A Unified Framework for Quantifying Privacy Risk in Synthetic Data

Matteo Giomi
matteo@statice.ai
Statice
Berlin, Germany

Franziska Boenisch
franziska.boenisch@vectorinstitute.ai
Vector Institute
Toronto, Canada

Christoph Wehmeyer
chris@statice.ai
Statice
Berlin, Germany

Borbála Tasnádi
borbala@statice.ai
Statice
Berlin, Germany

ABSTRACT

Synthetic data is often presented as a method for sharing sensitive information in a privacy-preserving manner by reproducing the global statistical properties of the original data without disclosing sensitive information about any individual. In practice, as with other anonymization methods, synthetic data cannot entirely eliminate privacy risks. These residual privacy risks need instead to be *ex-post* uncovered and assessed. However, quantifying the actual privacy risks of any synthetic dataset is a hard task, given the multitude of facets of data privacy.

We present Anonymeter, a statistical framework to jointly quantify different types of privacy risks in synthetic tabular datasets. We equip this framework with attack-based evaluations for the *singling out*, *linkability*, and *inference* risks, which are the three key indicators of factual anonymization according to data protection regulations, such as the European General Data Protection Regulation (GDPR). To the best of our knowledge, we are the first to introduce a coherent and legally aligned evaluation of these three privacy risks for synthetic data, as well as to design privacy attacks which model directly the singling out and linkability risks.

We demonstrate the effectiveness of our methods by conducting an extensive set of experiments that measure the privacy risks of data with deliberately inserted privacy leakages, and of synthetic data generated with and without differential privacy. Our results highlight that the three privacy risks reported by our framework scale linearly with the amount of privacy leakage in the data. Furthermore, we observe that synthetic data exhibits the lowest vulnerability against linkability, indicating one-to-one relationships between real and synthetic data records are not preserved. Finally, with a quantitative comparison we demonstrate that Anonymeter outperforms existing synthetic data privacy evaluation frameworks both in terms of detecting privacy leaks, as well as computation speed. To contribute to a privacy-conscious usage of synthetic data, we publish Anonymeter as an open-source library.[1]

KEYWORDS

anonymization, privacy risks, synthetic data, GDPR, differential privacy

1 INTRODUCTION

Societies in the digital era are faced with the challenge of striking a balance between the benefits that can be obtained by freely sharing and analyzing personal data, and the dangers that this practice poses to the privacy of the individuals whose data is concerned. Replacing original and potentially sensitive data with some synthetic data, *i.e.*, data that is artificially generated rather than coming directly from real individuals, is one of the approaches that attempt to resolve this tension [8, 46, 51]. Synthetic data captures population-wide patterns of the underlying potentially sensitive data while "hiding" the characteristics of the individuals. Popular approaches for synthetic data generation rely on deep generative models such as Generative Adversarial Networks (GAN) [23] and Variational Autoencoders (VAE) [27]. These generative models are trained on the original and potentially sensitive data to produce a synthetic dataset that preserves the *utility* of the original data.

Intuitively, if the generative models are able to generalize well, the synthetic data should not reflect the particular properties of any individual original record. This intuition underpins the use of synthetic data as a privacy-enhancing technology. Unfortunately, assuming that synthetic data simply carry no privacy risks—despite being tempting—is too simplistic. Generative models with enough capacity to express complex data patterns are often found to match the original data too closely [10, 55, 69]. As a consequence, the synthetic data will likely present some residual privacy risk. Reliably quantifying the privacy risk of synthetic data is therefore an important, yet still not settled, problem—even though such an assessment is not just desirable, but often represents a requirement imposed by legal frameworks, such as the European General Data Protection Regulation (GDPR [61]), the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA [41]), an California's Consumer Privacy Act (CCPA [57]).

Differential Privacy (DP) [16] provides a theoretical framework for upper-bounding privacy leakage. However, a gap exists between the worst-case privacy guarantee of DP and what can be empirically measured for practical attacks [38]. Moreover, due to the stochastic training and inference of generative models, the magnitude of the effective privacy risks exposed by the generated synthetic data

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

Proceedings on Privacy Enhancing Technologies 2023(2), 312–328

https://doi.org/10.56553/popets-2023-0055

cannot be quantified in advance [56]. Instead, the residual privacy risks need to be measured a posteriori in an empirical fashion from the generated data. However, since there exist various notions of *practical privacy* (*e.g.*, membership privacy, attribute privacy [63]), there is no unified metric for such a measurement. Instead, different metrics have been proposed to measure privacy risks for anonymized data [63]. Yet, interpreting and combining them in a meaningful way is still a topic subject to research.

In this work, we set out to close this gap by proposing Anonymeter, an empirical and statistical framework that measures privacy risks in anonymized tabular datasets. Anonymeter implements a general three-step procedure for risk assessment based on (1) performing privacy attacks against the dataset under evaluation, (2) measuring the success of such attacks, and (3) quantifying the exposed privacy risk in a well-calibrated and coherent manner. Each of the three steps is connected to the others via common interfaces to keep the framework modular and to allow the same risk quantification method to be shared by different privacy risks. For each privacy attack, the final risk is obtained by comparing the results of the privacy attack against two baselines; the first resulting from performing the same attack on a control dataset from the same distribution as the dataset under evaluation, and the second resulting from performing a random attack against the dataset under evaluation. While the latter provides insights into the strength of the main attack, the former makes it possible to measure how much of the attacker's success is simply due to the *utility* of the synthetic data, and how much is instead an indication of *privacy violations*. Within this framework, we propose three attacks to quantify the risks of *singling out*, *linkability*, and *inference*, respectively—the three privacy metrics defined by the Article 29 Data Protection Working Party [44], making our framework legally aligned.

We provide an experimental validation of our framework by testing its ability to detect different amounts of privacy leaks. Our results demonstrate that the framework is able to detect these leaks, that the reported risks scale linearly with the amount of privacy leaks present in the dataset, and that risk-computation is efficient even on large datasets. We also show (Section 6.4) that Anonymeter outperforms existing evaluation frameworks for synthetic data [56] in both computational performance and quality of privacy assessment. Furthermore, in our experiments (Section 6.3), synthetic data exhibits the highest risks to inference and singling out attacks, whereas the risk to linkability is comparably low over all datasets evaluated. This provides empirical evidence for the common intuition that generating synthetic data breaks the one-to-one links between data records (see, *e.g.*, [11, 20]). As expected, introducing DP into the training of the generative models [1] causes a general decrease in the privacy risks reported by Anonymeter; besides, a higher utility of the generated data corresponds to a higher reported risk, *i.e.*, the more the synthetic data is close to the original data the higher the risk reported by Anonymeter.

In this work, we propose implementations of singling out, linkability, and inference privacy attacks. Yet, Anonymeter's modularity allows for a simple and consistent integration of additional attack-based privacy metrics. Anonymeter is designed to be widely *usable* and to provide *interpretable* results, requiring minimal manual configuration and no expert knowledge besides basic data analysis skills. It is also *applicable* to a wide range of datasets and to both numerical and categorical data types. Anonymeter is *sensitive* and able to identify and report even small amounts of privacy leaks. Although developed for the specific case of synthetic data, our tool does not make any assumption on how the data is created, except for requiring consistency of attributes and data types, and can also be applied to assess other forms of anonymization and pseudonymization.

The software implementing our privacy evaluations is released as an open-source library.[2]

In summary, we make the following contributions:

- We propose Anonymeter, a statistical framework to derive well-calibrated privacy risks against different types of privacy attacks in a coherent way.
- To the best of our knowledge, we are the first to introduce a practical tool for jointly evaluating all the three privacy risks (singling out, linkability, and inference) defined by the main privacy regulations [44].
- For each risk, we provide a concrete attack algorithm, its implementation, and thorough experimental evaluation. Our framework compares favourably to previous work on privacy evaluation of synthetic datasets (see Section 6.4).
- We are the first to propose a practical implementation for the risk of singling out, and a linkability attack directly modeling the possibility of linking records belonging to specific individuals, rather than relying on related concepts, such as membership inference [56].
- We provide our framework, including the attack implementations, as an open-source library to support practical privacy-enhanced synthetic data usage.

2 NOTATION AND BACKGROUND

The following notation is used in the rest of the paper: a tabular dataset X is a collection of N records $\mathbf{x} = (x_1, ..., x_d)$, each with d attributes, drawn from a distribution $\mathcal{D}$. We use the subscripts ori and syn to denote original datasets, *i.e.*, collections of data records sampled from $\mathcal{D}$, and synthetically created datasets, respectively. More in detail, we denote an original dataset by $X_{\text{ori}} = \{\mathbf{x}^1_{\text{ori}}, ..., \mathbf{x}^N_{\text{ori}}\}$ and a synthetic dataset by $X_{\text{syn}} = \{\mathbf{x}^1_{\text{syn}}, ..., \mathbf{x}^M_{\text{syn}}\}$. We make use of matrix notation to indicate columns in the datasets: $X[:, i]$ is a vector of size N containing the i^{th} attribute of all the records, and $\mathbf{x}[i] = x_i$ is the value of the i^{th} attribute of record $\mathbf{x}$. Finally, $\mathcal{G}$ denotes the generative model from which the synthetic dataset X_{syn} is produced.

2.1 Synthetic Data Generation

In general, synthetic data is produced by a generative model $\mathcal{G}$ that is supposed to learn the distribution $\mathcal{D}$. However, since this distribution is usually unknown, the model $\mathcal{G}$ is instead trained on X_{ori} sampled from $\mathcal{D}$. Once trained, the model $\mathcal{G}(X_{\text{ori}})$ can be understood as a stochastic function that, without any input, generates synthetic data records $\mathbf{x}_{\text{syn}}$ [56]. By querying $\mathcal{G}$ multiple times, a full synthetic dataset X_{syn} can be sampled. Ideally, the generated synthetic data should reflect most of the statistical properties of the distribution $\mathcal{D}$. Yet, since $\mathcal{G}$ only has access to X_{ori}, and only learns

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

a partial representation of the data distribution, the generated data can only approximate $\mathcal{D}$.

Several methods exist to generate synthetic data. A possibility is to use statistical models, such as Bayesian networks [68] or Hidden Markov models [15, 40]. Such models generate explicit parametric representations of $\mathcal{D}$ and the features to be extracted from X_{ori} are determined beforehand. In contrast, deep learning models for synthetic data generation, such as GANs [23] and VAEs [27], learn which attributes to extract during a stochastic training process.

2.2 Privacy-Preserving Synthetic Data

One main reason to generate synthetic data is for the purpose of privacy-preserving data releases and data sharing: synthetic datasets are supposed to reproduce the properties of an original dataset X_{ori} from $\mathcal{D}$ without containing personal data from X_{ori} [2, 56]. Yet, recent studies indicate that through high-utility synthetic data it is still possible for an attacker to extract sensitive information about the original data [56].

In this work, we use CTGAN [66] to generate the synthetic data for our experiments. CTGAN is a GAN that uses a conditional generator to enable the generation of synthetic tabular data with both discrete and continuous-valued columns. The approach uses a mode-specific normalization as an improvement for non-Gaussian and multimodal data distributions [66]. Privacy guarantees can be integrated into CTGAN using DP [16]. In this case, the generative model is trained with a DP optimizer [1]. As a result of the post-processing robustness of DP, the synthetic data generated from such DP models also enjoy the same level of privacy guarantee as the trained generative model [2, 26, 45, 70].

