Anonymeter Application to CRC Diverse Communities Excerpts: A Privacy Perspective

Matteo Giomi[1]*, Nicola Vitacolonna[1], Omar Ali Fdal[1]

Abstract
Synthetic data is a practical approach to privacy-preserving information sharing, effectively reproducing the statistical properties of original datasets while reducing the exposure of individual records. While it is acknowledged that entirely eliminating privacy risks is challenging, synthetic data can help minimize those risks. However, not all synthetic data generators offer the same level of protection for all datasets. The residual risks need to be assessed on a case-by-case basis. Anonymeter is a useful framework in this context, specifically designed to assess and quantify privacy risks in synthetic data, in line with key GDPR indicators. In this study, we use Anonymeter to evaluate hundreds of datasets from the CRC Diverse Communities Excerpts. Our work uncovers various degrees of residual privacy risks across different synthetic datasets and synthetization algorithms. We also demonstrate correlations between Anonymeter's empirical risks and other measures of privacy and utility. Our results highlight the diversity of de-identified datasets submitted to the CRC program, and demonstrate Anonymeter's usefulness in carrying out large-scale privacy analysis.

Keywords
Anonymeter, Privacy Risk, Synthetic Data, SDNIST, GDPR

[1] *Anonos Inc.*
***Corresponding author**: matteo.giomi@anonos.com

1. Introduction

Synthetic data generation is an increasingly popular technique for preserving privacy in data sharing. However, just because the data is synthetic, it does not mean that it is without privacy risks [1, 2]. Understanding such risks is difficult as many factors come into play when it comes to data synthesis, and de-identification in general: the nature of the data, the anonymization algorithms and their specific parameters affect the residual risks of the protected datasets in complex ways. Besides, the achievable privacy level is constrained of course by the desired utility of the output for the downstream use-cases.

The NIST Collaborative Research Cycle (CRC [3]) represents a unique research opportunity by providing several curated datasets to be used as a common benchmark. Such datasets are subsets of a publicly available US Census Bureau dataset — the 2019 American Community Survey (ACS) Public Use Microdata Sample (PUMS) — called the *NIST Diverse Communities Data Excerpts* [4]. In particular, such subsets contain Public Use Microdata Areas (PUMAs) with challenging distributions, which can be used to expose interesting behaviour of synthetic data generation algorithms. As part of the first iteration of the CRC, the NIST collected over 350 recently de-identified versions of these datasets. Such de-identified versions may arguably be considered as representative of the state-of-the-art in the context of synthetic data. The NIST benchmark also provides extensive utility evaluation reports, which we use to analyze the privacy-utility trade-offs of the evaluated datasets.

Different ways of quantifying privacy risks have been proposed [5, 2, 1, 6]. In this study, we will use Anonymeter [1] to explore and quantify privacy risks of the synthetic datasets submitted in version 1.1 of the CRC. Anonymeter is an open source statistical framework[1] that evaluates the anonymity of synthetic tabular datasets by jointly quantifying different types of privacy risks. It provides legally aligned risk measures, and has been positively evaluated by the experts of the Commission Nationale de l'Informatique et des Libertés (CNIL [7]) [8].

2. Methods

In this section, we summarize the main concepts about privacy evaluations with Anonymeter and then describe the setup of our experiments.

2.1 Anonymeter privacy evaluation

Anonymeter empirically evaluates privacy risks using an adversarial methodology [1]. It is equipped with attack-based evaluations for singling out, linkability, and inference risks, which are originally based on the opinions of the Article 29 Data Protection Working Party [9] and commonly accepted as the three key indicators of factual anonymization and the three main threats to data privacy. Each risk evaluation is modeled

[1] https://github.com/statice/anonymeter

as an attacker in possession of the full synthetic dataset and, optionally, of some auxiliary information coming from the original dataset, and is given the task to come up with a set of guesses of the form:

- *singling out*: there is only one person with attributes X, Y, and Z.
- *linkability*: records A and B belong to the same person.
- *inference*: a person with attributes X and Y also has Z.

