with default parameters create synthetic data with high levels of utility that reproduce the original data without protection for the dislosive record. Therefore, CART-based models are not inherently immune to the utility-privacy trade-off. Finally, it is possible to increase protection by adding noise to the synthetic data with simple adjustments to the default parameters, but the cost is to reduce utility. The question is why one would reduce utility if there is no indication there was a disclosure problem? Given these results, it is important for users interested in reducing disclosure risk to better understand not only how SDGs generate synthetic data, but also how common privacy measures work. There is no one size fits all solution.

Acknowledgement. This work was supported by a grant from the German Federal Ministry of Education and Research (BMAS grant number 16KISA096) with funding from the European Union—NextGenerationEU. Reproducible files are located here: https://github.com/jonlatner/KEM_GAN/tree/main/latner/projects/simulation

Disclosure of Interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dankar, F.K., Ibrahim, M.: Fake it till you make it: Guidelines for effective synthetic data generation. Applied Sciences **11**(5), 21–58 (2021)
2. Fössing, E., Drechsler, J.: An evaluation of synthetic data generators implemented in the python library synthcity. In: International Conference on Privacy in Statistical Databases. pp. 178–193. Springer (2024)
3. Latner, J., Neunhoeffer, M., Drechsler, J.: Generating synthetic data is complicated: know your data and know your generator. In: International Conference on Privacy in Statistical Databases. pp. 115–128. Springer (2024)
4. Little, C., Allmendinger, R., Elliot, M.: Synthetic census microdata generation: A comparative study of synthesis methods examining the trade-off between disclosure risk and utility. Journal of Official Statistics p. 0282423X241266523 (2024)
5. Manrique-Vallier, D., Hu, J.: Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. Journal of the Royal Statistical Society Series A: Statistics in Society **181**(3), 635–647 (2018)
6. Nowok, B., Raab, G.M., Dibben, C.: synthpop: Bespoke creation of synthetic data in r. Journal of statistical software **74**, 1–26 (2016)
7. Raab, G.M.: Privacy risk from synthetic data: practical proposals. In: International Conference on Privacy in Statistical Databases. pp. 254–273. Springer (2024)
8. Raab, G.M., Nowok, B., Dibben, C.: Practical privacy metrics for synthetic data. arXiv preprint arXiv:2406.16826 (2024)
9. Reiter, J.P., Wang, Q., Zhang, B.: Bayesian estimation of disclosure risks for multiply imputed, synthetic data. Journal of Privacy and Confidentiality **6**(1) (2014)
10. Wagner, I., Eckhoff, D.: Technical privacy metrics: a systematic survey. ACM Computing Surveys (Csur) **51**(3), 1–38 (2018)

A Appendix

Table A.1: Frequency statistics for original and synthetic data

	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 A.2: 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

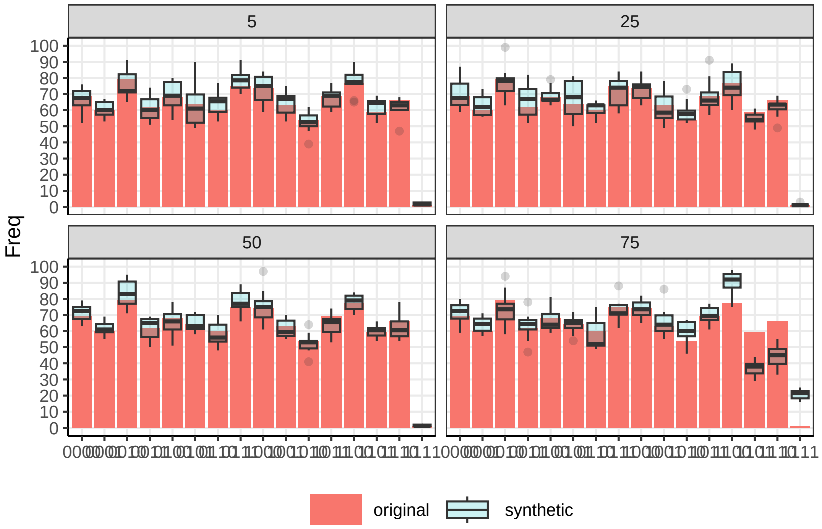
Table A.3: SD2011

Table A.4: Attribute disclosure measures for `depress` from keys: `sex` `age` `region` `placesize`

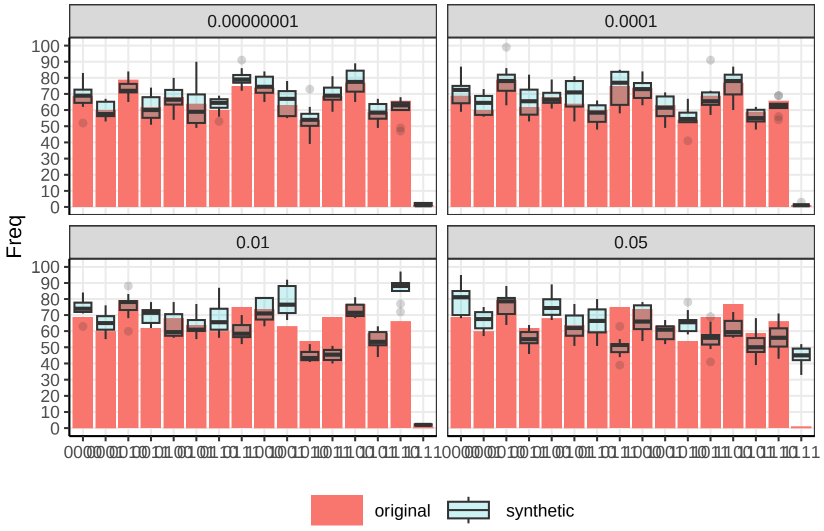
	Identity risk		Attribute risk	
Data	Raab et al., 2024	Modified	Raab et al., 2024	Modified
Original data	48.38	48.38	53.30	53.30
Synthetic 1	14.82	14.82	8.96	14.74
Synthetic 2	14.20	14.20	9.90	14.82
Synthetic 3	15.16	15.16	10.46	14.94
Synthetic 4	14.12	14.12	9.68	14.50
Synthetic 5	14.30	14.30	8.88	14.66
Average	14.52	14.52	9.58	14.73

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

Fig. A.1: Compare original and synthetic data with different hyperparameters



(a) Minimum bucket (default is 5)



(b) Complexity parameter (default is 10^{-8})