2.3 Privacy Metrics and Attacks in Synthetic Data

Privacy is a multi-faceted concept, which is reflected in the availability of dozens of different privacy metrics—see, e.g., [63] for an overview. In the concrete case of measuring the privacy leakage of synthetic data, many studies rely on similarity tests [67], on distance metrics calculating the mean absolute error between original and generated data records [33], or on measuring the number of identical records between original and synthetic datasets [30]. When the synthetic data is generated with DP guarantees, the DP privacy budget, usually denoted by ε, can also be used to report on the privacy of a synthetic dataset. However, for most of these metrics, it is unclear how they translate into privacy implications in practice and what concrete privacy risks exist for individual data records [38]. Therefore, using the success rate of concrete privacy attacks is becoming a common approach to quantifying the privacy of synthetic data, *e.g.* [56]. In this work, we integrate into Anonymeter the evaluation of *the three* privacy attacks that anonymization techniques must protect from, according to the GDPR [44]: *singling out*, *linkability*, and *inference*. We close the gap of prior evaluation frameworks *e.g.*, [56] that only jointly consider a subset of these legally essential risks. The importance to consider these three risks against anonymization results from their implication on individuals' privacy: singling out can be seen as a way to indirectly identify a person in a dataset. At the same time, it can serve as a stepping stone towards linkage attacks [12], which have been shown to yield complete de-anonymization of datasets in the past [36, 58]. Inference attacks, in turn, can disclose highly sensitive information on individuals, such as their genomics [7].

Singling Out happens whenever it is possible to deduce that within the original dataset there is a single data record with a unique combination of one or more given attributes [44]. For example, an attacker might conclude that in a given dataset X_{ori} there is exactly one individual with the attributes gender: *male*, age: *65*, ZIP-code: *30305*, number of heart attacks: *4*. It is important to note that singling out does not imply re-identification. Yet the ability to isolate an individual is often enough to exert control on that individual, or to mount other privacy attacks [9].

Linkability is the possibility of linking together two or more records (either in the same dataset or in different ones) belonging to the same individual or group of individuals [44]. This can be used for de-anonymization, as proven in the past [37, 59]. Due to statistical similarities between the generated data and the original data, linkability risks may still exist in synthetic datasets.

Inference happens when an attacker can confidently guess (infer) the value of an unknown attribute of the original data record [44]. An example of successful inference would consist in the attacker being able to confidently deduce that a record in the original dataset X_{ori}, with attributes "gender": *male*, "age": *65*, "ZIP-code": *30305* holds the secret attribute "number of heart attacks": *4*.

When measuring privacy risks, an important distinction has to be made between what an attacker can learn at a population level (generic information) and on an individual level (specific information). Generic information is what provides utility to the anonymized data; specific information enables the attacker to breach the privacy of some individuals [17]. Anonymeter distinguishes what the attacker learns from the anonymized dataset as generic information from specific inference, thus quantifying the privacy risk (see Section 4.2).

3 RELATED WORK

In this section, we present related work on privacy attacks and privacy evaluation of (synthetic) datasets.

Privacy Attacks: An extensive theoretical analysis and formalization of *singling out* has been carried out in [13]. Such analysis relies on the knowledge of the population distribution $\mathcal{D}$, which requires complex and careful modeling, *e.g.* [49]. That would make the evaluation harder to use and to interpret. We will instead approximate the baseline probability by carefully comparing the success of the attack on different samples from the data. To the best of our knowledge, we are the first to propose a singling out attack algorithm for tabular synthetic data, and to implement it in a practical and usable framework.

The concepts of *linkability* and *inference* attacks have extensively been studied both on synthetically generated datasets [56] and datasets that were anonymized with other approaches [25, 37, 59]. While many practical inference attacks exist, *e.g.* [37, 56], our work is the first one to devise a methodology to measure directly the linkability risk posed by the release of tabular "anonymized" data, and to implement it in a practical and efficient algorithm. In addition to empirical demonstration of the power of those attacks, theoretical analyses have been introduced, which formalize those concepts and show that the success rate of an attack can be improved when an attacker holds auxiliary information about rare attributes [34]. Finally, *background knowledge attacks* have been proposed, in which an attacker holds knowledge about target individuals and uses it to reconstruct their information [42] and to perform *membership inference* [29], *i.e.*, deduce the presence or absence of a particular individual in the original dataset.

Privacy Evaluation Frameworks: Similar to Anonymeter, other frameworks exist that rely on using privacy metrics and attacks to evaluate the privacy of synthetic datasets. Yet, to the best of our knowledge, we are the first in this line of work to leverage a control dataset for estimating specific privacy risk in contrast to generic population-wide risk. [30] uses similarity metrics between original and generated data records to quantify the risk of re-identification for data generated with GANs. [28] calculates a per-data record privacy by measuring a memorization coefficient based on distance measures between original and synthetic data records in a latent space. [3] introduces the framework of differential testing, similar to the notion of DP, which deems a dataset anonymized when the inference accuracy from the synthetic data is about the same whether the user's record is included in the original dataset or not. In the medical domain, [43] and [18, 33] evaluate privacy vs. utility of generative models and synthetic datasets, respectively. The first two works rely on the notion of membership inference to quantify the privacy risks, the last one on equivalence classes and intra-record distances. Finally, attack-based approaches have been used to quantify the strength of database query anonymization mechanisms [4].

The framework for privacy evaluation in synthetic data that is most closely related to ours is [56]. It operates in the same scenario as we do by treating the data generation mechanism as a black box and by evaluating the privacy leakage directly on the generated synthetic data. The privacy leakage is measured in terms of a *privacy gain* (*PG*), which is defined as the difference between the *advantage* of an adversary trying to guess secrets about a target data record given the original data compared to an adversary with access only to the synthetic data. The attack is formalized as an adversarial game. For a given target, such games are repeated many times, each time requiring training of a new generative model and sampling several synthetic datasets. As a consequence, the computational cost of the attack is very high which severely limits its usability. As detailed in Section 6.4, even without taking the training of the generative models into account, [56]'s methodology is ten times slower than Anonymeter.

Additionally, our privacy risk metric has several advantages over the Privacy Gain defined in [56]. First, it is closer to the legal terminology and therefore easier to understand and explain than *PG*. Second, contrary to our work, in [56] there is no separation between how much of the attacker's advantage is due to actual privacy violations and how much of it is instead due to the utility of the synthetic data. Third, *PG* is measured from the evaluation of a few records, typically hand-picked outliers. It is not clear how a characterization of the privacy of a restricted sample of records can be extrapolated to a whole dataset. The privacy risk computed by Anonymeter is instead obtained by attacking thousands of different targets, making it a more robust metric compared to *PG*. Anonymeter also naively provides an estimate of the statistical uncertainties on the reported risk. Moreover, as we will see in Section 4.2, Anonymeter incorporates sanity checks ensuring that the risk estimate is based on attack models that are actually effective.

Finally, [56] does not include any evaluation of the singling out risk. Also, in [56] the linkability *PG* is not measured directly, but instead is indirectly derived from a membership inference game.[3]

4 ANONYMETER FRAMEWORK

We propose a framework which provides a coherent assessment of diverse privacy risks based on dedicated privacy attacks.

4.1 Threat Model and Assumptions

To provide a conservative privacy risk assessment, we consider the strongest threat model in which the attacker is in full possession of the synthetic dataset. Moreover, the attacker holds additional partial but correct knowledge, called *auxiliary information*, about a subset of the original records (the *target* records). This accounts for practical scenarios where overlapping data sources are common. Depending on the amount and quality of the auxiliary information, more or less powerful attacks can be modeled. We rely on simple yet effective heuristics to choose the auxiliary knowledge for our respective attacks (see Section 5 for details). The targeted original records are chosen at random from the original dataset X_{ori}. That is, we make no assumption on how the attacker would choose the targets, resulting in a more robust evaluation of the overall privacy offered by the synthetic data. If needed, however, the framework can easily be adapted to attack specific records, for example to measure privacy risks for some specific sub-population in the data, or particular individuals.

We treat the data generation mechanism as a black box that *cannot* be accessed or queried by the attacker who only receives the synthetic dataset. Concerning the original dataset X_{ori}, we assume that it consists of N records drawn independently from the population $\mathcal{D}$, and that each record refers to a different individual. The full original dataset is split into two disjoint sets X_{train} and X_{control}. That is, $X_{\text{ori}} = X_{\text{train}} \cup X_{\text{control}}$ and $X_{\text{train}} \cap X_{\text{control}} = \emptyset$. The synthetic dataset X_{syn} is sampled from a generative model trained on X_{train} exclusively: $X_{\text{syn}} \sim \mathcal{G}(X_{\text{train}})$. In Appendix A we show that, in the scope of Anonymeter where X_{control} consists of at least a few thousand data records, training the generative model on X_{train} instead of on the full X_{ori} does not affect the utility of the resulting synthetic data. To evaluate privacy, Anonymeter requires all three datasets, X_{train}, X_{syn}, and X_{control}. All the datasets have the same number of attributes d, but the number of records might differ.

4.2 Framework Architecture

Privacy risks in our framework are estimated following a common procedure, see Figure 1. In this modular design, the analysis is made up of three sequential steps: (i) the *attack phase* in which the privacy attacks are carried out, (ii) the *evaluation* phase in which the success of the attacks is measured, and (iii) the *statistical risk*

[3] However, linkability cannot be fully measured through membership inference. This is because linkability is a stronger privacy breach. In fact, establishing a successful linkage between a target t and some records in a dataset X implies membership of t in X. Yet, the opposite is not true, *i.e.*, even if membership of t in X is established, the attacker still does not know which records of X are directly linked to the target.

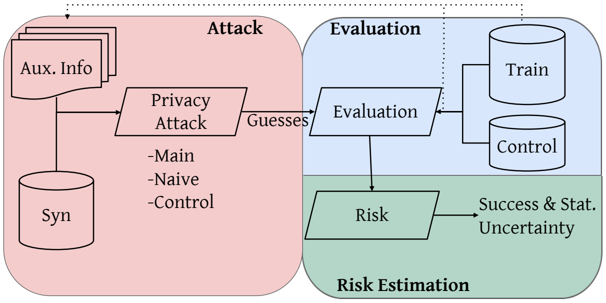

Figure 1: Schematic overview of our framework. The attacker is given access to the entire generated synthetic dataset and some auxiliary information. (1) In the *attack phase*, the framework performs three different attacks (main, naive, and control) each of which outputs guesses on the original private data. (2) The correctness of these guesses is then evaluated in the *evaluation phase* against the original training data (main and naive attack) or the control data (control attack). (3) Based on the evaluation, a final statistical risk quantification is output in the *risk estimation* phase.

estimation phase quantifying the privacy leakage for the attack. How the concrete attacks and evaluation phases are carried out differs for each evaluated privacy risk, but the way the risks are derived from the results of the evaluation phase is common to all cases. This consistent view improves the interpretability of the different attack results and makes Anonymeter more modular.

Attack Phase: the attack phase consists of executing three different attacks. First, the "main" privacy attack in which the attacker uses the synthetic dataset X_{syn} to deduce private information of records in the training set X_{train}. Second, a "naive" attack is carried out based on random guessing, to provide a baseline against which the strength of the "main" attack can be compared. Finally, to distinguish the concrete privacy risks of the original data records (*i.e.*, specific information) from general risks intrinsic to the whole population (*i.e.*, generic information), a third "control" attack is conducted on a set of control records from X_{control}. For all the risks, each of the three attacks is formulated as the task of making a set of guesses: $\mathbf{g} = \{g_1, ..., g_{N_A}\}$ on N_A original *target* records. As an example, a singling out guess could state that "there is just one person in the original dataset who is male, 65 years old and lives in area 30305". The naive attack draws its guesses at random, using the synthetic dataset only to know the domain of the dataset attributes. The main and the control attacks generate the guesses trying to actively leverage the synthetic dataset (and the auxiliary information, when available) to gain information. They both share the same attack algorithm, but in the main privacy attack such guesses are evaluated against X_{train}, whereas in the control attack the guesses are evaluated against X_{control}. Note that X_{control} is completely independent of the synthetic data generated from X_{train} (see Section 4.1). Hence, if the attacker is successful in guessing information about records in X_{control}, this must only be due to patterns and correlations that are common to the whole population X_{ori}, rather than being specific to some training record. Therefore, the difference between the success rate of the two attacks can provide a measure of privacy leakage that occurred by training $\mathcal{G}$ on X_{train}: Anonymeter reports a privacy leakage when the attacker is more successful at targeting X_{train} than X_{control}.

