Privacy Measurements in Tabular Synthetic Data: State of the Art and Future Research Directions

Alexander T.P. Boudewijn
Aindo
AREA Science Park, Trieste, Italy
alexander@aindo.com

Andrea Filippo Ferraris
University of Turin, Italy;
Data Valley consulting srl
andreafilippo.ferraris@unito.it

Daniele Panfilo
Aindo
AREA Science Park, Trieste, Italy
daniele@aindo.com

Vanessa Cocca
Data Valley consulting srl
v.cocca@datavalley.it

Sabrina Zinutti
Aindo
AREA Science Park, Trieste, Italy
sabrina@aindo.com

Karel De Schepper
Leuven, Belgium

Carlo Rossi Chauvenet
Bocconi University, Italy
carlo.rossi@unibocconi.it

Abstract

Synthetic data (SD) have garnered attention as a privacy enhancing technology. Unfortunately, there is no standard for assessing their degree of privacy protection. In this paper, we discuss proposed assessment approaches. This contributes to the development of SD privacy standards; stimulates multi-disciplinary discussion; and helps SD researchers make informed modeling and evaluation decisions.

1 Introduction and Relation to Prior Research

Synthetic data (SD) is rapidly gaining recognition as a privacy enhancing technology (PET) [1, 2], preserving analytic value whilst removing links to real individuals. The plethora of approaches makes SDset's degree of individuals' privacy protection is hard to assess. In this paper, we codify common technical assessment frameworks for individual's privacy in SDsets. This raises interdisciplinary awareness of privacy in SD and helps SD researchers make informed modeling and assessment choices.

Several surveys mention privacy protection as an SD use case, but do not cover its assessment in a detailed manner [3–5]. Reviews of privacy in AI fail to metion SD [6, 7]. Surveys, reviews, and experimental comparisons of SD techniques provide little consideration of privacy metrics [8–10, 5, 11–14]. Legal analyses of SD are scarce and do not cover quantitative, case-by-case privacy assessment methods [15, 16].

2 Definitions and Notation

To the best of our knowledge, there is no widely accepted definition of SD. Following Jordon et al. [4], we propose Definition 2.1.

Definition 2.1. (Synthetic data, [4]) ***Synthetic data*** *(SD) is data that has been generated using a purpose-built mathematical model or algorithm (the "**generator**"), with the aim of solving a (set of) data science task(s).*

NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research.

The *generator* can be inferred through deep learning, (e.g. Generative Adversarial Networks (GANs) [17–20]; Variational Autoencoders (VAEs) [21–24]); agent-based and mathematical modelings [25, 26]; autoregressive approaches through traditional AI, e.g. decision tree learning [27, 28]; diffusion models [29, 30]; nearest neighbor-based methods [31, 32]; Bayesian networks [33]; clustering [34]; and large language models [35].

We let D denote a database describing **data subjects** through attributes $A(D)$. Rows $d \in D$ are $|A(D)|$-tuples with a value $v(d, a)$ for each attribute $a \in A(D)$. Attribute $a \in A(D)$ is **categorical** if its domain is finite and **numeric** if its domain is a subset of $\mathbb{R}$. We use the terms **row** and **record** interchangeably. We denote by $\mathcal{G}$ a generator, and by $\hat{D} \sim \mathcal{G}(D)$ denote that synthetic dataset $\hat{D}$ was obtained from generator $\mathcal{G}$ trained on D. **Seed-based** generators are a subclass of generators that produce one unique synthetic record, denoted $\mathcal{G}(d)$ for every given real record d (the **seed**). This is opposed to most models (e.g. GANs, VAEs) that represent the overall properties of datasets probabillistically, and then produce synthetic data by randomly sampling from the obtained distribution, breaking the one-to-one correspondence between real and synthetic records.

3 Synthetic Data Privacy Risks

Three key risks identified by the WP 29 [36], act as benchmark for a proper anonymization, namely: *Singling Out* (isolating records), *Linkability* (linking records concerning the same data subject in one or more datasets), and *Inference* (deducing, with significant probability, the value of an attribute). Privacy risks in SD can be a consequence of various factors. The most important ones are detailed below.

Model and data properties. Improperly trained Generators may overfit, memorizing and reproducing fixed patterns rather than inferring stochastically [37, 38]. Records that emerge in isolation, with little variability around their attribute values are difficult to generalize. As such, datasets with outliers; sparse datasets; and datasets with underrepresented strata are more at risk of memorization than more homogeneous sets [39, 40]. By their natures, such sets also have large singling-out susceptibility.

The approach to data synthesis. Most generators represent overall datasets stochastically, and obtain synthetic records by random sampling. This removes links between real data subjects and synthetic records. However, some methods (e.g. [32, 31]) create one specific synthetic record for each real record. This poses greater risk, as the link between data and data subject is maintained.

GANs may infer the minimal information needed to deceive the discriminator, failing to capture the nuances and variability of real data (mode collapse [12, 41, 42]). The SD then resembles a small selection of real data subjects well, but not the population as a whole. The SD becomes "cluttered" around specific real records, leaking information about them (see Appendix A, Figure 1).

The threat model. A threat model is the information leveraged by an adversary besides the SD (see Figure 2). They can be: 1) *No box*: the adversary accesses the SD only. 2) *Black box*: the adversary also has limited generator access (e.g. no access to the model class or parameters, but access to the model's input-output relation). 3) White box: the adversary has full generator access (model class and parameters). 4) *Uncertain box* [43]: the adversary has stochastic model knowledge (model class and knowledge that parameters stem from given probability distributions). 5) Any of the aforementioned, along with *auxiliary information*; in the context of SD formalized through Definition 3.1.

Definition 3.1. *Let D be a dataset with attributes $A(D)$. An adversary has **auxiliary information** if they know the values of some subset $A' \subseteq A(D)$ of attributes of some subset $D' \subseteq D$ of records.*

4 Mathematical Privacy Properties

4.1 Differential Privacy

Differential privacy (DP) [44] is a property of information-releasing systems. A DP system does not release data directly, but a derivative obtained through processing. The system is considered DP if the released information does not change significantly when a single record is removed from the database. DP is formally defined in Definition 4.1.

Definition 4.1. (Differential Privacy, [44]) *A randomized algorithm $\mathcal{M}$ is (ε, δ)-**differentially private** ((ε, δ)-**DP**) if for all $S \subseteq A(P)$:*

$$\mathbb{P}\left[\mathcal{M}(D) \in S\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(D') \in S\right] + \delta \tag{1}$$

for all databases D, D' such that $\exists d \in D : D' = D \setminus \{d\}$.

Generators are data releasing systems and can thus be DP: suppose we have two real datasets D and D' with $D' = D \setminus \{d\}$. Then generator $\mathcal{G}$ is DP if a data controller with access to $\hat{D} \sim \mathcal{G}$ cannot infer whether $\mathcal{G}$ was trained on D or D' (Appendix B, Figure 3). Appendix B.2 details approaches to train generators with built-in mechanisms to guarantee output data is DP. Importantly, in this context, DP is a *property of generators*, and not of the synthetic data they may produce.

4.2 k-Anonymity

Privacy risks persist even if identifying attributes like names are removed: combinations of attribute values may still single out an individual. The concept of k-anonymity was introduced to avoid thereby incurred risks [36, 45–47]. A dataset is k-**anonymous** if at least k individuals share each combination of attribute values. Further restrictions (l-diversity [48]; t-closeness [49]; (α, k)-anonymity [50]) offer additional protection.

Synthetic data based on autoregressive models can incorporate k-anonymity directly in the generation process [27]. For example, in data generated by decision trees, pruning can guarantee that each combination of attribute values is sampled at least k times in mathematical expectation [51]. Unlike DP, k-anonymity is a property of deidentified or synthetic datasets, not the algorithms producing them.

