Empirically testing privacy in machine learning

A review and some challenges for practice[*]

—draft; comments are welcome—

Flavio Hafner[†]

Chang Sun[‡]

March 29, 2024

Abstract

How reliably do synthetic data from machine learning generators protect the privacy of the original data? We first review the empirical testing of privacy in machine learning, focusing on the intuitions for and assumptions behind such tests. We then discuss whether and how these tests could be used when evaluating generative algorithms for deployment at statistical agencies and in the health care system. Privacy audits assume strong adversaries and can identify bugs in the training algorithm. They could thus become standard in ensuring the correct implementation of a given algorithm. Membership inference attacks assume realistic threat scenarios, but they do not scale in the size of the training data. Their ability to consider the specific context of releasing privacy-sensitive models could make them informative for practitioners—if the scaling challenge is solved. We suggest some ideas for future work and discussion in this direction.

Keywords: empirical privacy estimation; differential privacy; machine learning

[*] This work was supported by ODISSEI—Open Data Infrastructure for Social Science and Economic Innovations.

[†] Netherlands eScience Center, `f.hafner@esciencecenter.nl`

[‡] Institute of Data Science, Maastricht University, `chang.sun@maastrichtuniversity.nl`

Research in the social and health sciences often relies on tabular data from statistical offices or from the health care system. Access to these data is restricted to protect sensitive personal information. In recent years, the machine learning community has explored ways to generate synthetic microdata that do not violate privacy protection and could thus be shared more widely—which could for instance make it easier for researchers to reproduce statistical analyses generated by others with these data sets (Mukherjee et al., 2023).

The technology enabling such synthetic data generators is Differential Privacy (DP) (Dwork and Roth, 2013), which has been applied to machine learning with the Differentially Private Stochastic Gradient Descent algorithm (DP-SGD, Abadi et al. (2016)). The algorithm provides a theoretical guarantee of privacy, but there are several reasons why an empirical counterpart could be useful.

First, DP-SGD is more complex than DP mechanisms for releasing aggregate data. Thus, before generators for synthetic micro data can be used in practice, they need to gain the trust of data owners (Cummings and Sarathy, 2023) and be carefully vetted for potential privacy leakage. Second, additional tests may be necessary, for instance because an adversary may have additional data sets available for a potential attack that are not considered in DP-SDG directly (Cummings et al., 2023). Further, the theoretical bound may be too conservative for realistic scenarios, and given the privacy-utility trade-off inherent to DP, relying on a more realistic threat model could boost the statistical utility of the algorithm.

An emerging method to quantify privacy leakage empirically are privacy attacks. There now exist software packages to perform such tests on generative (Houssiau et al., 2022*a*; Qian, Cebere and van der Schaar, 2023) or predictive models (Kumar and Shokri, 2020), making it easy to use these tests for benchmarking generators.

Conducting these tests and interpreting their results, however, requires a good understanding of the method and the underlying assumptions, particularly because they always only approximate the theoretical level of privacy, and some tests may require an impractically large number of computations. In our experience, these considerations are not straightforward to grasp for people unfamiliar with the research field. Moreover, it is an open question to what extent (large-scale) privacy testing should and could be relied upon in practice in the first place (Jagielski, Ullman and Oprea, 2020; Yoon, Drumright and Van Der Schaar, 2020).

Our paper thus aims to be an entry point for people familiarizing themselves with these methods and to stimulate discussion from stakeholders such as statistical agencies and researchers in the social and health sciences. We also want to take stock of the rapid development of the literature in recent years—because with it have emerged both computationally feasible privacy audits and evidence of bugs in implementations of supposedly privacy-protecting machine learning algorithms (Nasr et al., 2023; Stadler, Oprisanu and Troncoso, 2022).

In addition, even within machine learning community there is no standardized approach to evaluating the privacy-preserving properties of newly proposed algorithms—although some of these pre-date the development of privacy testing we discuss. For instance, in the context of generative adversarial networks, evaluations range from sophisticated membership inference attacks (Park et al., 2018) to scoring on the distance to the closest synthetic record (Xu and Veeramachaneni, 2018; Sun, van Soest and Dumontier, 2022; Choi et al., 2017) to no privacy evaluation (Fang, Dhami and Kersting, 2022; Xie et al., 2018).[1]

Our synthesis focuses on centralized machine learning with a trusted third party (as opposed to federated learning). Compared to existing reviews of differential privacy (Cummings et al., 2023; Ponomareva et al., 2023), we focus on a very particular aspect of the topic, but explain in more detail the "how" behind empirical privacy testing. Moreover, focusing on generative models, we provide novel suggestions for future work compared to the existing reviews that focus on predictive models. Our discussion of privacy testing is closely related to Houssiau et al. (2022*a*), but we additionally discuss—from a practical perspective—the assumptions behind and challenges for the tests, and include more recent methods as well as privacy audits.

[1] Instead of using differentially private stochastic gradient descent, Yoon, Drumright and Van Der Schaar (2020) modify the loss function when training the GAN model.