Evaluation Phase: In the evaluation phase, the guesses from the attack phase are compared against the truth in the original data to estimate the privacy risk. The outcome of the evaluation phase is a vector of bits $\mathbf{o} = \{o_1, ..., o_{N_A}\}$, where $o_i = 1$ if the i^{th} guess g_i is correct, otherwise $o_i = 0$. Each attack defines the criteria for a guess to be considered correct. In the singling out example from above, the guess would be considered correct if there indeed exists exactly one such individual in the original data.

Risk Quantification Phase: In the risk quantification phase success rates of the "main" privacy attack are derived from the evaluation together with a measure of the statistical uncertainties due to the finite number of targets. Under the assumption that the outcome o_i of each attack is independent from the others, $\mathbf{o}$ can be modeled as Bernoulli trials. We define the true privacy risk $\hat{r}$ as the probability of success of the attacker in these trials. The best estimate r of the true attacker success rate $\hat{r}$ and the accompanying confidence interval $\hat{r} \in r \pm \delta_r$ for confidence level α are estimated via the Wilson Score Interval [65]:

$$\begin{aligned} r &= \frac{N_S + z_\alpha^2/2}{N_A + z_\alpha^2} \\ \delta_r &= \frac{z_\alpha}{N_A + z_\alpha^2}\sqrt{\frac{N_S(N_A - N_S)}{N_A} + \frac{z_\alpha^2}{4}}, \end{aligned} \tag{1}$$

with $N_S = \sum_{i=1}^{N_A} o_i$ being the total number of correct guesses, and z_α the probit, *i.e.*, the inverse of the cumulative distribution function of the normal distribution, corresponding to the confidence level α. Using Equation (1) and the number of successful guesses for each attack, we evaluate the success rates for the "main", "naive", and "control" attacks as $(r_{\text{train}} \pm \delta_{\text{train}})$, $(r_{\text{naive}} \pm \delta_{\text{naive}})$, and $(r_{\text{control}} \pm \delta_{\text{control}})$, respectively.

The success rate of the naive attack provides a baseline to measure the strength s of the attack which can be defined as the difference between the success rate of the main attack against training records and the success rate of the naive attack:

$$s = r_{\text{train}} - r_{\text{naive}} \tag{2}$$

with the error on s obtained via error propagation as $\delta_s = \sqrt{\delta_r^2 + \delta_{naive}^2}$.

If the attack is weaker than the naive baseline, *i.e.*, $r_{\text{naive}} \geq r$, the attack is said to have failed. This can happen in case of incorrect modeling, for instance when the attacker is given too little auxiliary information or auxiliary information that is uncorrelated with the guess's objective, or when the synthetic dataset has little utility and it is actually misleading for the attack. In such cases, Anonymeter warns the user that the results are considered void of meaning and must be excluded from the analysis. See Table 2 in Section 6.3 for details on the fractions of generated valid attacks for our three concrete instantiations of singling out, linkability and inference risks. Note that these fractions depend on the concrete attacks and their respective instantiations. Excluding invalid attacks is important in practice, because it avoids the situation in which "no risk" is reported due to the incorrect modeling of the attacks.

For the "control" attack, the attack's success rate is evaluated on control records (r_{control}) using Equation (1). Intuitively, if the synthetic dataset contains more information on the training records than on those in the control set, this implies $r_{\text{train}} \geq r_{\text{control}}$. From these two success rates the specific privacy risk R is derived as as:

$$R = \frac{r_{\text{train}} - r_{\text{control}}}{1 - r_{\text{control}}}. \tag{3}$$

The numerator in the above expression corresponds to the excess of attacker success when targeting records from X_{train} versus the success when the targets come from X_{control}. The denominator represents the maximum improvement over the control attack that a perfect attacker ($r = 1$) can obtain, and helps contextualizing the difference at the numerator by acting as a normalization factor. For example, suppose that out of 100 guesses the attacks against the training and control sets are correct 90 and 80 times, respectively: $r_{\text{train}} = 0.9$ and $r_{\text{control}} = 0.8$. Of the 90 correct guesses of the main privacy attack, 80 could be explained as being due the utility of the dataset, leaving the remaining 10 correct guesses to indicate privacy violations. This 0.1 excess in the success rate r_{train} translates in a $R = 0.5$ risk, since the best possible attack can only score 100 out of 100 guesses, *i.e.*, its rate can only be 0.2 higher than r_{control}. Similar ways of normalizing the risk are proposed in [4, 32], with the difference that in our work the normalizing baseline (r_{control}) is derived from attacking a control set of records, rather than from the success of the naive attack.

If both success rates are identical, access to the synthetic data does not give the attacker any benefit to gain information about the training data: the success of the attack can be explained by the general utility of the synthetic data, *i.e.*, it is a consequence of general inference. If, however, the success rate on training data exceeds the one on control data, this shows that information has been leaked from the synthetic data.

Properties of our Privacy Risk: There are no unified requirements on properties that privacy metrics should possess [63]. Our privacy risks align best with the three main requirements of the privacy metrics formulated by [53]. First, we evaluate the *correctness* of the guesses generated by the attacker; second, we report the *uncertainty* of the risks through confidence intervals; and third, as demonstrated empirically in Sections 6.1 and 6.4, our proposed metric is accurate as it is able to measure the actual percentage of data leaked from the synthetic dataset. The experiments of Sections 6.1 and 6.4 also show that our metric is *meaningful* in the sense that $R \sim 0$ if (and only if) the evaluated dataset is independent of the original data (noninterference), and that the reported risk increases proportionally with the amount of privacy leaks [54]. Finally, our privacy risks are based on probabilities, namely the probabilities of making correct guesses on the sensitive data, as suggested for privacy metrics by [5].

4.3 Practical Privacy Evaluation Bounds

In general, an attack-based privacy analysis provides a *lower* bound for the privacy risk (in contrast to theoretical frameworks, such as DP [16] that provide *upper* bounds, *i.e.*, worst-case guarantees). Therefore, the computed privacy risks are just as representative as the employed attacks are. Yet, in practice, using attack-based approaches to quantify privacy leakage has become state-of-the-art in several domains, such as machine learning, *e.g.* [35, 39].

To overcome potential limitations of an attack-based approach, we model powerful and knowledgeable attackers: we always assume the attacker holds knowledge of the entire synthetic dataset—the worst case scenario in which the synthetic data is released to the public. Note that in many practical applications, the value of the datasets which are processed, see *e.g.* [14, 21, 22, 62, 64], discourages such scenarios. In addition, for the linkability and inference estimation, partial but correct auxiliary knowledge of some original records is also available to the attacker. Finally, we evaluate the attack strength by comparing to a baseline attack based on the uninformed guesses from the naive attack. This adds the context needed to interpret the results correctly. In particular, the results are only valid if the "main" attack is able to outperform the baseline scenario.

5 OUR FRAMEWORK'S PRIVACY ATTACKS

This section proposes concrete instantiations of three privacy attacks to assess the fundamental risks of *singling out*, *linkability*, and *inference* within our framework. We chose to implement attacks measuring these specific three risks due to their importance in the legislation: according to the GDPR [44], any successful anonymization technique must provide protection against such risks. For each privacy risk, we present the design and implementation of both the attack and the evaluation phases. The risk quantification phase is common to all attacks.

5.1 Singling Out

The singling out attack is given the task to create N_A many predicates based on the synthetic data that might single out individual data records in the training dataset. As stated above, it produces guesses like: *"there is just one person in the original dataset who is male, 65 years old and lives at area 30305"*. The intuition behind this approach is that attributes (or combinations thereof) that are rare or unique in the synthetic data might also be rare or unique in the original data. Therefore, access to the synthetic data would allow for generating more meaningful predicates than uninformed guessing.

Attack Phase: For the attack phase, we introduce two algorithms, namely the univariate `PredicateFromAttribute` and the multivariate `MultivariatePredicate`, that can be used to generate the N_A many singling-out predicates (*guesses*). While the univariate algorithm creates predicates using single attributes, the multivariate algorithm relies on the combination of several attributes.

Algorithm 1 samples all unique attribute values in the synthetic dataset as predicates. For categorical attributes or in case of missing values, such unique values are values that appear only once in the dataset. For numerical continuous attributes, we use the maximum and minimum value of the respective attribute and create the predicate based on being smaller than the minimum or larger than the maximum value. The intuition behind this approach is to exploit outlier values in all the one-way marginals. Such univariate predicates are specially designed to exploit privacy leaks in pre- and post-processing, *e.g.*, when numerical values sampled from the generative models are scaled to ranges derived from the original dataset, or when high-cardinality categories such as identifiers or

Algorithm 1: Creating a univariate singling out predicates for attribute a.

```
def PredicateFromAttribute($X$, $a$):
    predicates ← [];
    if Sum ($X_{syn}[:, a]$ == $NaN$) == 1 then
        predicates + = "$a$ Is NaN";
    end
    if IsContinuous ($a$) then
        predicates + = "$a \leq \min(X_{syn}[:, a])$";
        predicates + = "$a \geq \max(X_{syn}[:, a])$";
    end
    for $v$ In Set ($X_{syn}[:, a]$) do
        if Sum ($X_{syn}[:, a]$ == $v$) == 1 then
            predicates + = "$a$ == v";
        end
    end
    return predicates;
```

addresses are preserved. By running Algorithm 1 for all attributes in a dataset, we obtain a large collection of univariate singling-out predicates. The attacker picks N_A of them at random to use them as guesses.

Algorithm 2: Create a multivariate predicate from the attributes of record $\mathbf{x}$.

```
def MultivariatePredicate($X$, attributes, $\mathbf{x}$):
    predicates ← [];
    for $a$ In attributes do
        p ← "";
        if ($\mathbf{x}[a]$ == $NaN$) then
            p ← "$a$ Is NaN";
        end
        if IsContinuous ($a$) then
            if $\mathbf{x}[a] \geq$ median ($X[:, a]$) then
                p ← "$a \geq \mathbf{x}[a]$";
            else
                p ← "$a \leq \mathbf{x}[a]$";
            end
        else
            p ← "$a$ == $\mathbf{x}[a]$";
        end
        predicates += p;
    end
    predicate ← LogicalAnd (predicates);
    return predicate
```

Algorithm 2 creates predicates as the logical combinations of univariate predicates created from randomly selected synthetic data records. It starts by drawing a random record $\tilde{\mathbf{x}}$ from the synthetic dataset and considering a random set of attributes $\{a_1, \ldots, a_d\}$. A multivariate predicate is then formulated as the logical AND of the univariate expressions derived from the values of $\tilde{\mathbf{x}}$.[4] If attribute a_i is categorical or not a number, the expression sets a_i to be equal to $\tilde{\mathbf{x}}[a_i]$. If a_i is numerical, the expression sets for values of a_i either greater or equal or smaller or equal than $\tilde{\mathbf{x}}[a_i]$. The sign of the inequality depends on whether $\tilde{\mathbf{x}}[a_i]$ is above or below the median of $X_{\text{syn}}[a_i]$, respectively. This latter condition helps creating predicates with a higher chance of singling out. The attacker evaluates each of these predicates on the synthetic dataset and adds them to the set of guesses only if they are satisfied by a single record in X_{syn}. The fraction of generated predicates by the multivariate algorithm that passes this selection depends on the dataset and the number of attributes used to generate the guesses. For the experiments and datasets presented in Section 6.3 this fraction is globally $\sim 24\%$, *i.e.*, to obtain N_A singling-out predicates, roughly $4N_A$ predicates must be generated.