4.3 Plausible Deniability

A degree of plausible deniability is inherent in synthetic datasets, as their records do not pertain to real data subjects. Two approaches have emerged to formalize the notion of plausible deniability [52, 53], of which one is most relevant to (seed-based) synthetic data.

Definition 4.2. (Plausible deniability (PD), [52]) *Let D be a dataset and let $\mathcal{G}$ be a generator that converts any real individual record $d \in D$ into a corresponding synthetic record $\hat{d} = \mathcal{G}(d)$. For any dataset D with $|D| > k$, and any record $\hat{d}$ such that $\hat{d} = \mathcal{G}(d_1)$ for $d_1 \in D$, we say that $\hat{d}$ is releasable with (k, γ)-**plausible deniability**, if there exist at least $k - 1$ distinct records $d_2, ..., d_k \in D \setminus \{d_1\}$ such that for all $i, j \in \{1, 2, ..., k\}$:*

$$\gamma^{-1} \leq \frac{\mathbb{P}\left[\hat{d} = \mathcal{G}(d_i)\right]}{\mathbb{P}\left[\hat{d} = \mathcal{G}(d_j)\right]} \leq \gamma \tag{2}$$

Intuitively put, a generator producing synthetic records from a particular seeds has PD if, for each synthetic record generated from a specific seed, k other seeds could have resulted in *roughly the same* (quantified through γ) synthetic record. Like DP and unlike k-anonymity, PD is therefore a property of (seed-based) generators, though it shares intuition with both other properties.

5 Statistical Privacy Indicators

5.1 Identical Records, Distances, and Nearest Neighbors

Most indicators quantify how many synthetic records are identical, or suspiciously similar to particular real records. Unlike DP and PD, these indicators measure *properties of synthetic datasets*, not their generators. The proportion of synthetic records that coincide with real records is referred to as the *identical match share* (IMS) [54–56]. The IMS is therefore generalized to similarity metrics, and further to Nearest neighbor (NN)-based methods. The latter two can be classified based on the properties detailed below. Table 3 of Appendix C classifies approaches along these properties.

Similarity metrics. Table 2 in Appendix C contains an overview of commonly invoked measures.

Metric evaluation. A complicating factor in evaluating similarity metrics in structured datasets is the multitude of datatypes. The following approaches exist to do so: 1) *binning numeric attributes*; treating them as categorical; and using a metric for categorical values. 2) *Aggregation of multiple metrics*, applying one metric per type and integrating the results. 3) *Ignoring attributes*, for instance by considering only numerical attributes with a metric appropriate for them. 4) *Evaluating distances in embedding spaces*, in which all information is preserved, but represented in normalized, numeric attributes, e.g. through t-SNE [57], discriminant analysis [58], factor analysis [59], or representation learning [60].

Evaluated distances. For a given synthetic record $\hat{d} \in \hat{D}$, we can find its closest real record $d \in D$. We call the distance between these records the *synthetic to real distance (SRD) of* $\hat{d}$, denoted by $\texttt{SRD}(\hat{d})$ (see equation (3), where `Dist` is some similarity metric).

$$\texttt{SRD}(\hat{d}) := \min_{d \in D} \texttt{Dist}(\hat{d}, d) \quad \forall \hat{d} \in \hat{D} \tag{3}$$

In an analogous fashion, the smallest synthetic to synthetic (`SSD`), real to synthetic (`RSD`), and real to real distance (`RRD`) can be defined. These are all visualized in Figure 4 of Appendix C.

Use of holdout sets. To compute the RRD, the real data D can be partitioned into two subsets D_1 and D_2. For a real record $d_1 \in D_1$, the RRD is then the smallest distance to any record $d_2 \in D_2$, as in equation (4). This "holdout set" provides a baseline for comparing SD against [55].

$$\texttt{RRD}(d_1) := \min_{d_2 \in D_2} \texttt{Dist}(d_1, d_2) \quad \forall d_1 \in D_1 \tag{4}$$

Statistics. The *Distance to closest record* (DCR) compares the SRD and RRD distributions. Real data subjects may be at risk if, for some synthetic record $\hat{d}$, we have $\texttt{SRD}(\hat{d}) < \texttt{RRD}(d^*)$, with $d^* := \arg\min_{d \in D} \texttt{Dist}(\hat{d}, d)$. The DCR is sensitive to realistically replicated outliers, as they have large RRDs (see Appendix C, Figure 5). Risks are expressed statistically through proportions [24, 61] and medians, means and standard deviations of "suspiciously close" synthetic records, with [62, 55, 63, 24] or without [64, 29, 30] using a hold-out set. Small percentiles are also often invoked [63, 54, 65, 32]. E.g., Mami et al. [63] compute the proportion P of synthetic records that closer to real records than the smalles 5% of RRDs, using a holdout set.

Panfilo et al. [24, 61] use an *inferential statistical test* to assess whether the SRD and RRD stem from the same distribution. Yale et al. [66–68] introduced the *adversarial accuracy* and *Privacy loss*, including the SSD and RSD for a baseline. Some distance-based indicators are for seed-based SD only: *distance-based record linkage* [69–71]; the *hidden rate* [32]; and *local cloaking* [32].

5.2 Other Statistical Indicators

Taub et al.[72] introduce the *targeted correct attribution probability* (TCAP) indicator. This TCAP is essentially an indicator of parameter inference attack success rates. It quantifies the frequency with which synthetic parameter values correspond to real ones in l-diverse equivalence classes. Emam et al. [73] derive a related probabilistic approach to quantify the risks of the WP29 attacks, using real holdout sets as baselines. Esteban et al. [74] and Rashidian et al. [75] propose using the maximum mean discrepancy (MMD) as a privacy metric, inferentially testing whether the generator overfits.

6 Computer Scientific Experimental Privacy Assessment

Computer scientific privacy assessment is the deliberate conducting of SD-informed privacy attacks, and the measurement of their effectiveness, to quantify SD's degree of protection. Attack frameworks are classified in Table 4 (Appendix D), based on threat models and the factors outlined below.

The use of specific threat models is an important innovation of the computer scientific approach. Mathematical properties and statistical indicators pertain only to either generators or synthetic data (but not both). Computer scientific attacks, on the other hand, allow for flexibility in modeling how much knowledge an adversary may have about generators.

6.1 Attack Frameworks

Vulnerable Record Discovery (VRD). Some methods conduct attacks by identifying seemingly vulnerable synthetic records. Giomi et al. [76] propose looking for synthetic records with unique (combinations of) attribute values. Singling out attacks are then claims that these records are also unique real records. Carlini et al. [77] study the extent to which generators overfit, quantifying the likelihood of synthetic records being memorized secrets.

Adversarial Machine Learning. Model inversion, membership inference attacks (MIAs), and shadow modeling ("model stealing") compromise confidentiality without technological system misuse [78–80]. In a MIA, an adversary infers whether a given target record was in the training dataset of a given ML model. This can be through classifier models [81–84, 66, 85–87]. A *shadow model* (SM) is constructed by an adversary to mimic a given model. SMs may mimic a given generator to conduct MIAs [83, 43, 82, 88]: data is provided to the SM and the real model. By comparing their outputs, the adversary determines whether a given record was in the training set of the real model [89, 87, 84, 85]. Shadow modeling requires at least a black box threat model.

Combined approaches. Recent approaches use VRD to make informed decisions for potential targets in membership inference attacks, reducing the computational burden [29, 84, 90, 91].