The risk evaluation consists of three phases: first, the attack phase, where a large number of attacks (N_{attacks}) are performed, each consisting of a guess about a record in the target population. Then, in the evaluation phase, the success rates of the attacks are computed by comparing the guesses with the true information about the targets. Finally, in the risk quantification phase, a risk value and its confidence interval are derived.

During the attack phase, three kinds of attacks are carried out. The *main* privacy attack is the one in which the attacker uses the synthetic data to guess information on target records from the original data. Then, a *control* attack is also carried out, in which the target records are selected from a *control* dataset, that is, a sample of records from the same distribution as the original data, which were not used to generate the synthetic data. Finally, a *naive* attack which does not use the synthetic data but instead guesses at random is also performed, to provide a baseline against which the strength of the main attack is measured.

The success rate of the *main* attack (r_{main}) is a measure of the amount of correct information the synthetic data leaks. Crucially, it does not differentiate between general information about populations in the dataset and information specific to particular individuals. The former is fundamental to the utility of the synthetic data, and only the latter constitutes a privacy concern. By comparing the success rate of the main attack with that of the control attack, the two contributions can be discriminated. By construction, the synthetic data does not contain any private information about records in the control set since the latter was not used in the synthesis. Thus, the success rate of the *control* attack (r_{control}) measures general population-level information contained in the synthetic data. If $r_{\text{main}} \simeq r_{\text{control}}$ then the attacker's information gain about the original data is negligible: the correctness of the guesses can be explained by population-level utility. If, however, $r_{\text{main}} > r_{\text{control}}$ then specific information about original records was leaked during the synthesis. The privacy risks reported by Anonymeter are defined as $(r_{\text{main}} - r_{\text{control}})/(1 - r_{\text{control}})$: the denominator is the maximum improvement over the control attack that a perfect attacker can obtain, and helps contextualizing the numerator.

2.2 Analysis parameters and datasets

Anonymeter was applied to the analysis of the de-identified datasets submitted as part of the version 1.1 of the CRC [10]. These datasets are generated from the three subsets of the 2019 ACS:

- *National*: 27254 records drawn from 20 PUMAs from across the United States;
- *Massachusetts*: 7634 records drawn from five PUMAs of communities from the North Shore to the west of the greater Boston, Massachusetts area;
- *Texas*: 9276 records drawn from six PUMAs of communities surrounding Dallas-Fort Worth, Texas area.

As control datasets, corresponding subsets coming from the 2018 ACS were used. Such datasets match the original datasets listed above, but they refer to the previous year. We restricted our analysis to submissions relative to three different feature sets: "all feature", "simple feature", and "demographic-focused" for a total of 167 datasets generated with 12 different synthetic data libraries. See Appendix A for details. We excluded from the analysis the submission coming from the "LostInTheNoise", and "Sarus SDG" libraries, as well as those from the "sub-sample" 1% and 5% approach, because they did not cover all the three datasets.

For a comprehensive estimation of the privacy risks, each submission was evaluated by modeling attackers with different strengths. Anonymeter allows the user to do so by specifying the amount of auxiliary information n_{aux} made available to the attacker. For the linkability and inference attacks, n_{aux} defines the number of attributes known to the attacker; for singling out, it represents the number of attributes used to isolate individuals in the target dataset. Increasing the auxiliary information generally results in stronger attacks. In our study, n_{aux} was varied from 1 to $n_{\text{aux}} - 1$ for inference, from 2 to n_{aux} for linkability, and from 1 to 10 columns for singling out, where n_{aux} was the number of columns in the considered dataset. The singling out analysis was not extended beyond ten columns, since the singling out attack does not perform well for larger values of n_{aux} (see Figure 7 in Appendix E of [1]). For each risk, we evaluated the attacks for 10 levels of auxiliary information. For each probed value of n_{aux}, $N_{\text{attacks}} = 1000$ attacks were performed. We found that this setting provided a good compromise between computation time and statistical accuracy. For additional details on the analysis parameters, see Appendix B. In total, across all submissions and analysis parameters, more than 32600 risk values were measured: $\approx$1600 for singling out, $\approx$6800 for linkability, and $\approx$24200 for inference. The entire analysis took a couple of days on a 32-cores instance. The code used to perform this analysis is publicly available[2].