We note that starting from randomly selected synthetic records and attributes ensures that the attack predicates explore the whole parameter space while not overfitting to the synthetic dataset.

To quantify the strength of the attack, we implement an algorithm that generates random predicates in order to measure the probability of creating predicates that single out an individual by chance. Such predicates are created as the joined logical AND of univariate predicates of the form: $a \; \Pi \; v$ where a is a randomly chosen attribute, Π a comparison operator[5] selected at random, and v is a value sampled uniformly from the support of $X_{\text{syn}}[:, a]$.

The randomly generated predicates from this algorithm are not evaluated on the synthetic dataset, and reach the evaluation phase of the analysis without undergoing the selection phase. An experimental evaluation of our attack strength as a function of the number of attributes is presented in Figure 7 in the Appendix. It illustrates the effect of the random predicates on the reported risk.

Evaluation Phase: For the univariate and multivariate algorithms, as well as for the naive attack, the results are sets of N_A predicates. These singling out guesses are evaluated on the original dataset to check whether they represent singling-out predicates in the original data as well.

Risk Quantification Phase: As for each of the three privacy attacks, the output of the evaluation is used for risk quantification. To derive a unique singling out risk estimate, both the univariate and multivariate attack algorithms are run, and the one with the best performance (highest risk) is chosen to provide a more conservative privacy assessment.

In contrast to the other privacy attacks, in the case of the singling out attack, care must be taken when comparing the results of the attack against the training set r_{train} and the control set r_{control}. The ability of the attack to single out a record is strongly dependent on the size of the dataset. If, as it is often the case in practice, the control dataset is smaller than the training set, the number of predicates that successfully single out in the control dataset is lower by construction than in the case of singling out in the training set. To be able to measure the true privacy risk with Equation (3) we need to know how many predicates would have singled out in a population of size N_{train}, given the number of predicates that single

[4] For example, a multivariate singling-out predicate using five attributes and generated from the Adults dataset looks like this: "race='Asian-Pac-Islander' & relationship='Husband' & education='Bachelors' & capitalloss≥1902 & country='Philippines'".

[5] $=, \neq, >, <, \geq, \leq$.

Synthetic data

Gender	Age	ZIP Code	Heart Attacks	...
...	...	...	...	...
...	...	...	...	...

External dataset A

...	Gender	ZIP Code
...	...	...
...	...	...

External dataset B

...	Age	Heart Attacks
...	...	...
...	...	...

Table 1: Setting of the linkability risk evaluation. The original external datasets A and B (bottom, left and right tables, respectively) share some attributes with the synthetic data (top) which can then be used to establish a link between datasets A and B.

out in a population of size N_{control} (where $N_{\text{control}} \leq N_{\text{train}}$). We do this by developing a model based on the Bernoulli distribution which is then fitted to the data to derive the scaling factor needed to compare r_{train} and r_{control} accounting for the different sample sizes. This model is presented in Appendix B.

5.2 Linkability

The linkability attack tries to solve the following task: *"Given two disjoint sets of original attributes, use the synthetic dataset to determine whether or not they belong to the same individual"*. We assume that there exist two (or more) external datasets A and B containing some of the attributes of a set of original data records and that these attributes are also present in the synthetic data. An example of such a scenario is depicted in Table 1.

Attack Phase: In the linkability attack, the target records of the attack are a collection T of N_A original records randomly drawn from X_{ori}. We assume that the attacker has some knowledge on the targets, *i.e.*, the values of the attributes in datasets A and B: $T[:, A]$ and $T[:, B]$. The goal of the attack is then to correctly match records of $T[:, B]$ to each record in $T[:, A]$, or vice versa.

To do so, for every record in $T[:, A]$ the attacker finds the k closest synthetic records in $X_{\text{syn}}[:, A]$. The resulting indices are $\mathbf{l}^A = (\mathbf{l}_i^A, ..., \mathbf{l}_{N_A}^A)$, where each $\mathbf{l}_i^A$ is the set of indexes of the k synthetic records that are nearest neighbours of the i^{th} target in the subspace of feature set A. The same procedure is repeated on the feature set B, resulting in the indexes $\mathbf{l}^B$ of $X_{\text{syn}}[:, B]$. To solve the nearest neighbour problem, we use a simple brute force approach using the Gower coefficient [24] to measure the distance between records. We choose this distance since it naturally supports inputs with both categorical and numerical attributes. For categorical attributes, this distance is 1 in the case of a match (or if the two values are both NaNs[6]), and 0 otherwise. For numerical attributes the distance is equivalent to the L_1 distance, with the values scaled so that $|x_i - x_j| \leq 1 \ \forall x_i, x_j \in \mathbf{x}$.

The attack procedure is then repeated using the synthetic dataset to establish links between N_A target records drawn from the control set. This results in the two sets of indexes $\mathbf{l}_{\text{control}}^A$ and $\mathbf{l}_{\text{control}}^B$. Finally, a naive attack is implemented to provide a measure probability of finding the correct link by chance. For this $\mathbf{l}_{\text{naive}}^A$ and $\mathbf{l}_{\text{naive}}^B$ are obtained by drawing indexes uniformly at random from the range $[0, n_{\text{syn}} - 1]$ where n_{syn} is the size of the synthetic dataset.

Evaluation Phase: For each of the N_A targets, we check whether both identified nearest neighbor sets share the same synthetic data record. If they do, the synthetic record allows an attacker to link together previously unconnected pieces of information about a target individual in the original dataset. The attacker scores a success for every correctly established link. The outcome $\mathbf{o}$ of this evaluation is:

$$o_i(\mathbf{l}_i^A, \mathbf{l}_i^B) = \begin{cases} 1 & \text{if } \mathbf{l}_i^A \cap \mathbf{l}_i^B \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This evaluation is performed on the outputs of the three attacks: $(\mathbf{l}^A, \mathbf{l}^B)$ for the attack on training records, $(\mathbf{l}_{\text{control}}^A, \mathbf{l}_{\text{control}}^B)$ for the attack against the control set, and $(\mathbf{l}_{\text{naive}}^A, \mathbf{l}_{\text{naive}}^B)$ for the naive attack. By default, the linkability attack is performed with $k = 1$, that is, it considers only the first nearest neighbor. Extending the search to larger values of k helps relax the definition of successful linkage by tolerating a certain degree of ambiguity. This strengthens the attack and is helpful for evaluating synthetic data where no direct one-to-one link between data records might exist.

5.3 Inference

For the inference attack, we assume that the attacker knows the values of a set of attributes (the *auxiliary information*) for some target original records. The task of the attacker is to use the synthetic dataset to make correct inferences about some *secret* attributes of the targets.

Attack Phase: The core of the inference attack is a nearest neighbor search, similar to what is done in [37]: for each target record, the attacker looks for the closest synthetic record on the subspace defined by the attributes in the auxiliary information. The nearest neighbour algorithm is the same as that discussed in Section 5.2. The values for the secret attribute of the closest synthetic record constitutes the guess of the attacker which can then be evaluated for correctness. The attack is then repeated against the targets from the control set. Finally, the probability of making a correct inference by chance is measured by implementing a naive inference attack where the attacker's guesses are drawn randomly from the possible values of the secret attribute.

Evaluation Phase: For evaluation, we say that the attacker has made a successful inference if, for a given secret attribute, the attacker's guess is correct. Comparing the guesses with the true values of the secret in the original data, we count how many times the attacker has made the correct inference. If the secret s_i is a categorical variable, a correct inference requires recovering the exact value. For numerical secrets, the inference is correct if the guess is within a configurable tolerance δ from the true value:

$$o_i(s_i, g_i, \delta) = \begin{cases} \frac{|s_i - g_i|}{s} \leq \delta & \text{if } i \text{ numerical continuous} \\ 1 & \text{if } s_1 = g_i \text{ and } i \text{ categorical} \\ 0 & \text{if } s_1 \neq g_i \text{ and } i \text{ categorical.} \end{cases} \tag{5}$$

[6] Of the three possible ways in which NaNs can be compared ("NaNs are equal to anything", "NaNs are equal to nothing", and "NaN is equal to NaN") considering only "NaNs equals to NaNs" gives a broader distribution of distances, which helps identifying close-by records, and gives more effective comparisons in the presence of to suppressed/missing values.

Note that since the same δ is applied to the main and the control attack, the choice of the particular value of δ has little impact on the results of the inference analysis.

6 EXPERIMENTAL EVALUATION

For practical evaluation, we use three tabular datasets, namely *"Adults"* [48], 1940 US Census [52] (*"US Census"* in the following), Texas Hospital Discharge Data Public Use Data [60] (*"Texas"* in the following). The Adults dataset contains records with 15 mixed-type columns, the Census dataset has 97 mixed-type columns, and the Texas Hospital dataset holds data records with 193 mixed-type columns. The records in the Adults dataset are assigned at random to the training and control dataset with an 80%-20% split (39074 training records and 9768 in the control set). The other two datasets are much larger than Adults, and we selected a subset of 50000 records chosen at random for the training dataset, and an equal number of records for the control set. To facilitate synthetic data generation and produce synthetic data with high utility, we consider *subsets* of 37 and 28 columns for the US Census and the Texas dataset, respectively. They are listed in Appendix C. The choice of which columns are included in this analysis has been based on criteria such as cardinality, number of empty rows, and the presence of other columns which are almost exact duplicates.

6.1 Linearity of Risk Estimates

This set of experiments demonstrates the ability of our Anonymeter to detect and measure various amounts of privacy leaks.

To assess our privacy risk estimates based on some "ground truth", we first evaluate them without relying on synthetic data since no synthetic data (or anonymization mechanism) can be assumed free of privacy risks. Instead, we split the original dataset in three non-overlapping populations: a training, a control, and a release set of records. The latter one plays the role of the synthetic dataset in these experiments. Since the three sets are independent, no information specific to training or control data can be found in the release set. Based on the training and the release set, we create a "leaky synthesiser". The leaky synthesizer outputs a dataset X_{syn}^{leaky} of size m composed by a fraction $f_l \in [0, 1]$ of the training data while the remaining records are taken from the release set. The fraction of leaks f_l is the main parameter of this analysis: it represents how many training records (mf_l) are directly disclosed. If $f_l = 0$ the leaky synthesizer will exclusively output data records from the release set, thus offering maximum privacy to the training records (noninterference). When $f_l = 1$ the leaky synthesizer outputs solely training records, so that privacy is maximally violated.

For these experiments, we set the number of generated predicates N_A to 2000. N_A is one of the free parameters of the analysis and its main influence is to control the statistical uncertainties on the risk estimates. Increasing N_A makes the risk estimate more robust, such that in a practical setting, N_A should be set to yield acceptably small confidence intervals.

The other main parameter of the analysis is the amount of attributes that are used to mount the attacks. This corresponds to the auxiliary information for the linkability and inference attacks, and to the number of attributes in the singling-out predicates in the multivariate singling out case. For brevity, in the following we refer to this parameter simply as auxiliary information for all cases. Increasing auxiliary information results in stronger attacks. We vary the auxiliary information from the minimum value of one attribute (two in the case of linkability) to all the available attributes (minus one, the secret, for the inference attack), thus modeling increasingly more powerful attacks.