6.2 Attack Mechanisms

Nearest Neighbors (NN). Suppose an adversary has auxiliary information about a target, but does not know the value of one of its attributes. They may then assign the missing value based on the target's k synthetic NNs (see Appendix D, Figure 6) [81, 86, 68, 76, 43, 92, 93]. Experimentally, small k values perform well, particularly $k = 1$ [86, 81, 76, 93]. More auxiliary information means better attacks: Experimentally, access to one extra attribute of auxiliary information roughly increases accuracy by 30% across datasets and generators [86].

Machine Learning (ML). Techniques from ML can guide the attack process. For instance, classifiers can be trained to re-identify real data subjects [89, 87, 83, 84]. Classifiers and regression models can also be used to estimate parameter values of real records based on synthetic ones [81, 86].

Information Theory (IT). IT concepts like Shannon entropy [94] and mutual information can be used to identify the degree to which records in an anonymized dataset deviate from more common ones, in which case they have a higher likelihood of being memorized by an overfit generator [40, 77, 95, 96].

6.3 Baselines and Effectiveness Estimation

Absolute Measurements. Efficacy metrics not requiring baselines include the probability with which records can be singled out [77, 96, 95]; and the proportion of real records for which information can be re-identified [81, 86]. ML-based attacks can be evaluated through ML metrics (see [37]). For instance, MIAs based on classification are evaluated with ROC AUCs [84, 68, 83]; F-1 scores [82, 83] and Precision and recall [81].

Random baseline. Giomi et al. [76] propose a random baseline, to evaluate attack efficacy with SD access to that of uninformed (random) guesses (see Appendix D, Figure 7). For some methods, the mathematical expectation of random hypotheses is known a priori and implicitly integrated in scoring (see, e.g. [93]).

Control baseline. Giomi et al. [76] propose using a *control baseline*, by splitting the real data into a training set and a control set. The generator is trained using the training set, having no access to the control set. The estimated success rate of attacks on the training data (Figure 2) is compared to that of attacks on the control data (Figure 8). If the former is large, the SD may leak information. However, if the latter is also large, this was generic, population-level information, as opposed to specific secrets of specific data subjects ("*relating to in content*" in the legal analysis of López and Elbi [16]).

Deliberate secret insertion. Deliberate secrets can be inserted in the training data [77], or the SD after generation [76]. Re-identifying these secrets casts light on the ease of inferring sensitive information from synthetic data.

6.4 Relation to WP29 Attack Types

Singling out. VRD directly implements singling out attacks, identifying SD records that likely result from overfit generators (typically outliers). This works even under a no box threat model. MIA can also model singling out: the adversary quantifies the likelihood of a combination of attribute values being a unique real record.

Linkage. Attacks with NN as mechanism require auxiliary information. They may be interpreted as linkage attacks: the adversary has to obtain the auxiliary information from some other data source, linking it to the SD. The *Anonymeter* framework [76] and information theory-based VRD [95] are the only methods to explicitly model linkage attacks.

Inference. NN-based attacks and MIA can function as inference attacks. E.g. if the SD is used to evaluate the rare disease treatment efficacy, the training set contains patient data. If an adversary can determine that Giovanna is in the training dataset, they can infer that she has the disease.

7 Discussion

7.1 The Assessment Frameworks

Mathematical privacy properties. In DP, there is no consensus on the choice of parameters (ε, δ). Large parameter values offer weak privacy guarantees. A given ε can result in different degrees of protection for different use cases, making it hard to choose values in practice [97]. Furthermore, DP SD is still susceptible to linkage and inference attacks [36, 98], possibly providing a false sense of security. DP is a property of generators, not their produced data.

Achieving k-anonymity involves considerable information destruction [99] and is NP-hard to achieve optimally through generalization [100]. In a court ruling in California, k-anonymity was shown to offer sufficient protection, only once the analytic utility is completely removed [101]. This is corroborated by findings that with a combination of only fifteen parameter values, over 99% of a population can be re-identified [102]. Unlike DP, k-anonymity is a property of deidentified datasets, not the methods that produce them.

PD is only Applicable to seed-based methods, ruling out most classes of generators. Like k-anonymity, an individual record is considered protected if it is indistinguishable from a fixed number of other records. The difference is that PD is developed for synthetic data specifically, with "indistinguishable" the probability of stemming from multiple seeds. The notion of probabilistic lack of impact of an individual record also shares its intuition with DP, with which PD is closely related. To date, PD has gained little traction in practice.

Statistical privacy indicators. Distance-based indicators (by far the most common indicators) are difficult to interpret. The multitude of options and involved modeling decision adds to this confusion. Evaluation of distances can be a particular difficulty, as structured data has mixed data types, for which different similarity metrics may be appropriate. Choice of similarity metric and its evaluation may have an impact on results. Statistical indicators measure properties of synthetic data, not their generators.

Computer scientific privacy experiments. Deliberate attacks are most commonly MIAs. The results of such assessments provide crucial insight into data privacy. However, they often only work under threat models with considerable information. No box, no auxiliary information approaches are more rare, and typically confined to outlier detection (VRD).

Most attack-based approaches require auxiliary information, Arguably making them linkage attacks. Most attack approaches rely on distances (nearest neighbors) or ML as an attack mechanism, so we hypothesize that distance-based indicators are very reliable predictors of their efficacy, quantifying risks in a more all-encompassing manner.

Unlike the other approaches, computer scientific experiments can leverage several threat models. This allows them to take into consideration properties of both the synthetic data and their generators.

7.2 Relation to Synthetic Data Risks

A core SD risk is generator memorization, particularly around outliers; in sparse or small datasets; or through mode collapse. All frameworks address this risk: mathematical properties center around uniqueness of records. DP measures the impact of individual training records, with outliers clearly having large individual impacts. Furthermore, k-anonymity and PD specifically foster datasets with severely limited uniqueness of individual records.

Distance-based indicators are sensitive to outliers, as their synthetic neighbors have small SRDs, while the corresponding real outliers have large RRDs (Figure 5). Other statistical indicators are measures of memorization by their very nature. In computer scientific experiments, VRD deliberately seeks for outliers, while MIAs are nearly exclusively effective for outliers (see, e.g. [87]).

To the best of our knowledge, no research was conducted to assess whether seed-based generators inherently pose greater risks than other generators. Intuitively, this seems evident, as they do not remove the links between (synthetic) records and real data subjects.

7.3 Suggestions for Future Research

Standardizing privacy assessment. Synthetic data privacy is a multifaceted subject encompassing several disciplines such as mathematics, computer science, ethics, policy-making, law, and philosophy. There is a pressing need for increased interdisciplinary research to gain an inclusive understanding of synthetic data as a PET. Conventional assessment standards should be developed, so that research findings are easy to interpret, compare, and contrast. Consensus should be formed over whether privacy is a property of synthetic datasets, the generators that produce them, or some combination of both.

Synergies between assessments. A comparison (deductive or experimental, e.g. comparing multiple assessments on the same SDsets) between mathematical, statistical, and empirical privacy approaches would indicate consistency, and identify merits and weaknesses. For replicability, experiments should use open-source generators and publicly available datasets (e.g. from the UCI ML repository [103]). As attacks are often distance (NN)-based, insights from indicators should be integrated in simulated attacks (e.g. involved distance metrics and their evaluation methods; distances; use of holdout set for a baseline; statistical interpretation of results).

Outlier protection. Following Tai et al. [39], research should address outlier protection in SD, e.g. by binning and aggregating attribute values (cf. Section 5.1 on metric evaluation); through innovation; or by invoking other PETs. Outlier detection methods (e.g. [104–106]) can be used for VRD.

Incorporating privacy into generators. While DP is incorporated in various generators, this is not true for privacy metrics and empirical privacy approaches. Future research should focus on incorporating the latter two, for instance by incorporating metrics in loss functions, or through combinatorial optimization. Combining SD with outlier-protecting PETs should be considered.