3. Results

3.1 Impact of dataset and feature selection

The average risks across all the submissions for the three datasets are of the order of $\sim 5\%$. In particular, the average risks are 0.059, 0.054, and 0.051 for the Massachusetts, Texas, and National dataset, respectively. The standard error on the mean on these aggregates is 0.001 in all cases. Despite no dataset being distinctly worse than the others, the average risk

[2] https://github.com/AnonosDev/anonymeter-sdnist/tree/main

is highest for the Massachusetts dataset, suggesting that those records are harder to protect compared to the other datasets. To check whether that is a consequence of the size of the dataset – Massachusetts is the smallest dataset, with 7634 rows — samples of 7634 rows of each submission were taken and reevaluated. Although the absolute risks of the resized datasets were slightly reduced, the ranking was confirmed. See more details in Appendix C.

Figure 1 breaks down the aggregated risks by risk type across all submissions for each dataset. All datasets exhibit rather large values for the inference and singling out risks (6%–8%), while the linkability risk is significantly lower than those. This is consistent with the analysis presented in Giomi et al. [1], and indicates that the process of generating synthetic samples from a dataset indeed breaks the one-to-one mapping between synthetic records and the original ones, preventing linkage. The singling out risk seems to show a reversed trend compared to the average risks reported above, which highlights the importance of taking into account all the three risk types.

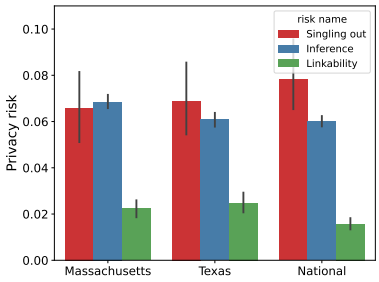

Figure 1. The average singling out, linkability, and inference risks for the different datasets. Throughout this work, and unless otherwise stated, the error bars represent the 95% confidence interval on the mean value, computed using bootstrapping.

In Figure 2, the analysis results are separated according to the different feature sets of the submissions. As visible, which features are present in the datasets does have an impact on the privacy risks. This effect is most apparent in the case of the linkability risk, where restricting the datasets to only demographic-focused features drastically reduces the risks. On the other hand, for the National dataset, the demographic feature set shows higher inference and singling out risk.

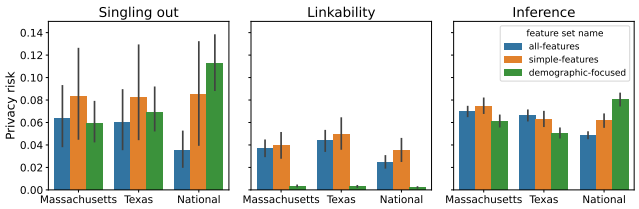

Figure 2. The three privacy risks measured by Anonymeter for the different datasets, separated per feature set.

3.2 Correlation with other privacy and utility metrics

As part of the CRC program, the submitted de-identified datasets are analyzed with the SDNIST data evaluation library [11]. In particular, the utility of the datasets is evaluated using the *k-marginal* utility metric, which is a measure of the difference between real and synthetic data with respect to multidimensional bin counts. A simple measure of privacy leaks counting the fraction of original records found in a de-identified dataset is also present. In Figure 3 we investigate how SDNIST's metrics correlate with our empirical evaluations of the privacy risks. As expected, higher privacy risks correspond both to larger fractions of duplicated records and to larger utility values. Singling out and inference show a higher correlation with the other metrics ($\rho = 0.69$ and $\rho = 0.47$, respectively) than the linkability risk ($\rho = 0.26$). This further suggests that the process of generating synthetic records significantly weakens the risk of linkability.