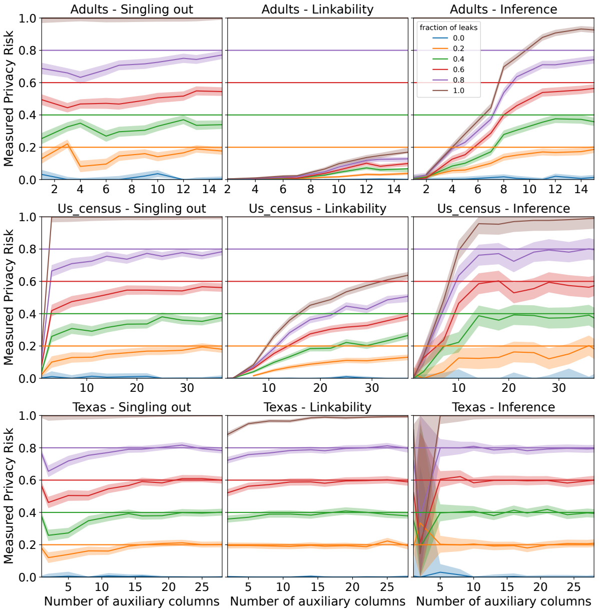

Figure 2: Estimated privacy risks as a function of the attacker's auxiliary knowledge for different amounts of artificial privacy leakage. The results from different datasets are arranged vertically: Adults dataset (top), US Census (middle), and Texas (bottom). The horizontal panels show the different privacy risks. From left to right, singling out, linkability, and inference. The shaded areas correspond to the 95% confidence interval, as measured with the methods presented in Section 4.2.

Figure 2 presents the main results of our experiments. First of all, we observe that by increasing the fraction of artificial privacy leaks—*i.e.*, by replacing a certain amount of data records in the release set by training data records—the privacy risk increases accordingly. Second, the results highlight that for all three datasets, the measured privacy risks increase linearly with an increased amount of auxiliary attributes known to the attacker. As depicted in Figure 2, as soon as the auxiliary information covers 10 to 20 attributes, the attack is able to fully exploit the whole fraction of the implanted privacy leaks, and the reported risks saturate to match the fraction of artificially introduced leaks. The linear behaviour of the risk estimates is of great importance when it comes to communicating and interpreting the results of Anonymeter: the framework matches the intuition that the reported privacy leakage grows linearly with the privacy leakage encountered in the data.

When comparing the risk estimates of the different attacks over the three datasets, we observe that the privacy risks in the Texas dataset are higher than in the other two datasets. This is likely due to the many columns with a large number of unique values: 12 (resp.,

8) of the 28 columns in the dataset have more than 100 (resp., 1000) unique values. These quasi-identifiers make attacking this dataset particularly easy. We note that the linearity between the reported risk and the amount of privacy leaks is not fully preserved for the linkability attack in the Adults and the US Census dataset (see the second column, second and third row in Figure 2). Especially for the Adults dataset, the reported linkability risks is low, ≤ 20%, even for high fractions of privacy leaks. This result is probably due to the intrinsic properties of the dataset, namely the relatively small number of columns and a prevalence of low cardinality categorical attributes. These factors result in many ambiguous links that the attacker cannot leverage.

Equally important, the reported risks are roughly zero when no privacy leaks occur. This "zero-point" of the risk measure derives directly from the fact that the reported risk corresponds to the excess of success by an attacker in comparison to the control baseline—see Equation (3). This property is of great help in practical settings, since it eliminates any possible ambiguity in the interpretation of the measured risks. Any non-zero risk reported by Anonymeter has to be seriously inspected. In particular, since the framework provides a per-column estimate of the privacy risks, it helps identify which attributes, alone or in combinations, exhibit the largest risks (and thereby expose a lot of information on the original data). As an example, Figure 8 in the Appendix provides an overview on how the inference risk varies depending on which secret attribute the attacker tries to infer. This information provided by the framework can support the decision-making around synthetic data release: based on the setup, high-risk columns could be removed or perturbed to decrease the privacy risk before releasing the data.

6.2 Synthetic Data Generation

This section describes how we generate the synthetic data for privacy evaluation using CTGAN [66]. We rely on the original implementation of the algorithm[7], as well as a DP version of the model.[8]

6.2.1 Synthetization Setup. For the synthetization with CTGAN, we use the default hyperparameters as specified in the documentation.[9] Training DPCTGAN requires more tuning, as we experimentally verified that no usable synthetic data is obtained when running with default parameters. Through a hyperparameter search, we identified the best learning rates for the generator and discriminator to be 10^{-5} and 10^{-3}, respectively. For each generator update, the discriminator is updated five times, as proposed in [6]. Moreover, as for CTGAN, we use the log frequency for the categorical levels in conditional sampling (`log_frequency` is `True`). Finally, for all datasets, we use a batch size of 64.

DP Hyperparameters: Since DPCTGAN cannot be applied to continuous data, the data must be pre-processed before training the model. For details on the pre-processing and its impact on the DP guarantees, see Appendix D. We set the noise multiplier in the DPCTGAN training to 0.5 and the clip norm to 10. With these hyperparameters, we run DPCTGAN for different numbers of epochs (10, 20, 50, 100, 200, respectively) and select the synthetic dataset with the highest utility (see Section 6.2.2 for an overview on how the utility is computed). For Adults and US Census the highest utility data results from 100 training epochs, while for Texas 10 epochs yield the best utility. The rightmost column in Table 3 and Figure 9 in the Appendix provide an overview on the utility of each generated dataset. The achieved privacy levels depend on the batch size, on the noise multiplier and on the number of training epochs, and are $\varepsilon = 32$ for Adults, $\varepsilon = 28$ for US Census and $\varepsilon = 11$ for the Texas dataset. Note that, even though from a theoretical point of view such large values for the privacy budgets might not yield meaningful privacy guarantees, it has been shown that the use of DP, even in the large ε regime, still results in increased privacy protection in practice [38].

6.2.2 Measuring Utility. A utility score between 0 and 100 is attributed to each synthetic dataset as a measure of its data quality. Such score measures three different aspects of the data: marginal distribution similarity, pairwise dependency similarity, and similarity in terms of query counts. Each aspect is quantified separately and the utility score is obtained by averaging them. The marginal distribution similarity is measured using the Jensen-Shannon divergence [19] between original and synthetic data if the variable is categorical, or the Kolmogorov-Smirnov statistic [31] for continuous variables. If the data consists of multiple variables, the marginal score is an average over the single variable scores. The pairwise scores measure how similar correlation, correlation ratio, or mutual information are for a pair of variables in the original and synthetic data. The score is computed based on the absolute error between the original and synthetic statistics. If more than one pair of variables is examined, the pairwise score is averaged over all pairs. For the query counts similarity, we create a number of random queries and measure how many rows in the original and synthetic data match each query. We then compute the score as the correlation coefficient between the original and synthetic query response sizes. Figure 9 in Appendix E summarizes the utility scores obtained for each dataset and synthetization algorithm. Note how the DP version of the same algorithm gives lower utility.

6.3 Evaluating Privacy Risks of Synthetic Data

In this set of experiments, we apply Anonymeter to the evaluation of the three privacy risks on the generated synthetic data.

6.3.1 Evaluation Setup. For a comprehensive evaluation of the privacy risks of the synthetic data we explore a variety of settings for the attacks algorithms, similar to the experiments of Section 6.1. In particular, different amount of auxiliary information is evaluated. For each setting, we use $N_A = 2000$, repeat the analysis twice and then average the results over both runs.

For the singling out evaluation we execute both the univariate and multivariate algorithm. For the latter, we repeat the attack with a number of attributes varying between 3 and 12. The final privacy risk is derived from the most successful attack. For linkability different amounts of auxiliary information are considered, ranging from two columns to the maximum number of columns d in the datasets. For each of the setups, the privacy risks are computed by considering the k nearest neighbors within the synthetic dataset with $k \in \{1, 2, \ldots, 10\}$. Considering $k > 1$ neighbours makes it

[7] https://github.com/sdv-dev/SDV

[8] https://github.com/opendp/smartnoise-sdk/tree/main/synth

[9] https://sdv.dev/SDV/user_guides/single_table/ctgan.html#how-to-modify-the-ctgan-hyperparameters

Dataset	*Method*	**Singling out**	**Linkability**	**Inference**
Adults	*CTGAN*	100%	68%	86%
	DPCTGAN	90%	52%	77%
Texas	*CTGAN*	84%	79%	86%
	DPCTGAN	24%	41%	60%
US Census	*CTGAN*	90%	82%	88%
	DPCTGAN	88%	73%	85%

Table 2: Fraction of risk estimates which resulted in a valid attack, *i.e.*, for which the success rate of the main privacy attack is above that of the random baseline algorithm. Percentages are given over the 120, 480, and 3460 computed privacy estimates for singling out, likability, and inference, respectively.

possible to compensate for the fact that synthetic data might not keep one-to-one links between data records, as described in Section 5.2. To evaluate inference we mount a set of attacks against each column in the dataset using between 1 and $d-1$ other columns as auxiliary information.

In total, the above settings result in 120, 480, and 3460 computed privacy estimates for the singling out, linkability, and inference risks, respectively, each corresponding to a different combination of attack settings, dataset, and generative model.

6.3.2 Quality Cuts. In the following, we present the results only for those settings for which the main privacy attack is stronger than the random baseline (see Section 4.2, *valid attacks*). The fraction of such settings is presented in Table 2 for the synthetic datasets generated by CTGAN and DPCTGAN. Furthermore, we exclude attack runs from the results where the success rate r_c in the control attack is greater than 0.9, *i.e.*, attacks where the population-level privacy leakage is very high. In such cases, any additional privacy risk induced through the synthetic dataset that appears when $r_t \geq r_c$ must be relatively small. Accurately measuring such small effects often requires larger values of N_A than those used in our analysis due to the underlying statistical randomness. Applying these criteria led to excluding approximately 7% and 22% of the results on the inference analysis for the Texas and US census datasets, respectively. Results of the singling out and linkability attacks, and any results for the Adults dataset are not affected by this cut.

6.3.3 Evaluation Results. We evaluate the privacy risks for all three attacks on the synthetic datasets generated by CTGAN and DPCTGAN. Figure 3 shows that, for the Adults and the US Census datasets, the synthetic data generated with DP guarantees exhibit smaller privacy risks than their non-DP counterpart. When considering the ratios of different privacy risk scores between the non-DP and DP synthetic data for Adults and US Census, we observe that the ratio for inference (Adults: 1.52, US Census 2.18) and linkability (Adults: 1.82, US Census: 4.53) are higher than for singling out (Adults: 1.24, US Census 1.25). This suggests that DP mitigates the former two privacy risks to a higher extent than the singling out risk. We attribute the cause of such difference to the fact that, regardless of how the generative model is trained, the values of the categorical attributes in the synthetic dataset are sampled from those present in the original dataset (see [50] and the discussion in Appendix D for details). Of the three privacy attacks, singling out is the one which puts the greater emphasis on the exact values of these attributes.

However, for the Texas dataset, despite having a smaller DP level ε, the mitigation of the privacy risk offered by DP is not as effective. We suspect that this is either a result of the DP guarantees not being protective enough, or of the pre-processing described in Appendix D which preserves too detailed information on the respective categorical diagnostic codes in the dataset, or of a combination of both effects together with the intrinsic properties of that particular dataset (number of attributes, their distributions and their ranges).

Furthermore, we see that risks reported for linkability are low over all three datasets. This supports the common intuition that synthetic data indeed does not preserve one-to-one links between data records.

We also display the privacy risks as a function of the auxiliary information available to the attacker. For linkability and inference risk, the amount of auxiliary information is given by the number of auxiliary columns, whereas for singling out, the amount of auxiliary information corresponds to the number of attributes used to create the predicates. Figure 11 in Appendix E shows the respective risks for each attack.

Figure 10 in Appendix E depicts the linkability risk not only as a function of the number of auxiliary columns, but also as a function of the number k of nearest neighbors used to determine the linkage for both CTGAN and DPCTGAN. Although a clear trend can be observed, showing an increase in the reported linkability risk as the number of auxiliary information columns grows, the effect of including a higher number of neighbors in the nearest neighbor search is less consistent. This further supports the thesis that synthetic data indeed helps thwart linkability attacks.