Assessment for advanced data formats. Most covered approaches to privacy assessment in structured data were developed for data contained in a single table. More research is required to assess privacy in relational datasets, with information contained in multiple, interconnected tables. So-called "*profiling attacks*" re-identify subjects not by their literal records, but by latent behavioral patterns [107]. Such attacks may play a more considerable role in the context of relational databases.

Distribution-level confidentiality. The outlined frameworks are developed to assess the upholding individuals' right to privacy. In practice, properties of datasets as a whole may additionally be confidential. They may for instance be trade secrets (e.g. the total number of annual transactions of a financial institution). The reader is referred to [108–110] for contemporary assessment frameworks of confidentiality on the level of overall dataset properties.

References

[1] M. Anders, C. Ivanov, R. Riemann, X. Lareo, and S. Leucci, "Techsonar: 2022 - 2023 report," 2022.

[2] S. Judah, A. White, S. Sicular, L. Clougherty Jones, G. De Simoni, T. Friedman, M. Beyer, J. Heizenberg, and S. Parker, "Gartner predicts 2021: Data and analytics strategies to govern,

scale and transform digital business," 2020.

[3] B. van Breugel and M. van der Schaar, "Beyond privacy: Navigating the opportunities and challenges of synthetic data," 2023.

[4] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, "Synthetic data – what, why and how?," 2022.

[5] S. I. Nikolenko, "Synthetic data for deep learning," 2019.

[6] E. D. Cristofaro, "An overview of privacy in machine learning," 2020.

[7] C. C. Aggarwal and P. S. Yu, *A General Survey of Privacy-Preserving Data Mining Models and Algorithms*, pp. 11–52. Boston, MA: Springer US, 2008.

[8] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, vol. 493, pp. 28–45, 2022.

[9] A. Pathare, R. Mangrulkar, K. Suvarna, A. Parekh, G. Thakur, and A. Gawade, "Comparison of tabular synthetic data generation techniques using propensity and cluster log metric," *International Journal of Information Management Data Insights*, vol. 3, no. 2, p. 100177, 2023.

[10] T. E. Raghunathan, "Synthetic data," *Annual Review of Statistics and Its Application*, vol. 8, no. 1, pp. 129–140, 2021.

[11] C. Task, K. Bhagat, and G. Howarth, "SDNist v2: Deidentified Data Report Tool," Mar. 2023.

[12] A. Figueira and B. Vaz, "Survey on synthetic data generation, evaluation methods and gans," *Mathematics*, vol. 10, no. 15, 2022.

[13] M. Hittmeir, A. Ekelhart, and R. Mayer, "Utility and privacy assessments of synthetic data for regression tasks," *2019 IEEE International Conference on Big Data*, 2019.

[14] G. M. Raab, "Utility and disclosure risk for differentially private synthetic categorical data," in *Privacy in Statistical Databases* (J. Domingo-Ferrer and M. Laurent, eds.), (Cham), pp. 250–265, Springer International Publishing, 2022.

[15] S. M. Bellovin, K. Dutta, Preetam, and N. Reitinger, "Privacy and synthetic datasets," *Stanford Technology Law Review*, 2018.

[16] C. A. López and A. Elbi, "On the legal nature of synthetic data," *Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[18] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *International Journal of Information Management*, p. 100004, 01 2021.

[19] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," *arXiv*, 2018.

[20] E. Jang, S. Gu, and B. Poole, "Modeling tabular data using conditional gan," *arXiv*, 2019.

[21] D. Kingma and M. P. Welling, "Auto-encoding variational bayes," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

[22] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2016.

[23] H. Huang, R. He, Z. Sun, T. Tan, *et al.*, "Introvae: Introspective variational autoencoders for photographic image synthesis," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[24] D. Panfilo, A. Boudewijn, S. Saccani, A. Coser, B. Svara, C. R. Chauvenet, C. A. Mami, and E. Medvet, "A deep learning-based pipeline for the generation of synthetic tabular data," *IEEE Access*, pp. 1–1, 2023.

[25] E. Bonabeau, "Agent-based modeling: Methods and techniques for sim- ulating human systems," *Proceedings of the national academy of sciences*, 2002.

[26] R. Carmona and F. Delarue, *Probabilistic Theory of Mean Field Games with Applications*. Springer, 2018.

[27] B. Nowok, G. M. Raab, and C. Dibben, "Synthpop: Bespoke creation of synthetic data in r," *Journal of Statistical Software*, vol. 74, no. 11, p. 1–26, 2016.

[28] S. Mahiou, K. Xu, and G. Ganev, "dpart: Differentially private autoregressive tabular, a general framework for synthetic data generation," 2022.

[29] T. Ceritli, G. Ghosheh, V. Chauhan, T. Zhu, A. Creagh, and D. Clifton, "Synthesizing mixed-type electronic health records using diffusion models," 02 2023.

[30] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "Tabddpm: Modelling tabular data with diffusion models," 2022.

[31] M. Beigi, A. Shafquat, J. Mezey, and J. Aptekar, "Simulants: Synthetic clinical trial data via subject-level privacy-preserving synthesis," *AMIA - Annual Symposium proceedings. AMIA Symposium, 231–240*, 2022.

[32] M. Guillaudeux, O. Rousseau, J. Petot, Z. Bennis, C.-A. Dein, T. Goronflot, N. Vince, S. Limou, M. Karakachoff, M. Wargny, and P.-A. Gourraud, "Patient-centric synthetic data generation, no reason to risk reidentification in biomedical data analysis," *NPJ Digital Medicine*, no. 6(37), 2023.

[33] D. Kaur, M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon, "Application of bayesian networks to generate synthetic health data," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 801–811, 2021.

[34] M. Li, D. Zhuang, and J. M. Chang, "Mc-gen:multi-level clustering for private synthetic data generation," 2022.

[35] V. Borisov, K. Sessler, T. Leemann, M. Pawelczyk, and G. Kasneci, "Language models are realistic tabular data generators," in *The Eleventh International Conference on Learning Representations*, 2023.

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

[37] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[38] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," 2017.

[39] B.-C. Tai, S.-C. Li, and Y. Huang, "K-aggregation: improving accuracy for differential privacy synthetic dataset by utilizing k-anonymity algorithm," in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, pp. 772–779, IEEE, 2017.

[40] A. Sen, C. Task, D. Kapur, G. Howarth, and K. Bhagat, "Diverse community data for benchmarking data privacy algorithms," 2023.

[41] E. Richardson and Y. Weiss, "On GANs and GMMs," 2018.

[42] A. Gainetdinov, "GAN mode collapse explanation," in *TowardsAI.net*, 2023.

[43] F. Houssiau, J. Jordon, S. N. Cohen, O. Daniel, A. Elliott, J. Geddes, C. Mole, C. Rangel-Smith, and L. Szpruch, "Tapas: a toolbox for adversarial privacy auditing of synthetic data," 2022.

[44] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[45] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[46] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[47] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners," *Trans. Data Privacy*, vol. 7, p. 337–370, dec 2014.

[48] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, 03 2007.

[49] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115, 2007.

[50] R. Wong, J. Li, A. Fu, and K. Wang, "(α, k)-anonymous data publishing," vol. 33, p. 209–234, 10 2009.

[51] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, and G. Epelde, "Reliability of supervised machine learning using synthetic data in healthcare: A model to preserve privacy for data sharing (preprint)," *JMIR Medical Informatics*, vol. 8, 03 2020.

[52] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," 2017.

[53] S. Rass, S. König, J. Wachter, M. Egger, and M. Hobisch, "Supervised machine learning with plausible deniability," *Computers & Security*, vol. 112, p. 102506, 2022.