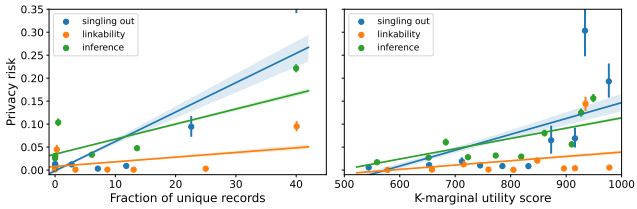

Figure 3. Anonymeter privacy risks vs privacy (left) and utility (right) metrics of the CRC submission.

Submissions from "Genetic SD", "smartnoise", "synthcity", "rsynthpop", "LostInTheNoise" and "Sarus SDG"[3], include algorithms satisfying the theoretical privacy definition of Differential Privacy (DP) [12]. For those algorithms, we can inspect the correlation of the measured privacy risks and the DP privacy budget ε, see Figure 4. For $\varepsilon \lesssim 5$ the measured risks are low and only mildly increasing. As the DP guarantee gets looser ($\varepsilon \gtrsim 10$) the empirical analysis is able to uncover larger and larger privacy risks. In Appendix D we present a more in-depth analysis of these results, showing that the degree of correlation between the DP budget and the empirically measured risks varies greatly across the different libraries.

3.3 Comparing different libraries and algorithms

The CRC program offers the opportunity to compare the results of different data synthetization algorithms, which we did with respect to the empirical privacy risks (see Figure 5 and Appendix E). Unsurprisingly, the sub-sampling approach (i.e., releasing 40% of the original data) results in the highest risk, estimated at ∼30% by Anonymeter (the theoretical value is 40%). Another group of algorithms, including Synthcity, RSynthpop, and Sdcmciro, shows moderately high risks around 10%. The rest of the submissions all have comparable and low risk values, around a few percent points. These results indicate that it is very difficult to reduce the privacy

[3] In this section we include also results from "LostInTheNoise" and "Sarus SDG", which are not included in the experiments of the rest of the paper — see Section 2.2.

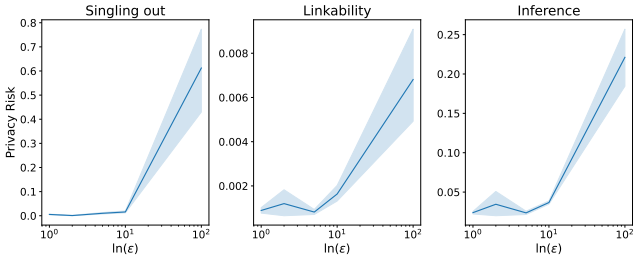

Figure 4. Aggregated privacy risks for differentially private submissions as a function of the privacy budget ε.

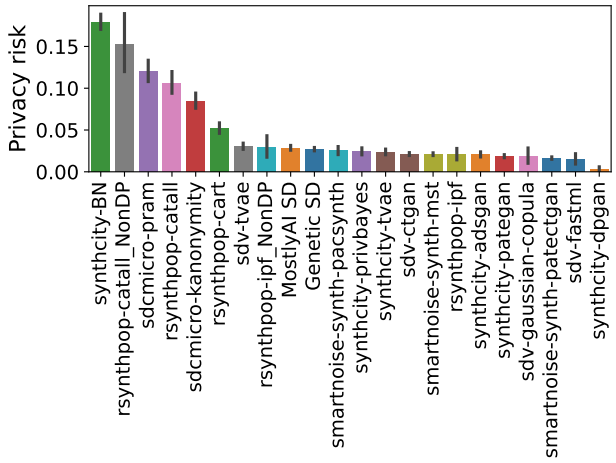

Figure 5. Comparison of the privacy risks for different libraries, aggregated across all the submissions.

risks below a few percent points with the current generation of synthetic data algorithms.

Finally, as far as the privacy-utility trade-off is concerned, RSynthpop, sdcMicro, MostlyAI-SD, and sub-sampling result in good levels of utility (see Figure 6 and Appendix F). However, with the exception of MostlyAI-SD, their submissions tend to result in larger privacy risks, too.

4. Outlook and conclusions

Our analysis reveals that while synthetic data techniques provide a measure of privacy protection, they are not infallible. While significant differences exist between various algorithms, none is able to decrease the privacy risks below a few percent points, a value which can be considered the current "state of the art".