As seen in Section 6.2.2, synthetic data generated with CTGAN has more utility than those from DPCTGAN. Table 3 compares the changes in the measured privacy risks among data generated with the respective generator models. In both the Adults and US Census datasets, we observe that the data generated with DP guarantees exhibit lower privacy risks. This might also be explained by synthetic data with higher utility more closely matching the original data, thereby exposing an increased risk to the various privacy attacks. In the Texas dataset, however, the privacy risks are similar for both CTGAN and DPCTGAN and they remain low even if CTGAN offers better utility.

6.4 Quantitative Comparison with Previous Work

In this experiment, we compare our methods with [56] which we refer to as *Synthetic Data Release* (SDR [47]).

To evaluate privacy risks, we rely on the "leaky synthesizer" methodology presented in Section 6.1. We instantiate six different leaky synthesizers, each one corresponding to a different f_l between 0 and 1 and use SDR to evaluate their linkability and inference privacy risks, leaving all the analysis settings unchanged from those presented in [56]. The analysis uses the Texas dataset included in the SDR code release [47], which contains a smaller set of columns than those used in the rest of this paper (18 instead of 28). The leaky datasets created for evaluating SDR are saved to disk, and are then evaluated using Anonymeter. For this evaluation, we vary

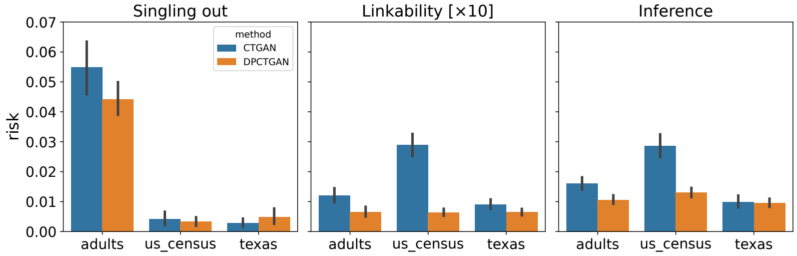

Figure 3: Overview of the privacy risks scores for different datasets and synthetization methods. For the sake of visualization, the linkability risk score has been increased by a factor ten.

Dataset	*Method*	**Singling out**	**Linkability**	**Inference**	**Utility**
Adults	*CTGAN*	0.0548 ± 0.0134	0.0012 ± 0.0003	0.0161 ± 0.0031	88
	DPCTGAN	0.0441 ± 0.0084	0.0007 ± 0.0002	0.0106 ± 0.0021	79
Texas	*CTGAN*	0.0028 ± 0.0020	0.0009 ± 0.0002	0.0098 ± 0.0030	88
	DPCTGAN	0.0049 ± 0.0037	0.0007 ± 0.0002	0.0095 ± 0.0019	71
US Census	*CTGAN*	0.0042 ± 0.0032	0.0029 ± 0.0006	0.0286 ± 0.0057	74
	DPCTGAN	0.0033 ± 0.0022	0.0006 ± 0.0002	0.0131 ± 0.0023	63

Table 3: Measured risks and utility scores for the different attack types, datasets, and generative models. The reported risks are obtained by averaging over all the attack settings (see Section 6.3.1) and the reported uncertainties shows the 95% confidence interval computed via bootstrapping.

the number of auxiliary information used to mount the attack between 10 and the maximum number of available attributes. For the linkability attack, we consider linkability at the level of the closest neighbour only (see Section 4.2), as done for the experiments presented in Section 6.1.

As described in Section 2, the metric used in SDR is the Privacy Gain (PG). In our work, we instead use the privacy risk (R). These two quantities are defined differently and therefore care must be taken when comparing them: in particular, one should not assume that $PG = x$ has the same meaning as $1 - R = x$. However, both serve the same purpose and one should expect that a decreasing $1 - R$ would be matched by a reduced PG—and vice versa, so that a high level comparison is indeed possible. Figure 4 presents the results of such a comparison between SDR and Anonymeter for different values of f_l.

For the linkability attack (top row), the success of SDR is strongly influenced by the selection of the target records. For randomly selected targets, the PG is always high and shows only a mild dependency on f_l. When the evaluation targets selected outlier points (the same as in [56]) we see a step-like variation of PG with the f_l: $PG \sim 1$ for $f_l \lesssim 0.5$ and $PG \sim 0$ for $f_l \gtrsim 0.5$. For the inference attack, the performance of SDR varies greatly depending on whether the secret that is being guessed is categorical (RACE, middle row) or continuous (LENGTH_OF_STAY, bottom row). In the former case, PG goes from ~ 0.4 to $\lesssim 0$ across the range of privacy leaks, again showing a sharp drop at $f_l \sim 0.5$. When the secret is continuous however, PG remains ~ 1 for all values of f_l, even if the synthesizer leaks the entire training data.

In contrast, the results of Anonymeter, for both inference and linkability, vary linearly with f_l between $1 - R = 1$ for $f_l = 0$ and $1 - R = 0$ for $f_l = 1$, as expected. This result aligns with the experiments presented in Section 6.1 and confirms that Anonymeter outputs "well behaved" privacy metrics.

The results support our claim that Anonymeter is a robust way to capture various degrees of privacy leaks. Perhaps more importantly, Anonymeter never fails to report $R > 0$ when privacy leaks are present. Anonymeter also offers better scalability to large datasets than SDR. This is because SDR requires training dozens of models and generating thousands of synthetic datasets, which restricts the practical usability of the method to datasets with a maximum of tens of thousands of data records. For example, for each of the 6 leaky synthesizers, SDR created 2550 and 1250 synthetic datasets to analyze inference and linkability, respectively. In contrast, Anonymeter only requires one realization of the synthetic dataset and can evaluate the privacy of large synthetic datasets with millions of rows within less than one day of computing time using rather inexpensive general-purpose virtual machines with 64 virtual CPUs. For the analysis just described, SDR was 10 times slower than Anonymeter in wall clock time. Such performance difference would have been even more pronounced for a "real" synthesizer using time-consuming training algorithms.

7 DISCUSSION AND FUTURE DIRECTION

The evaluation of Anonymeter on singling out, linkability, and inference risks highlights the effectiveness of our framework to provide a coherent assessment of legally meaningful privacy metrics. Not only does Anonymeter allow us to analyze general privacy leakage as a function of the attacker's power, but at the same time it helps identify concrete privacy violations in the synthetic datasets. In particular, Anonymeter significantly outperforms existing frameworks for privacy evaluation of synthetic data in both the detection

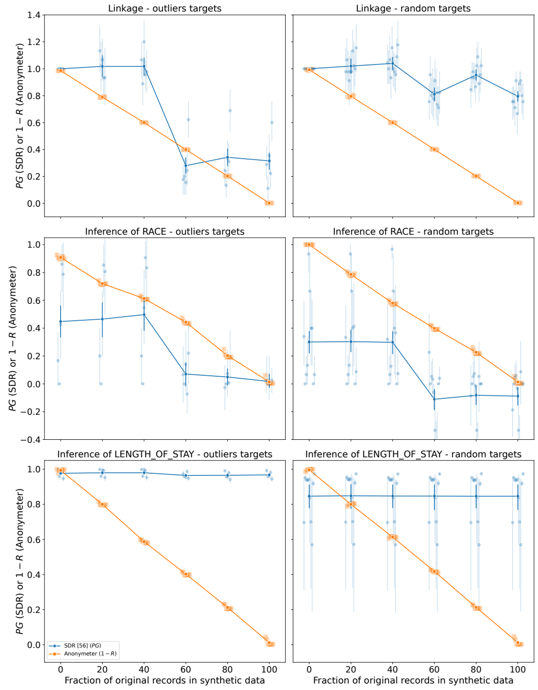

Figure 4: Privacy gain PG (blue) and $1-R$ (orange) as a function of the amount of privacy leaks. The plots on the left side are obtained from the analysis of hand-selected target records (the same as those chosen in [56]), those on the right uses 10 randomly chosen targets. The blue points report the PG of each target. The orange squares present the privacy risk estimated for various amounts of auxiliary information. Although the two metrics are reported on the same axis, care must be taken when comparing their values. See text for details.

of privacy leakage and computational complexity. This is a crucial step on the way towards leveraging the full potential of using synthetic data while keeping track of the privacy implications.

Moreover, the modular nature of Anonymeter facilitates the future integration of new and potentially stronger attacks for evaluating the three privacy risks analyzed in this work. Privacy attacks that evaluate other aspects of privacy, such as membership inference, can also be integrated. This flexibility allows the framework to adapt to and to meet future requirements from emerging and changing privacy regulations.

Another advantage of Anonymeter is that it separates the evaluation of the success rate of the privacy attacks from the calculation of the reported privacy risks. Due to the statistical nature of Anonymeter's risk quantification phase (where each attack simply yields a boolean array), the privacy risk is deduced from the main attack and the baselines, which provide the necessary context for turning attack success into expressive privacy risks.

Since Anonymeter treats the synthetic data generation mechanism as a black box and solely utilizes the generated dataset, the framework can be used on datasets anonymized with other methods. Anonymeter can even be applied to an original dataset to identify individual data records with high privacy risks. This assessment can, among others, serve as a pre-filtering mechanism to identify—and, for example, remove—high-risk data records before training a generative model on the original data. This can lead to reduced privacy risks for the generated synthetic dataset.

In this work, we apply Anonymeter solely to perform privacy assessments of entire datasets. However, the framework can be directly applied to quantify the privacy risks associated with a particular individual (or subgroup of individuals) in the dataset. Therefore, the generation of guesses in the attack phase just has to specify the respective individual(s) instead of using random targets. To provide a more fine-grained risk assessment over the entire dataset, the target selection in our framework could also rely on identifying targets with high privacy risks and selecting these for generating the guesses. This would help approximate the upper bound on privacy leakage in the dataset more closely than an assessment over randomly chosen targets.

Future work could also look into the extension of Anonymeter for assessing additional trustworthiness aspects in synthetic datasets, such as their fairness, or existing biases towards certain subgroups of individuals, thereby supporting a joint assessment of these different aspects together with utility and privacy of the data.

8 CONCLUSION

Synthetic data has the potential to mitigate existing tensions between the need to share and utilize sensitive datasets and the privacy concerns of the individuals whose data is included in these datasets. The fact that the actual privacy leakage in such datasets is hard to quantify, hinders leveraging the high potential of the data. To close this gap, we propose Anonymeter, a statistical framework for jointly quantifying different privacy risks in synthetic datasets. Within this framework, we implement concrete attacks to measure the privacy risks of singling out, linkability, and inference—the three risks that anonymization methods must mitigate to be legally compliant with existing privacy legislation. Our work is the first to propose practical attacks directly measuring the singling out and linkability risks posed by the release of a synthetic dataset. The experiments we conduct highlight that Anonymeter is able to report privacy risks in a coherent and fine-grained manner, making the framework a valuable resource for identifying privacy leakage and quantifying the corresponding risks. We also demonstrate that Anonymeter significantly outperforms prior work both in finding privacy leaks as well as in usability. By making Anonymeter available as an easy to use open source library, we hope to contribute to a more mature and privacy-aware usage of synthetic data. In particular, we want to help avoid dangerous situations in which, due to the lack of usable privacy analysis, synthetic data is simply "trusted" to be private.

ACKNOWLEDGMENTS

We thank the authors of [56] for making their work easy to reproduce and to interact with, and the SDV and Smartnoise teams for providing implementations of CTGAN. We thank Nicola Vitacolonna for the help and support during the revision process. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Oct 2016). https://doi.org/10.1145/2976749.2978318

[2] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2018. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 510–526.