[54] J. Hradec, M. Craglia, M. Di Leo, S. De Nigris, N. Ostlaender, and N. Nicholson, *Multipurpose synthetic population for policy applications*. JRC technical report, 2022.

[55] M. Platzer and T. Reutterer, "Holdout-based empirical assessment of mixed-type synthetic data," *Frontiers in Big Data*, 2021.

[56] D. Sinha Roy, C. Pandey, V. Tiwari, and J. J. P. C. Rodrigues, "Transforming internet traffic prediction with 5gt-trans: A synthetic data and transformer-based approach," *SSRN preprint*, 2023.

[57] L. van der Maaten and G. Hinton, "Viualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[58] C. J. Huberty, "Discriminant analysis," *Review of Educational Research*, vol. 45, no. 4, pp. 543–598, 1975.

[59] G. Saporta, "Simultaneous analysis of qualitative and quantitative data," in *Societa Italiana di Statistica. XXXV riunione scientifica*, vol. 1, pp. 62–72, CEDAM, 1990.

[60] T. Shenkar and L. Wolf, "Anomaly detection for tabular data with internal contrastive learning," in *International Conference on Learning Representations*, 2021.

[61] D. Panfilo, *Generating Privacy-Compliant, Utility-Preserving Synthetic Tabular and Relational Datasets Through Deep Learning*. University of Trieste, 2022.

[62] J. Weldon, T. Ward, and E. Brophy, "Generation of synthetic electronic health records using a federated gan," 2021.

[63] C. A. Mami, A. Coser, E. Medvet, A. T. P. Boudewijn, M. Volpe, M. Withworth, B. Svara, G. Sgroi, D. Panfilo, and S. Saccani, "Generating realistic synthetic relational data through graph variational autoencoders," *Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[64] S. An and J.-J. Jeon, "Distributional learning of variational autoencoder: Application to synthetic data generation," 2023.

[65] Z. Zhao, A. Kunar, H. V. der Scheer, R. Birke, and L. Y. Chen, "Ctab-gan: Effective table data synthesizing," 2021.

[66] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. Bennett, "Assessing privacy and quality of synthetic health data," in *Business Information Systems Workshops, Lecture Notes in Business Information Processing, Springer International Publishing*, pp. 1–4, 05 2019.

[67] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. Bennett, "Generation and evaluation of privacy preserving synthetic health data," *Neurocomputing*, vol. 416, 04 2020.

[68] A. Yale, S. Dash, K. Bhanot, I. Guyon, J. Erickson, and K. Bennett, *Synthesizing Quality Open Data Assets from Private Health Research Studies*, pp. 324–335. 11 2020.

[69] V. Torra, J. M. Abowd, and J. Domingo-Ferrer, "Using mahalanobis distance-based record linkage for disclosure risk assessment," in *Privacy in Statistical Databases* (J. Domingo-Ferrer and L. Franconi, eds.), (Berlin, Heidelberg), pp. 233–242, Springer Berlin Heidelberg, 2006.

[70] J. Herranz, J. Nin, P. Rodríguez, and T. Tassa, "Revisiting distance-based record linkage for privacy-preserving release of statistical datasets," *Data & Knowledge Engineering*, vol. 100, pp. 78–93, 2015.

[71] J.-S. Lee and S.-P. Jun, "Privacy-preserving data mining for open government data from heterogeneous sources," *Government Information Quarterly*, vol. 38, no. 1, p. 101544, 2021.

[72] J. Taub, M. J. Elliot, and G. M. Raab, "Creating the best risk-utility profile : The synthetic data challenge," 2019.

[73] K. El Emam, L. Mosquera, and J. Bass, "Evaluating identity disclosure risk in fully synthetic health data: Model development and validation," *Journal of Medical Internet Research*, vol. 22, no. 11, 2020.

[74] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional gans," 2017.

[75] S. Rashidian, F. Wang, R. Moffitt, V. Garcia, A. Dutt, W. Chang, V. Pandya, J. Hajagos, M. Saltz, and J. Saltz, "Smooth-gan: Towards sharp and smooth synthetic ehr data generation," in *Artificial Intelligence in Medicine* (M. Michalowski and R. Moskovitch, eds.), (Cham), pp. 37–48, Springer International Publishing, 2020.

[76] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, "A unified framework for quantifying privacy risk in synthetic data," 2022.

[77] N. Carlini, C. Liu, Úlfar Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," 2019.

[78] R. S. S. Kumar, D. O. Brien, K. Albert, S. Viljöen, and J. Snover, "Failure modes in machine learning systems," 2019.

[79] R. S. Siva Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia, "Adversarial machine learning-industry perspectives," in *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 69–75, 2020.

[80] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," 2021.

[81] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," 2018.

[82] A. Kuppa, L. Aouad, and N.-A. Le-Khac, "Towards improving privacy of synthetic dataset," 06 2021.

[83] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proceedings of the VLDB Endowment*, vol. 11, pp. 1071–1083, 06 2018.

[84] B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar, "Membership inference attacks against synthetic data through overfitting detection," 2023.

[85] B. Oprisanu, G. Ganev, and E. D. Cristofaro, "On utility and privacy in synthetic genomic data," 2022.

[86] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Slaes, "Generation and evaluation of synthetic patient data," *BMC Medical Research Methodology*, 2020.

[87] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data – anonymisation groundhog day," 2022.

[88] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro, "Knock knock, who's there? membership inference on aggregate location data," 2017.

[89] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

[90] M. Meeus, F. Guepin, A.-M. Cretu, and Y.-A. de Montjoye, "Achilles' heels: Vulnerable record identification in synthetic data publishing," 2023.

[91] H. Hu and J. Pang, "Membership inference attacks against gans by leveraging over-representation regions," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, (New York, NY, USA), p. 2387–2389, Association for Computing Machinery, 2021.

[92] J. Jordon, D. Jarrett, J. Yoon, T. Barnes, P. Elbers, P. Thoral, A. Ercole, C. Zhang, D. Belgrave, and M. van der Schaar, "Hide-and-seek privacy challenge," 2020.

[93] J. Jordon, D. Jarrett, E. Saveliev, J. Yoon, P. Elbers, P. Thoral, A. Ercole, C. Zhang, D. Belgrave, and M. van der Schaar, "Hide-and-seek privacy challenge: Synthetic data generation vs. patient re-identification," in *Proceedings of the NeurIPS 2020 Competition and Demonstration Track* (H. J. Escalante and K. Hofmann, eds.), vol. 133 of *Proceedings of Machine Learning Research*, pp. 206–215, PMLR, 12 2021.

[94] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[95] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *arXiv preprint cs/0610105*, 2006.

[96] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, 2008.

[97] J. Lee and C. Clifton, "How much is enough? choosing ε for differential privacy," in *Proceedings of the 14th International Conference on Information Security*, ISC'11, (Berlin, Heidelberg), p. 325–340, Springer-Verlag, 2011.

[98] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data-a privacy mirage," *arXiv preprint arXiv:2011.07018*, 2020.

[99] O. Angiuli and J. Waldo, "Statistical tradeoffs between generalization and suppression in the de-identification of large-scale data sets," in *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, pp. 589–593, 2016.

[100] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, (New York, NY, USA), p. 223–228, Association for Computing Machinery, 2004.

[101] Supreme Court of California, "RICHARD SANDER et al., Plaintiffs and Appellants, v. STATE BAR OF CALIFORNIA et al., Defendants and Respondents.," 2013.

[102] L. Rocher, J. Hendrickx, and Y.-A. Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communications*, vol. 10, 07 2019.

[103] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[104] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, 07 2009.