This work highlights the variability in privacy protection and utility across different techniques. The privacy risks are influenced not only by the generative algorithms, but also by the nature and shape of the data: the number of rows and which features are selected in the analysis all play a role in determining the residual risks. These complex relationships highlight the need for evaluating the privacy risks of the generated synthetic datasets. Our analysis demonstrates the value of Anonymeter as an essential tool for quantifying and

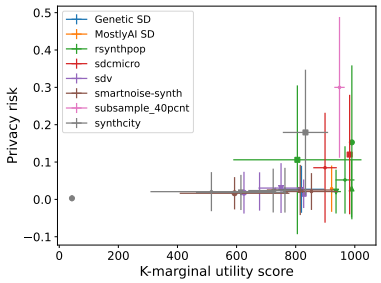

Figure 6. Privacy risk vs utility of various libraries, with standard deviations. Colors denote different libraries. Marker shapes distinguish between different algorithms for the same library, if there is more than one.

understanding such risks.

This paper contributes to the understanding of privacy risks and aligns with the CRC's objective of fostering robust and transparent research practices in data privacy. This is also a real-world application of the Anonymeter, a practical algorithm-agnostic framework for assessing privacy risks in synthetic data.

References

[1] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A Unified Framework for Quantifying Privacy Risk in Synthetic Data. *arXiv e-prints*, page arXiv:2211.10459, November 2022.

[2] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic Data – Anonymisation Groundhog Day. *arXiv e-prints*, page arXiv:2011.07018, November 2020.

[3] The NIST Collaborative Research Cycle (CRC) program homepage.

[4] Christine Task, Karan Bhagat, Streat Damon, and Gary Howarth. NIST Diverse Community Excerpts Data, December 2022.

[5] Florimond Houssiau, James Jordon, Samuel N. Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. Tapas: a toolbox for adversarial privacy auditing of synthetic data, 2022.

[6] Khaled El Emam, Lucy Mosquera, and Jason Bass. Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *J Med Internet Res*, 22(11):e23139, November 2020.

[7] Commission nationale de l'informatique et des libertés.

[8] Official CNIL opinion on Anonymeter.

[9] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques. 2014.

[10] The NIST Collaborative Research Cycle (CRC) Research Acceleration Bundle v1.1.

[11] Christine Task, Karan Bhagat, and Gary Howarth. Sdnist v2: Deidentified data report tool, December 2023.

[12] Cynthia Dwork. Differential privacy. *Automata, languages and programming*, 2006.

A. Submission census

The following table counts the number of de-identified datasets analyzed in this work, divided by library and dataset.

library name	target dataset	number of submissions
Genetic SD	ma2019	6
Genetic SD	national2019	7
Genetic SD	tx2019	6
LostInTheNoise	national2019	1
MostlyAI SD	ma2019	2
MostlyAI SD	national2019	2
MostlyAI SD	tx2019	2
Sarus SDG	national2019	1
rsynthpop	ma2019	2
rsynthpop	national2019	12
rsynthpop	tx2019	2
sdcMicro	ma2019	3
sdcMicro	national2019	5
sdcMicro	tx2019	3
sdv	ma2019	5
sdv	national2019	12
sdv	tx2019	4
smartnoise-synth	ma2019	14
smartnoise-synth	national2019	14
smartnoise-synth	tx2019	14
subsample_1pcnt	national2019	2
subsample_40pcnt	ma2019	3
subsample_40pcnt	national2019	3
subsample_40pcnt	tx2019	3
subsample_5pcnt	national2019	2
synthcity	ma2019	12
synthcity	national2019	15

Table 1. Number of analyzed submissions for different libraries and datasets.

B. Analysis details

For all privacy attacks, Anonymeter has two main parameters: N_{attacks} and n_{aux}.