[3] Gergely Acs, Claude Castelluccia, and Daniel Le Métayer. 2016. Testing the robustness of anonymization techniques: acceptable versus unacceptable inferences. In *The Brussels Privacy Symposium* . brussels, Belgium. https://hal.inria.fr/hal-01399858

[4] Aircloak. 2017. Aircloak Challenge version 7. https://aircloak.com/wp-content/uploads/Aircloak-Challenge.pdf

[5] Christer Andersson and Reine Lundin. 2007. On the fundamentals of anonymity metrics. In *IFIP International Summer School on the Future of Identity in the Information Society*. Springer, 325–341.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. https://doi.org/10.48550/ARXIV.1701.07875

[7] Erman Ayday and Mathias Humbert. 2017. Inference attacks against kin genomic privacy. *IEEE Security & Privacy* 15, 5 (2017), 29–37.

[8] Steven M Bellovin, Preetam K Dutta, and Nathan Reitinger. 2019. Privacy and synthetic datasets. *Stan. Tech. L. Rev.* 22 (2019), 1.

[9] Franziska Boenisch, Reinhard Munz, Marcel Tiepelt, Simon Hanisch, Christiane Kuhn, and Paul Francis. 2021. Side-channel attacks on query-based data anonymization. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 1254–1265.

[10] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2018. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. *arXiv e-prints*, Article arXiv:1802.08232 (Feb. 2018), arXiv:1802.08232 pages. arXiv:1802.08232 [cs.LG]

[11] Aloni Cohen. 2022. Attacks on Deidentification's Defenses. https://doi.org/10.48550/ARXIV.2202.13470

[12] Aloni Cohen and Kobbi Nissim. 2020. Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences* 117, 15 (2020), 8344–8352.

[13] Aloni Cohen and Kobbi Nissim. 2020. Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences* 117, 15 (2020), 8344–8352.

[14] Tech Crunch. 2022. *The market for synthetic data is bigger than you think*. https://techcrunch.com/2022/05/10/the-market-for-synthetic-data-is-bigger-than-you-think/

[15] Alberto Dainotti, Antonio Pescapé, Pierluigi Salvo Rossi, Francesco Palmieri, and Giorgio Ventre. 2008. Internet traffic modeling by means of Hidden Markov Models. *Computer Networks* 52, 14 (2008), 2645–2662.

[16] Cynthia Dwork. 2006. Differential Privacy. (2006).

[17] Dwork, Cynthia and Roth, Aaron. 2013. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* 9 (01 2013). https://doi.org/10.1561/0400000042

[18] Khaled El Emam, Lucy Mosquera, and Jason Bass. 2020. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *J Med Internet Res* 22, 11 (16 Nov 2020), e23139. https://doi.org/10.2196/23139

[19] Bent Fuglede and Flemming Topsoe. 2004. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.* IEEE, 31.

[20] Benjamin Fung, ke Wang, Rui Chen, and Philip Yu. 2010. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Comput. Surv.* 42 (06 2010). https://doi.org/10.1145/1749603.1749605

[21] Gartner. 2022. *Gartner Unveils Top Predictions for IT Organizations and Users in 2022 and Beyond.* https://www.gartner.com/en/newsroom/press-releases/2021-10-19-gartner-unveils-top-predictions-for-it-organizations-and-users-in-2022-and-beyond

[22] Gartner. 2022. *Innovation Insight for Synthetic Data.* https://www.gartner.com/en/documents/4011129

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[24] J. C. Gower. 1971. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27, 4 (1971), 857–871. http://www.jstor.org/stable/2528823

[25] Javier Herranz, Jordi Nin, Pablo Rodríguez, and Tamir Tassa. 2015. Revisiting distance-based record linkage for privacy-preserving release of statistical datasets. *Data & Knowledge Engineering* 100 (2015), 78–93. https://doi.org/10.1016/j.datak.2015.07.009

[26] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.

[27] Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *arXiv e-prints*, Article arXiv:1312.6114 (Dec. 2013), arXiv:1312.6114 pages. arXiv:1312.6114 [stat.ML]

[28] Aditya Kuppa, Lamine Aouad, and Nhien-An Le-Khac. 2021. Towards Improving Privacy of Synthetic DataSets. In *Privacy Technologies and Policy*, Nils Gruschka, Luís Filipe Coelho Antunes, Kai Rannenberg, and Prokopios Drogkaris (Eds.). Springer International Publishing, Cham, 106–119.

[29] Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. 2013. Membership privacy: a unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 889–900.

[30] Pei-Hsuan Lu, Pang-Chieh Wang, and Chia-Mu Yu. 2019. Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics* (Seoul, Republic of Korea) *(WIMS2019)*. Association for Computing Machinery, New York, NY, USA, Article 16, 6 pages. https://doi.org/10.1145/3326467.3326474

[31] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[32] Luise Mehner, Saskia Nuñez von Voigt, and Florian Tschorsch. 2021. Towards Explaining Epsilon: A Worst-Case Study of Differential Privacy Risks. In *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. 328–331. https://doi.org/10.1109/EuroSPW54576.2021.00041

[33] Ofer Mendelevitch and Michael D. Lesh. 2021. Fidelity and Privacy of Synthetic Medical Data. arXiv:2101.08658 [cs.LG]

[34] Martin M. Merener. 2012. Theoretical Results on De-Anonymization via Linkage Attacks. *Trans. Data Privacy* 5, 2 (aug 2012), 377–402.

[35] Sasi Kumar Murakonda and Reza Shokri. 2020. ML Privacy Meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning.

[36] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125.

[37] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-Anonymization of Large Sparse Datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP '08)*. IEEE Computer Society, USA, 111–125. https://doi.org/10.1109/SP.2008.33

[38] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. 2021. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. *CoRR* abs/2101.04535 (2021). arXiv:2101.04535 https://arxiv.org/abs/2101.04535

[39] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papemoti, and Nicholas Carlin. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 866–882.

[40] BO Ngoko, H Sugihara, and T Funaki. 2014. Synthetic generation of high temporal resolution solar radiation data using Markov models. *Solar Energy* 103 (2014), 160–170.

[41] Office of the Privacy Commissioner of Canada. 2019. The Personal Information Protection and Electronic Documents Act (PIPEDA). https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/, last accessed on: Dec. 2nd 2021.

[42] Iyiola E. Olatunji, Jens Rauch, Matthias Katzensteiner, and Megha Khosla. 2021. A Review of Anonymization for Healthcare Data. arXiv:2104.06523 [cs.CR]

[43] Bristena Oprisanu, Georgi Ganev, and Emiliano De Cristofaro. 2021. On Utility and Privacy in Synthetic Genomic Data. *arXiv e-prints*, Article arXiv:2102.03314 (Feb. 2021), arXiv:2102.03314 pages. arXiv:2102.03314 [q-bio.GN]

[44] Article 29 Data Protection Working Party. 2014. Opinion 05/2014 on anonymisation techniques. (2014). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[45] Youyang Qu, Shui Yu, Jingwen Zhang, Huynh Thi Thanh Binh, Longxiang Gao, and Wanlei Zhou. 2019. GAN-DP: generative adversarial net driven differentially privacy-preserving big data publishing. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

[46] Jerome P Reiter. 2002. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* 18, 4 (2002), 531.

[47] Synhtetic Data Release. last accessed Sept. 2022 (latest commit is eba9a43). Code accompanying the ublication of Stadler et. al. 2020. https://github.com/spring-epfl/synthetic_data_release

[48] UCI Machine Learning Repository. 1996. Adults dataset. (1996). https://archive.ics.uci.edu/ml/datasets/Adult

[49] Luc Rocher, Julien Hendrickx, and Yves-Alexandre Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10 (07 2019). https://doi.org/10.1038/s41467-019-10933-3

[50] Luke Rodriguez and Bill Howe. 2018. Privacy-Preserving Synthetic Datasets Over Weakly Constrained Domains. https://doi.org/10.48550/ARXIV.1808.07603

[51] Donald B Rubin. 1993. Statistical disclosure limitation. *Journal of official Statistics* 9, 2 (1993), 461–468.

[52] Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. 2018. IPUMS USA: Version 8.0 Extract of 1940

Census for U.S. Census Bureau Disclosure Avoidance Research. *IPUMS* (2018). https://doi.org/10.18128/D010.V8.0.EXT1940USCB

[53] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. 2011. Quantifying location privacy. In *2011 IEEE symposium on security and privacy*. IEEE, 247–262.

[54] Geoffrey Smith. 2009. On the Foundations of Quantitative Information Flow. 288–302. https://doi.org/10.1007/978-3-642-00596-1_21

[55] Congzheng Song and Vitaly Shmatikov. 2019. Overlearning Reveals Sensitive Attributes. *arXiv e-prints*, Article arXiv:1905.11742 (May 2019), arXiv:1905.11742 pages. arXiv:1905.11742 [cs.LG]

[56] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2020. Synthetic Data – Anonymisation Groundhog Day. *arXiv e-prints*, Article arXiv:2011.07018 (Nov. 2020), arXiv:2011.07018 pages. arXiv:2011.07018 [cs.LG]

[57] State of California Department of Justice. 2018. California Consumer Privacy Act (CCPA). https://oag.ca.gov/privacy/ccpa https://oag.ca.gov/privacy/ccpa, last accessed on: Dec. 2nd 2021.

[58] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* 671, 2000 (2000), 1–34.

[59] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[60] Center for Health Statistics Texas Department of State Health Services. 2005. Texas Hospital Discharge Data Public Use Data. (2005). https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm

[61] The European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union (OJ L 119 4.5.2016, p. 1)* (2016). https://gdpr-info.eu

[62] Rob Toews. 2022. *Synthetic Data Is About To Transform Artificial Intelligence.* https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=726046797523

[63] Isabel Wagner and David Eckhoff. 2018. Technical Privacy Metrics: A Systematic Survey. *ACM Comput. Surv.* 51, 3, Article 57 (jun 2018), 38 pages. https://doi.org/10.1145/3168389

[64] Andrew White. 2021. *By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated.* https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/

[65] Edwin B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212. https://doi.org/10.1080/01621459.1927.10502953 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1927.10502953

[66] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *arXiv preprint arXiv:1907.00503* (2019).

[67] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. 2019. Assessing privacy and quality of synthetic health data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse.* 1–4.

[68] Jim Young, Patrick Graham, and Richard Penny. 2009. Using Bayesian networks to create synthetic data. *Journal of Official Statistics* 25, 4 (2009), 549.

[69] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv e-prints*, Article arXiv:1611.03530 (Nov. 2016), arXiv:1611.03530 pages. arXiv:1611.03530 [cs.LG]

[70] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)* 42, 4 (2017), 1–41.

A UTILITY OF SYNTHETIC DATA AND SIZE OF CONTROL DATASET

Anonymeter requires the user to set aside a small portion of the original data for the control dataset. We perform an ablation study to determine how the reduced size of the training set impacts utility. In this experiment, we synthesize the three datasets using CTGAN and the same setup as described in Section 6.2.1. For each of the three datasets, the training set is made of a varying fraction (from 10% to 99%) of original records. A synthetic dataset with the same number of samples as the original data is then created starting from the different sized training sets, and its utility is evaluated using the method proposed in Section 6.2.2, comparing it to the entire original dataset. The entire process is repeated three times and the results aggregated to account for the randomness involved in both training and sampling a generative model.

The results are presented in Figure 5. As expected, there is a degradation in utility with decreasing size of the training data. However, this effect is mild, especially if one considers that the control dataset needs to contain just as many records as the number of attacks ($N_{attacks}$) that will be performed during the evaluation. A few thousands records in the control set is generally enough to have small statistical uncertainties on the measured risks. The $N_{attacks}$ = 2000 as used in this work corresponds to a relative size of the control set of 4%, that is, 96% records can be used for training. As visible in Figure 5, there is no measurable difference in utility for such small control size fractions.

To conclude, our ablation study shows that, if the control dataset is relatively small (*i.e.*, less than 20% of the total), utility is hardly affected. The requirements of Anonymeter to have a control dataset does not represent a limitation to its practical use. The situation might be different in case of very small datasets or if the focus is on rare records (for example, in the case of anomaly detection). For both these cases, however, the use of synthetic data can be very problematic.