[105] G. Pang, A. Hengel, C. Shen, and L. Cao, "Deep reinforcement learning for unknown anomaly detection," 09 2020.

[106] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Muller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, pp. 756–795, may 2021.

[107] A. J. Tournier and Y.-A. de Montjoye, "Expanding the attack surface: Robust profiling attacks threaten the privacy of sparse behavioral data," *Science Advances*, vol. 8, no. 33, 2022.

[108] W. Zhang, O. Ohrimenko, and R. Cummings, "Attribute privacy: Framework and mechanisms," 2021.

[109] Z. Lin, S. Wang, V. Sekar, and G. Fanti, "Summary statistic privacy in data sharing," 2023.

[110] A. Suri, Y. Lu, Y. Chen, and D. Evans, "Dissecting distribution inference," 2022.

[111] L. Rosenblatt, X. Liu, S. Pouyanfar, E. de Leon, A. Desai, and J. Allen, "Differentially private synthetic data: Applied evaluations and enhancements," 2020.

[112] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 08 2016.

[113] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," 2018.

[114] M. L. Fang, D. S. Dhami, and K. Kersting, "Dp-ctgan: Differentially private medical data generation using ctgans," in *Artificial Intelligence in Medicine* (M. Michalowski, S. S. R. Abidi, and S. Abidi, eds.), (Cham), pp. 178–188, Springer International Publishing, 2022.

[115] R. Torkzadehmahani, P. Kairouz, and B. Paten, "Dp-cgan: Differentially private synthetic data and label generation," pp. 98–104, 06 2019.

[116] B. Xin, W. Yang, Y. Geng, S. Chen, S. Wang, and L. Huang, "Private fl-gan: Differential privacy synthetic data generation based on federated learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2927–2931, 2020.

[117] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," 2018.

[118] J. Yoon, J. Jordon, and M. van der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations*, 2019.

[119] N. Papernot, M. Abadi, Úlfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," 2017.

[120] D. Lee, H. Yu, X. Jiang, D. Rogith, M. Gudala, M. Tejani, Q. Zhang, and L. Xiong, "Generating sequential electronic health records using dual adversarial autoencoder," *Journal of the American Medical Informatics Association*, vol. 27, pp. 1411–1419, 09 2020.

[121] R. McKenna, G. Miklau, and D. Sheldon, "Winning the NIST contest: A scalable and general approach to differentially private synthetic data," 2021.

[122] R. McKenna, D. Sheldon, and G. Miklau, "Graphical-model based estimation and inference for differential privacy," 2019.

[123] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," 2012.

[124] R. Venugopal, N. Shafqat, I. Venugopal, B. M. J. Tillbury, H. D. Stafford, and A. Bourazeri, "Privacy preserving generative adversarial networks to model electronic health records," *Neural Networks*, vol. 153, pp. 339–348, 2022.

[125] P.-H. Lu, P.-C. Wang, and C.-M. Yu, "Empirical evaluation on synthetic data generation with generative adversarial network," WIMS2019, (New York, NY, USA), Association for Computing Machinery, 2019.

[126] A. V. Solatorio and O. Dupriez, "Realtabformer: Generating realistic relational and tabular data using transformers," 2023.

[127] T. Zhang, S. Wang, S. Yan, J. Li, and Q. Liu, "Generative table pre-training empowers models for tabular prediction," 2023.

[128] A. Kunar, "Effective and privacy preserving tabular data synthesizing," 2021.

[129] S. Norgaard, R. Saeedi, K. Sasani, and A. H. Gebremedhin, "Synthetic sensor data generation for health applications: A supervised deep learning approach," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1164–1167, 2018.

A Synthetic Data Risks

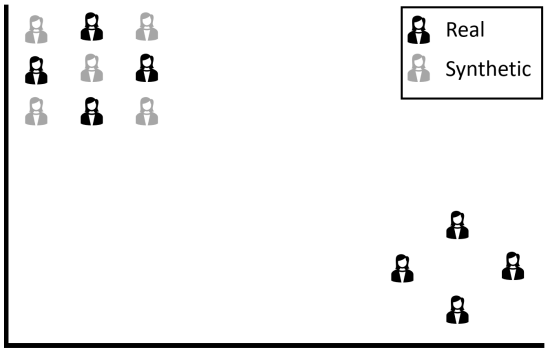


Figure 1: Mode collapse: the real records (visualized in black) are varied, and represented in two regions of the plane (top-left; bottom-right). The generator only learns to replicate synthetic records (visualized in grey) in the top-left group: merely inferring their patterns is sufficient to deceive the discriminator.

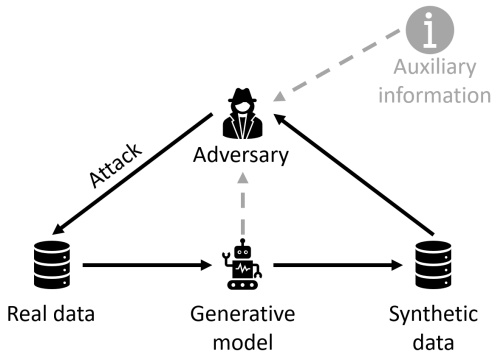


Figure 2: Visualization of an attack. Dotted lines indicate that the adversary may or may not leverage information sources, depending on the threat model: the adversary may use auxiliary information and/or information about the generative model (no, black, uncertain, or white box).

B Differentially Privacy for SD

B.1 Generators as Information Release Systems

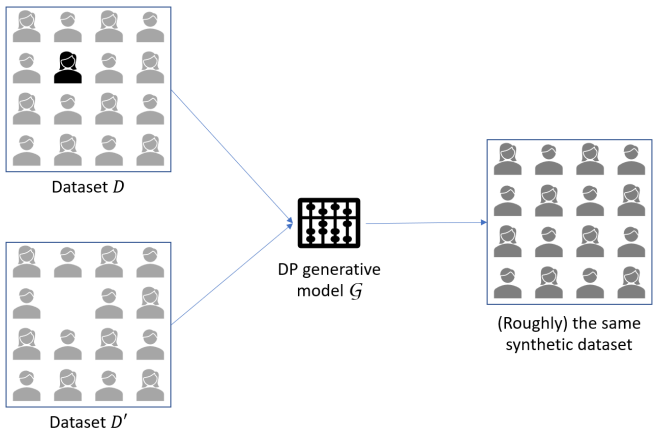


Figure 3: Differential privacy in synthetic data generative models: the SD released by generator $\mathcal{G}$ does not alter significantly if any single individual d (indicated in black) is removed from the training data. This protects the data subject with record d, as it makes it impossible to pinpoint information to them.

B.2 Overview of SD Methods with Built-In DP Mechanisms

Table 1 contains an overview of methods for generating SD, in which DP is incorporated directly. The resulting SD, when released, then automatically guarantees DP for the real dataset. Rosenblatt et al. [111] compare the performances of three approaches to DP-guaranteed SD generation (DP-GAN, PATE-GAN, MWEM, see below), concluding that PATE-GAN has the best utility for tabular data in ML applications.