N_{attacks} is the number of guesses that each attacker will make on its targets. In particular, for the inference and linkability attacks (the *main* and *control* attacks) N_{attacks} target records are randomly sampled (without replacement) from the respective target datasets. For the singling out attack, N_{attacks} is the number of singling out predicates generated by the attacker that are evaluated on the target datasets. N_{attacks} governs the precision on the risk estimates as the uncertainties on the measured success rates r_{main} and r_{control} scale as $\simeq 1/\sqrt{N_{\text{attacks}}}$ (they follow a Binomial distribution). Ideally one would want to set this this parameter to be as high as possible, within the limits imposed by the available computing resources.

The n_{aux} parameter governs the amount of information on the targets that the attacker can use. For the inference and linkability attacks, n_{aux} is the number of attributes of the original target records which are known to the attacker. The attacker will use this to match synthetic records to the target ones. For singling out, n_{aux} governs how many attributes are used in each predicate. Varying n_{aux} allows Anonymeter to model attackers of different strength. For inference, the actual columns used by the attacker are simply chosen as the first n_{aux} of the datasets, excluding the secret column that the attacker will try to guess. For linkability, each of the two vertical partitions of the data, among which links are to be found by the attacker, contains half of the n_{aux} columns with the highest number of unique values. Sorting the columns by number of unique values increases the strength of the attacks in the low n_{aux} regime. Finally, for singling out, the n_{aux} attributes used to create the predicates are selected at random for each predicate.

Besides these main parameters, which are common to all the attacks, there are two other parameters that govern the results of Anonymeter. The first one, k, is specific to the linkability attack and determines the number of nearest neighbors to consider when deciding if a link has been established (see Section 5.2 of [1] for more details). Higher values of k result in a looser definition of links and yield higher risk estimates. In this work, we computed linkability risks for k between 1 and 5, and aggregated the results. The second one, δ, controls when inference guesses on numerical attributes can be considered correct—see Section 5.3 of [1]. It bounds the maximum value of the relative difference between the true target value and the guess from the attacker. In this analysis we have used a value of $\delta = 0.05$.

C. Impact of dataset size on privacy risk

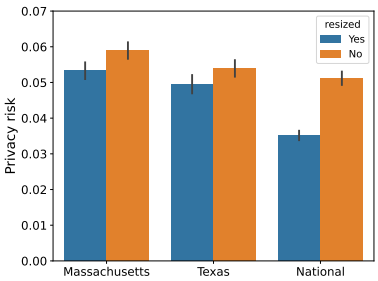

Figure 7. Comparison of the overall privacy risks measured for resized and not resized datasets.

The analyzed datasets have different sizes, ranging from $\sim$ 7600 records for the Massachusetts datasets to the $\sim$ 2700 of the Texas sample. To investigate the impact of the dataset size on the measured risks, we repeat the entire analysis but this time resizing all the datasets (original, control, and synthetic) to the size of the Massachusetts one.

The results are presented in Figure 7, showing the overall privacy risks for the different datasets. As visible, the relative

Library	Algorithm	Singling out	Linkability	Inference
Genetic SD	Genetic SD	0.3 ± 0.0	0.2 ± 0.0	3.3 ± 0.1
MostlyAI	MostlyAI	1.9 ± 0.2	0.5 ± 0.0	3.4 ± 0.2
rsynthpop	cart	5.9 ± 0.7	0.2 ± 0.0	7.4 ± 0.4
	catall	24.3 ± 4.7	0.5 ± 0.1	13.7 ± 0.9
	catall noDP	49.4 ± 9.3	0.6 ± 0.1	18.4 ± 2.0
	ipf	0.1 ± 0.1	0.1 ± 0.0	3.2 ± 0.5
	ipf noDP	1.3 ± 0.6	0.0 ± 0.0	4.4 ± 1.0
sdcMicro	kanonymity	28.4 ± 3.4	0.4 ± 0.0	10.1 ± 0.6
	pram	20.4 ± 2.1	0.9 ± 0.1	16.1 ± 0.9
sdv	ctgan	0.7 ± 0.1	0.1 ± 0.0	2.6 ± 0.1
	fastml	0.2 ± 0.1	0.0 ± 0.0	2.3 ± 0.5
	copula	0.0 ± 0.0	0.1 ± 0.0	2.7 ± 0.7
	tvae	3.6 ± 0.7	0.3 ± 0.0	3.8 ± 0.2
smartnoise	mst	1.7 ± 0.2	0.1 ± 0.0	2.7 ± 0.1
	pacsynth	0.4 ± 0.1	0.1 ± 0.0	3.9 ± 0.4
	patectgan	1.0 ± 0.1	0.0 ± 0.0	2.1 ± 0.1
subsample	subsample	55.8 ± 2.6	23.1 ± 0.9	30.4 ± 0.4
synthcity	adsgan	0.9 ± 0.2	0.1 ± 0.0	2.6 ± 0.2
	BN	6.8 ± 0.8	8.7 ± 0.7	21.1 ± 0.5
	dpgan	0.1 ± 0.0	nan $\pm$ nan	0.4 ± 0.2
	pategan	0.7 ± 0.1	0.1 ± 0.0	2.4 ± 0.1
	privbayes	0.8 ± 0.1	0.1 ± 0.0	3.7 ± 0.3
	tvae	1.0 ± 0.2	0.1 ± 0.0	3.1 ± 0.2