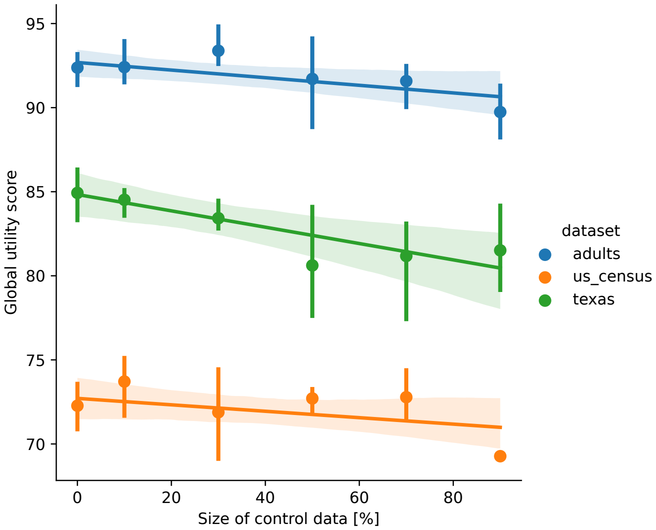

Figure 5: Measured utility as a function of the size of the control dataset for the Adults (blue), US Census (orange), and Texas (green) dataset.

B MODELING SINGLING OUT SUCCESS FOR DIFFERENT POPULATION SIZES

The success of a singling out attempt depends on the size of the target population. A birth date will likely single out an individual in a sample of 365 people, but not when considering all the inhabitants of a big city. In this section we present a model that allows us to compare the singling out success rates r_{train} and r_{control} in the general case in which the training set is larger than the control set.

More formally, let N_{train} and N_{control} be the number of records in the training and control set, respectively (generally $N_{\text{control}} < N_{\text{train}}$). The attacker builds N_A predicates generated from the synthetic dataset: m_{train} of these isolate a record in the train dataset, and m_{control} isolate a record in the control dataset. We want to derive $\tilde{m}_{\text{control}}$, the number of predicates that *would* isolate records in the control set, if the control set had N_{train} many records.

The probability that a predicate isolates a record in a population depends on the weight w of the predicate, *i.e.*, the probability that it is satisfied by exactly one record sampled from the data generating distribution $\mathcal{D}$. In particular, for a population of size n, this probability is given by [13]:

$$P(w, n) = nw(1 - w)^{n-1}. \tag{6}$$

As w depends on the distribution $\mathcal{D}$, which is unknown, Equation 6 cannot be used directly to compensate for the different size of the control set. However, the predicates we evaluate are fixed (generated by the attack) and so are their weights, since they are fully defined by the query and statistic of the population of interest. Moreover, these predicates single out in the synthetic dataset (which has size N_{train}), so their weights should be small, and of the order of $1/N_{\text{syn}}$. To focus on this low-weight regime we integrate $P(n, w)$ in the range $[0, w_{\text{eff}}]$. Our model then becomes:

$$S(n) = A \int_0^{w_{\text{eff}}} P(w, n) dw. \tag{7}$$

The parameter w_{eff} can be interpreted as the "effective weight" of the set of attack predicates, while A is a normalization constant to convert the probability into counts. To determine these parameters for the specific dataset and attack predicates, Anonymeter evaluates the attack predicates on a set of samples of control records of different sizes, in order to measure how the number of successful singling out predicates changes with the dataset size n. The evaluator automatically samples 10 different values for n from range $[1000, N_{\text{control}}]$, and for each it repeats the measurement on five different samples of control records, selected at random without replacement from the control set. This gives enough data points (five for each of the ten values of n) to fit our simple two parameter model with confidence. Once the parameters of $S(n)$ have been fitted, we can derive $\tilde{m}_{\text{control}}$ as:

$$\tilde{m}_{\text{control}} = \frac{S(N_{\text{train}})}{S(N_{\text{control}})} m_{\text{control}}. \tag{8}$$

In Figure 6 we present the results of a set of experiments done to evaluate the quality of our statistical model of the singling out success (Equation 7) which underpins our correction factor (Equation 8). The datasets used for the evaluation are formally introduced in Section 6. In all cases, the original dataset is split into two disjoint sets, called A and B. From set A, N_A singling out predicates are derived using the methods presented above. The attack predicates are then evaluated on different subsets of B. By varying the sizes of these subsets, we can measure how the the number of successful singling out predicates varies with the size of the target population. Figure 6 shows that the fitted statistical model $S(n)$ (Equation 7) provides very good agreement with the measured data points.

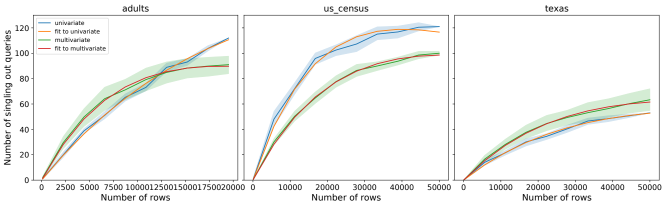

Figure 6: Number of successful singling out predicates as a function of the dataset size together with best fit model $S(n)$. The three panels refers to the different datasets: Adults (left), US census (middle), Texas (right).

C COLUMN SUBSETS INCLUDED IN THE ANALYSIS

For the Texas and US Census datasets, the following columns have been used throughout all experiments presented in this work:

Texas dataset: ADMIT_WEEKDAY, APR_DRG, APR_MDC, CERT_STATUS, COUNTY, ETHNICITY, FIRST_PAYMENT_SRC, ILLNESS_SEVERITY, LENGTH_OF_STAY, PAT_AGE, PAT_COUNTRY, PAT_STATE, PAT_STATUS, PAT_ZIP, PRINC_DIAG_CODE, PROVIDER_NAME, PUBLIC_HEALTH_REGION, RACE, RISK_MORTALITY, SEX_CODE, SOURCE_OF_ADMISSION, TOTAL_CHARGES, TOTAL_CHARGES_ACCOMM, TOTAL_CHARGES_ANCIL, TOTAL_NON_COV_CHARGES, TOTAL_NON_COV_CHARGES_ACCOMM, TOTAL_NON_COV_CHARGES_ANCIL, TYPE_OF_ADMISSION

US Census dataset: AGE, AGEMARR, AGEMONTH, BPL, CHBORN, CITIZEN, EDUC, ELDCH, EMPSTAT, FAMSIZE, HIGRADE, HISPAN, HRSWORK1, INCWAGE, IND1950, MARST, MBPL, MOMRULE_HIST, MTONGUE, NATIVITY, NCHILD, NSIBS, OCC1950, POPLOC, POPRULE_HIST, RACE, RELATE, RELATED, SEX, SPLOC, SPRULE_HIST, STEPMOM, STEPPOP, SUBFAM, UOCC, WKSWORK1, YNGCH

D DIFFERENTIAL PRIVACY AND DATA PRE-PROCESSING

Since DPCTGAN cannot be applied to continuous data, we encode the continuous columns of all datasets by binning them into 50 equally spaced bins, respectively. After synthetization, continuous values are obtained from each synthetic bin label by drawing random numbers from a uniform distribution spanning the interval defined by the corresponding bin edges. Note that even though the data synthetization with DP-CTGAN incorporates DP guarantees, the pre-processing itself does not. This is due to the ranges from which the bins are derived being obtained from the private data itself. This problem is common to many open source data synthesizers [56].

We note, however, that the problem is not limited to the continuous domain and affects also categories. Categorical data is also sampled from distributions whose support is derived from the input dataset itself. This leads to a violation of the DP guarantee which follows from the bad bins theorem [50]: *"There does not exist a differentially private mechanism that only returns elements from the active domain."*. This means that one cannot obtain DP for synthetic data if the synthesizer can only generate values that are actually present in the original data instead of also allowing for the presence of values which are possible but not present in the original dataset.

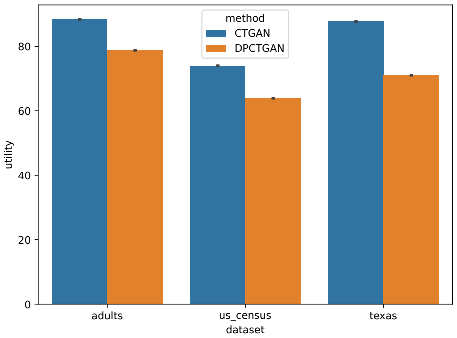

Figure 9: Utility scores for the different datasets and synthetization algorithms.

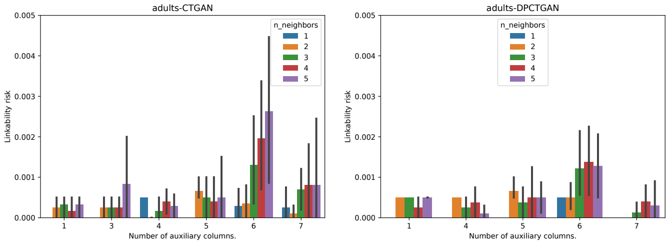

Figure 10: Measured linkability risk in the Adult dataset as a function of the amount of auxiliary columns used to mount the attack and and the number k of nearest neighbors considered in the evaluation phase. The two plots refer to synthetic data generated with CTGAN (left) and DPCTGAN (right). For more details, see Section 5.2.

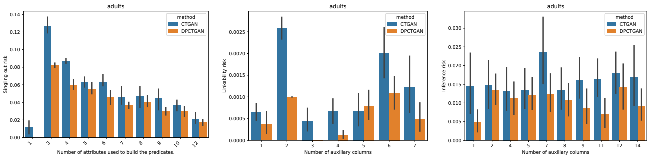

Figure 11: Singling out (left), linkability (middle), and inference (right) risks as a function of the amount of auxiliary information (in the case of linkability and inference) or number of attributes used in the predicates (for singling out) for the Adults dataset. See Section 6.3.1 for details.

Fixing all these issues is an interesting research topic, but out of the scope of this paper. The reader should however keep in mind that DP is not complete (and hence devoid of theoretical meaning).

E ADDITIONAL EXPERIMENTAL RESULTS

This section collects additional experimental results supporting the discussions presented in Sections 5.1, 6.1, 6.3, and 6.4.

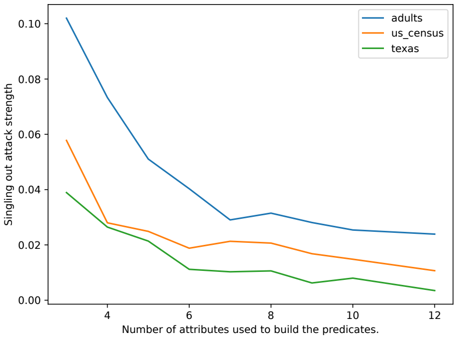

Figure 7: Strength of the multivariate singling out attack as a function of the number of attributes used to create the predicates. The singling out attack is stronger when it uses fewer attributes. As the number of attributes used to create the predicates increases, the likelihood of singling out by chance also improves, making the baseline attack more and more effective. This results in a decreasing strength of the multivariate singling out attack for a higher number of attributes.

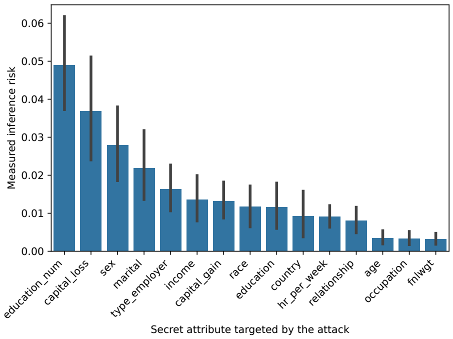

Figure 8: Inference risk for different attributes in the Adults dataset. The synthetic data has been generated with CTGAN from Section 6.2, and each column is attacked by using all the other columns as auxiliary information.