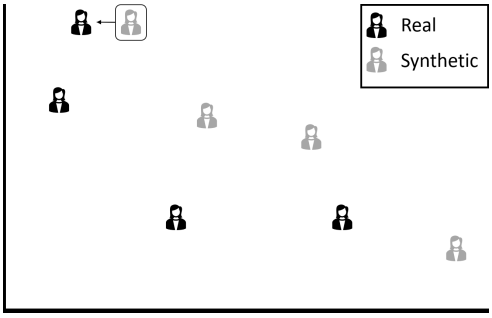
Method	SD technology	DP integration method	Reference(s)
Differential privacy GAN (DP-GAN)	GAN	Noise addition during training	[112, 113] [114, 115]
Federated learning GAN (FL-GAN)	GAN	Multiple clients train part of the GAN (see [117]) on non-overlapping, noise-added datasets	[116]
PATE-GAN	GAN	Multiple discriminators are trained, each on different subsets of data. Their classifications (real or synthetic) are aggregated. noise is added during the aggregation process (private aggregation of teacher ensembles, i.e. PATE, see [119])	[118]
Dual adversarial autoencoders (DAAE)	VAE, GAN	Noise addition during training	[120]
Private-PGM	Bayesian network	Sample marginals of the real data; infer high-dimensional distribution from these marginals through a probabilistic graphical model (PGM, see [122]); add noise during sampling.	[121]
Simulants	Nearest Neighbor	Noise addition during training	[31]
Multiple-level clustering generator (MC-GEN)	Clustering	Noise addition during training	[34]
Multiplicative Weights Exponential Mechanism (MWEM)	Optimization	Noise addition	[123]

Table 1: Overview of generative models with built-in differential privacy mechanism

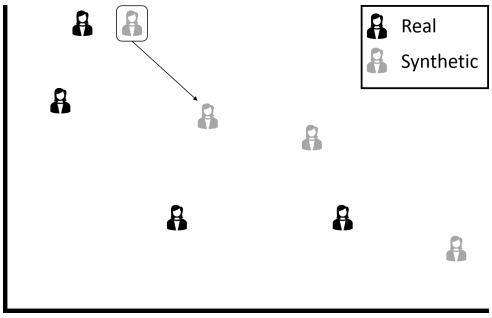
C Statistical Privacy Indicators

Name	Computation	Remark
$\mathcal{L}_1$-distance	$\texttt{Dist}_{\mathcal{L}_1}(d, d') = \sum_{a \in A(D)} \lvert v(d,a) - v(d',a) \rvert$	Numeric attributes; also known as *Manhattan distance*
Euclidean distance	$\texttt{Dist}_E(d, d') = \sqrt{\sum_{a \in A(D)} (v(d,a) - v(d',a))^2}$	Numeric attributes;
Hamming distance	$\texttt{Dist}_H(d, d') = \lvert \{a \in A(D) : v(d,a) \neq v(d',a)\} \rvert$	Categorical attributes;
Cosine similarity	$\texttt{Dist}_C(d, d') = \frac{\sum_{a \in A(D)} v(d,a) v(d',a)}{\sqrt{\sum_{a \in A(D)} v(d,a)^2 \cdot \sum_{a \in A(D)} v(d',a)^2}}$	Numeric attributes; Technically not a distance metric
Manhalobis distance	$\texttt{Dist}_M(d, d') = \sqrt{(d-d')S^{-1}(d-d')^T}$	For S the covariance matrix of real and synthetic distributions Categorical attributes; Generalization of Euclidean distance taking correlation into account.
Gower distance	Hamming distance for categorical attributes $+\mathcal{L}_1$ distance for numerical attributes	An aggregation of two metrics

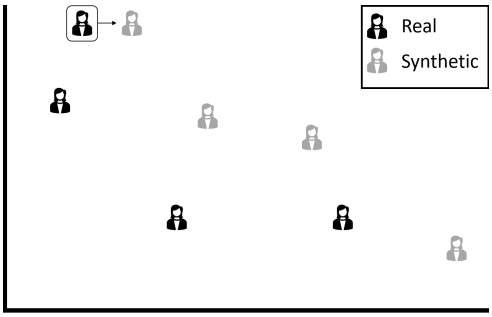
Table 2: Common distance and similarity metrics in synthetic data privacy assessment



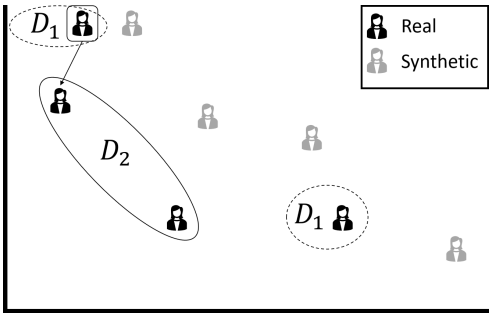
(a) Smallest synthetic to real distance (SRD) of selected synthetic record



(b) Smallest synthetic to synthetic distance (SSD) of selected synthetic record



(c) Smallest real to synthetic distance (RSD) of selected real record



(d) Smallest real to real distance (RRD) of selected real record

Figure 4: Distances evaluated for given synthetic and given real records in privacy indicators. Most indicators involve computing at least SR and RR distances for all synthetic and real data points. In (d), the record is in D_1 and is therefore only compared to real records in D_2.

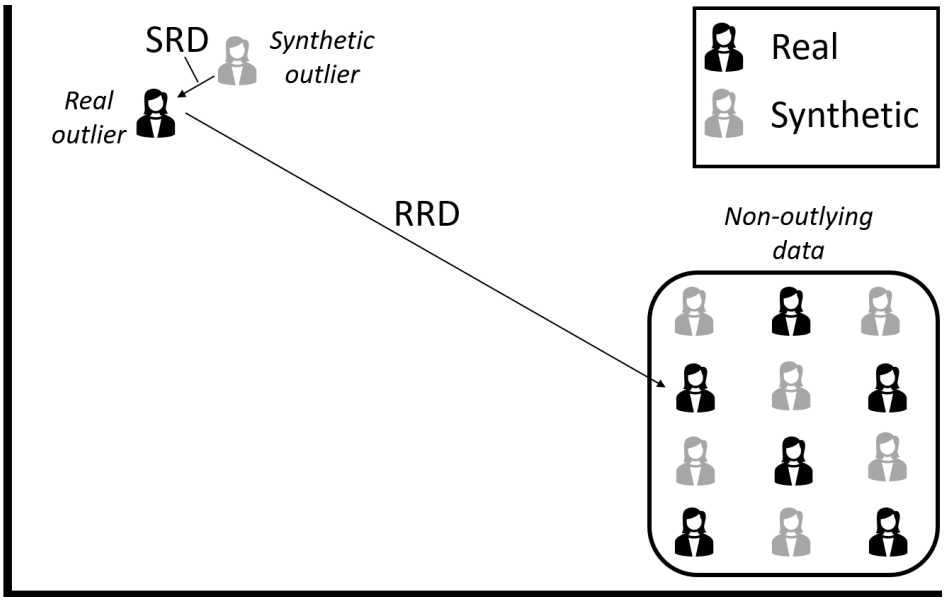


Figure 5: The DCR is sensitive to outliers: if the SD is accurate, it reproduces an outlier similar to the real outlier. The distance between the synthetic and real outliers (SRD) is then small. By definition, the distance between the real outlier and the closest other real record (RRD) is relatively large. Thus, $\texttt{SRD}(\hat{d}) < \texttt{RRD}(d^*)$.