Table 2. Privacy risks for the different libraries and algorithms. The reported values are the mean risks, aggregated over all datasets and analysis parameters, with their standard errors.

level of risks remains the same: the risks measured in the Massachusetts datasets are higher, and are the lowest in the National dataset. However, in general, reducing the size of the datasets diminishes the risk. This may be attributed to the fact that the strength of the adversary is dependent on the amount of information contained in the synthetic dataset. Smaller datasets carry less information, leading to lower privacy risks.

More experiments would be needed to investigate this further. Also, we note that resizing the dataset post synthetization is not ideal: the synthetic samples are generated from a model that has seen all the data. A more rigorous approach would be to resize the original data prior to synthetization.

D. Correlation with DP budget

In Section 3.2 we have explored the correlation between the theoretical DP privacy budget ε and the empirically measured risks reported by Anonymeter. As the values of ε vary across the different libraries (and some of the libraries submitted only de-identified datasets with a single ε value), a more in-depth analysis is desirable. In Figure 8 we present the correlation between the empirical risks and theoretical ε for the different libraries. As visible from the top panel, the correlation is much stronger for Rsynthpop: as the privacy budget increases, the empirical risks become larger and larger, reaching extremely high values, especially for singling out, for an epsilon of 100. Other libraries produce smaller risks and a milder correlation between the DP guarantee and empirical risks: see the bottom panel of Figure 8.

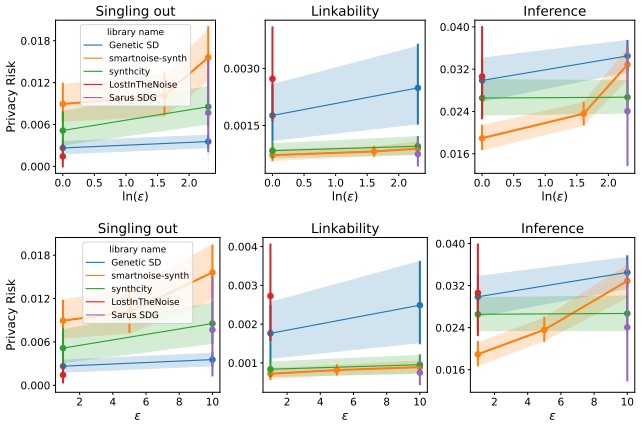

Figure 8. Correlation of the risks as measured by Anonymeter and the DP privacy budget ε for the different libraries. Note the log-scale on the x axis of the top panel. The bottom panel presents the same results, excluding the Rsynthpop library.

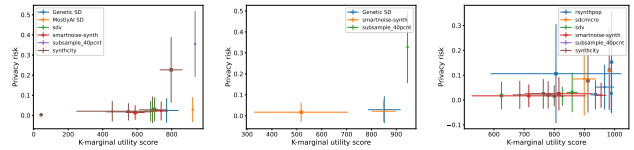

Figure 9. Privacy and utility of the submission relative to the "all-features" (left), "simple-features" (center), and "demographic-focused" (right) feature sets. Different colors identify different libraries. The marker style distinguishes the algorithms. The error bars report the standard deviation of the k-marginal score (on the x axis) and the privacy risks (on the y axis).

E. In-depth comparison of the different algorithms

In this section we explore how the measured privacy risks vary across the different libraries and generative algorithms: see Table 2. An updated version of this table is available at https://github.com/AnonosDev/anonymeter-sdnist/tree/main.

Finally, Figure 10 compares the performance of different algorithms within the same libraries, highlighting variations in privacy risk levels for the different datasets. As visible, for some libraries there is a great difference in privacy risks depending on the chosen algorithm. For SDV, for exampe, the TVAE algorithm consistently yields higher risks. The Synthcity library offer more algorithms than any other. Their Bayesian Network approach has a significantly worse privacy performance than the others, especially for linkability and inference. We also note that Synthcity's Bayesian Network

has a very high linkability risks, while for all other algorithms the linkability risks is very small.

F. Privacy utility trade-off

In Figure 9 we present privacy versus utility plots similar to Figure 6, separated by the three feature sets.

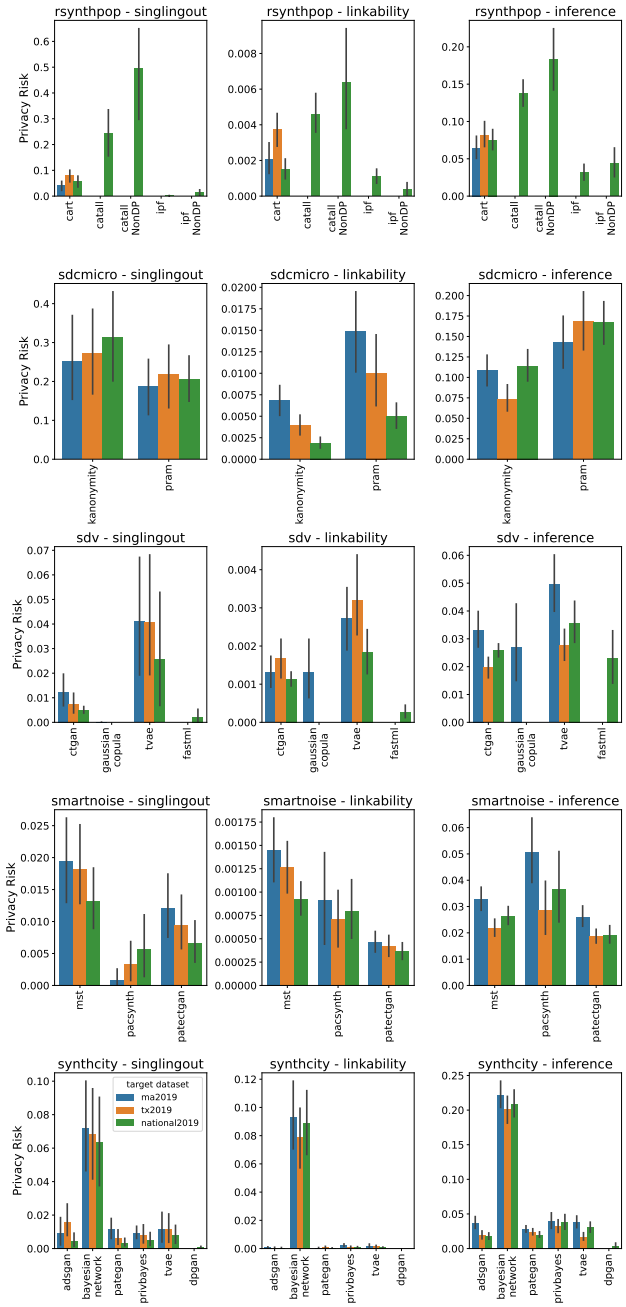

Figure 10. Comparison of the measured privacy risks for the different datasets and for each generative algorithm.