Method	**IMS**	**Distance based**	Similarity metric	Evaluation method	Statistic(s)	Holdout?	**NN**	**Other**
[73]	-	-	-	-	-	-	-	Statistics
[74]	-	-	-	-	-	-	-	MMD
[75]	-	-	-	-	-	-	-	MMD
[40]	Yes	-	-	-	-	-	-	-
[11]	Yes	-	-	-	-	-	-	-
[14]	Yes	-	-	-	-	-	-	-
[124]	-	PL	Euclidean	NS	PL	Yes	-	-
[68]	-	AA; PL	NS	NS	AA; PL	Yes	-	-
[65]	-	DCR	Euclidean	NS	Percentiles	No	Yes	-
[54]	Yes	DCR	NS	NS	Percentiles	NS	Yes	-
[125]	Yes	DCR	Euclidean	NS	μ, σ	No	-	-
[35]	-	DCR	$\mathcal{L}_1$ (num) Hamming (cat)	Aggr.	Histogram	No	-	-
[126]	-	DCR	NS	NS	Histogram	Yes	-	-
[64]	-	DCR	Euclidean	Ign.	μ, σ	No	-	-
[29]	-	DCR	Euclidean	NS	Median	No	-	-
[30]	-	DCR	NS	NS	Median	No	-	-
[56]	Yes	DCR	NS	NS	Percentile	No	-	-
[127]	-	DCR	$\mathcal{L}_1$ (num) + Hamming (cat)	Aggr.	Histogram	No	-	-
[128]	-	DCR	Euclidean	NS	Percentile	No	Yes	-
[129]	-	DCR	Cosine	NA	μ	yes	-	-
[62]	Yes	DCR	Euclidean	NS	μ, σ	Yes	-	-
[55]	Yes	DCR	Hamming	Bin.	p, μ	Yes	Yes	-
[63]	-	DCR	Euclidean	Emb.	Percentile	Yes	-	-
[24, 61]	-	DCR	Euclidean	Emb.	p, μ, σ Inferential*	Yes	-	-
[70]	-	DBRL	Euclidean	NS	p	No	-	-
[71]	-	DBRL	Euclidean; Manhalanobis	NS	p	No	-	-
[32]	Yes	DCR	Euclidean	Emb.	Percentile	No	Yes	Seed

Table 3: Indicators used in practice; "NS": Not specified; "num": for numeric attributes; "cat": for categorical attributes; "Aggr.": aggregating two metrics (one for numerical and one for categorical attributes); "Ign.": ignoring categorical attributes; "Emb.": evaluating indicators in an embedding space; "bin.": binning numeric attributes; "NA": Not applicable, for instance because data types are not mixed in the dataset(s) of the involved study; p: proportion; μ: mean; σ: standard deviation; seed: seed-specific distance-based indicators other than DBRL, e.g. local cloaking; hidden rate. *: the authors use a Kolmogorov-Smirnov test to test the null-hyopthesis: "the SRD and RRD distributions stem from the same underlying distribution", with α levels of 0.05 and 0.01.

D Empirical Privacy Assessment Frameworks

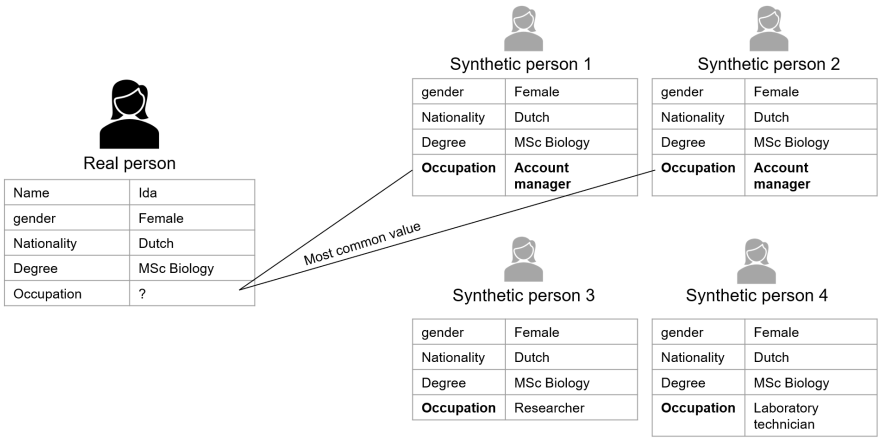


Figure 6: Nearest neighbor attack based on nearest neighbors (NN). The adversary knows Ida's gender, nationality and degree, but not her occupation. The adversary has access to a synthetic dataset. In this synthetic set, the adversary searches for the k (in this case four) records with the most similar profile based on the known attributes. The adversary then infers Ida Jansen's occupation based on the k closest synthetic records. The synthetic record's values can be aggregated in a number of ways. E.g. Ida can be assigned the most common value of the neighbors (in this case: account manager). For numeric attributes, averages can also be taken.

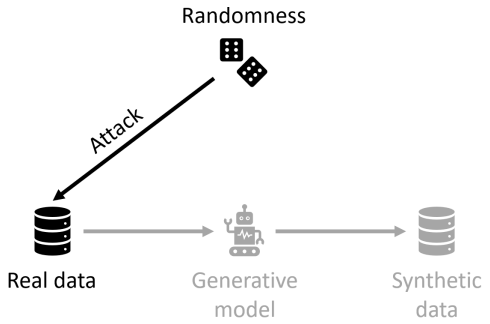


Figure 7: Random baseline to benchmark successful attacks against (cf. Figure 2)

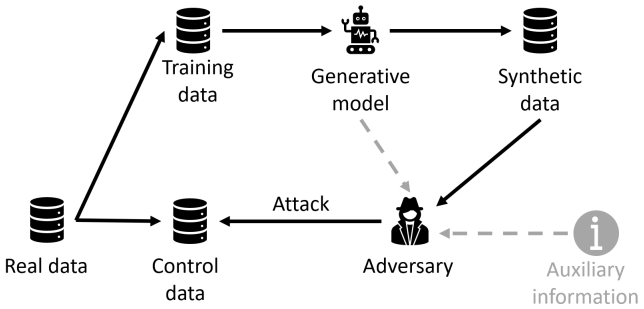


Figure 8: Control baseline: the adversary uses the available information to target a control data, not used in training the generative model

Method	Threat model	Baseline	Attack estimator	Attack technique	Attack type (WP29)
[77]	No box	SL (IT)	IT	VRD	S
[95]	Aux	A	IT	VRD	L
[81]	Aux	SL	NN	VRD	I*
[89]	Black; aux	SL (PR)	ML	SM; MIA	S, I*
[88]	Aux	SL	ML	MIA	S, I*
[76]					
Singling out	No box	R; G	-	VRD	S
Linkage	Aux	R; G	NN	VRD	L
Inference	Aux	R; G	NN	VRD	I*
[43]					
Neighborhood	Aux	R	NN	VRD	I*
Inference	Aux	R	ML	VRD	I*
Shadow model	Black	R	ML	SM; MIA	S, L, I
[86]					
Attribute inf.	Aux	SL	NN	VRD	I*
Membership inf.	No box	SL	NN	MIA	S, I
[68]	Black; aux	SL	NN	MIA	S*
[87]	Black; aux	M (DP)	ML	MIA	L
[83]	White; aux	SL (AUC,F1)	ML	SM; MIA	S*
[84]	Black; aux	SL (AUC)	NN; ML	SM; MIA	L
[82]	Black	SL (F1)	ML	SM; MIA	S
[85]					
Limited aux	Aux	R	ML	SM; MIA	S*
Aux	Aux	R	ML	SM; MIA	L
[90]	Black	SL (AUC)	ML	VRD; MIA	S; I
[29]	No box	A	NN	VRD; MIA	I

Table 4: Empirical privacy assessment frameworks. Threat model: aux - auxiliary information; black, white - black box, white box; Mult: experiments conducted with multiple threat models; Baseline: A - absolute (proportion or quantity of correct hypotheses, etc.), M - privacy metric (k - k-anonimity, DP - differential privacy), R - random, C - control, SL - metrics from supervised learning (PR - precision and recall, AUC - area under ROC curve, F1 - F1-score); Attack estimator: IT - information theory, NN - nearest neighbor, ML - machine learning; Attack technique: VRD - vulnerable record discovery through sorting, searching or sampling, SM - shadow modeling, MIA - membership inference attack; Attack type (WP29): S - singling out; L - linkage; I - inference, *: technically, any attack using auxiliary information is a linkage attack to some degree.