We first synthesize the literature of empirical privacy testing with membership inference attacks and privacy audits, which build on the hypothesis testing interpretation of differential privacy. Our synthesis is generic and applies to tests both for predictive and generative machine learning models. Methods for testing predictive models have progressed faster than for generative models, but we believe there are lessons to be applied from the former to the latter.

We then discuss practical challenges when using the tools from the empirical privacy testing literature for evaluating algorithms for deployment. We argue that current membership inference attacks with realistic adversaries are of little practical value. The reason is that if the goal is to assess the vulnerability of each record in the input data for privacy risk—which we believe could be a potential request from a data owner—, attacks with realistic adversaries do not scale in the size of the training data. This is related to a well-known trade-off between how realistic a privacy test is and its computational requirements. But it is worth emphasizing the importance of this aspect for use-cases at statistical agencies, where there are easily millions of records—in contrast to widely used data sets in some of the papers we discuss.[2]

In contrast, privacy audits for DP-SGD test whether the algorithm is implemented correctly. Because the audits scale much better in the size of the input data, we believe they could be considered for usage in a manner similar to how integration tests are used in software development.

Despite the current scaling challenges, membership inference attacks with realistic adversaries allow to take the specific context of a synthetic data release into account. This could still make them a useful tool if the scaling challenges are overcome, to which end we discuss some ideas for future work.

Theory

We start with the standard definition of differential privacy (DP):

[2] For instance, the CIFAR-10 dataset used in Carlini et al. (2022) has 60'000 images.

Definition 1 (Differential Privacy, Dwork and Roth (2013))**.** *A randomized algorithm $\mathcal{M}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subset Range(\mathcal{M})$ and for all neighboring databases D and D' that differ by at most one record:*

$$\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{S}] + \delta$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$.

Intuitively, this means that the distributions over all the outcomes $\mathcal{S}$ are similar when the input data sets D and D' only vary slightly—by one record. Higher privacy is associated with a lower ε: It provides an upper bound on the ability to distinguish the output of $\mathcal{M}(D)$ from the output of $\mathcal{M}(D')$; this bound fails to hold with probability δ. Further, a key property of differential privacy is that it is immune to postprocessing: any operation performed on the output of $\mathcal{M}$ has no worse privacy guarantees than the output of $\mathcal{M}$.

There are other definitions of differential privacy, and some allow one-to-one correspondences to the definition provided here. For instance, both Renyi-DP (Ponomareva et al., 2023, section 4.2.2) and cases of Gaussian DP (Dong, Roth and Su, 2019; Nasr et al., 2023) are convertible to (ε, δ)-DP.

There is a well-known trade-off between privacy and statistical utility: the higher the level of privacy protection, the more noise is added to the original data and therefore the less useful they are to draw statistical inference from.

Differentially private machine learning Both predictive and generative machine learning models could leak sensitive information about the training data: either from the model's predictions, or from the synthetic data generated by the model. For this reason, algorithms exist that train machine learning with differential privacy.

Currently, the main approach for training neural networks with DP is *Differentially Private Stochastic Gradient Descent (DP-SGD)*. It works by clipping the gradients of each individual sample to a maximum norm, and infusing Gaussian noise to the aggregated batch-level gradient. This yields the exact Gaussian mechanism at the batch level.

However, for computing the privacy loss from an end-to-end training algorithm, it is necessary to aggregate the privacy loss from each training step, taking also into account the randomization of records into different batches in each epoch (Nasr et al., 2023).

Because training requires access to the individual-level gradients, it is less able to exploit parallelism of the GPU and therefore slower. For convex objective functions, there are other methods that can yield much better statistical utility at high privacy levels (Ponomareva et al., 2023).

DP-SGD works with a threat model on the adversary that is often not discussed explicitly, but it has implications for how the results of privacy audits are interpreted: The adversary sees all gradients and parameter update steps during training; these are used to calculate the algorithms' level of privacy protection.

The hypothesis testing interpretation of differential privacy In order to connect the formal guarantees to empirical measures of privacy, one can interpret DP from a statistical hypothesis testing perspective (Wasserman and Zhou, 2009; Kairouz, Oh and Viswanath, 2015). Specifically, given an output Y from a randomized mechanism $\mathcal{M}$, consider the two hypotheses:

$$\begin{aligned} &H0 : Y \text{ was drawn from } \mathcal{M}(D) \\ &H1 : Y \text{ was drawn from } \mathcal{M}(D') \end{aligned} \tag{1}$$

If $\mathcal{M}$ satisfies (ε, δ)-DP, then an attacker attempting to distinguish between the two hypotheses faces a trade-off between type-I error rate α and type-II error rate β:

$$\begin{aligned} \alpha + e^{\varepsilon}\beta \geq 1 - \delta \\ e^{\varepsilon}\alpha + \beta \geq 1 - \delta \end{aligned}$$

Thus, if one can empirically conduct this hypothesis test, and given a value for δ, it possible to estimate a lower bound on ε with two methods. First, the system of inequalities implies

$$e^{\varepsilon}(\delta) \geq \max\left(\frac{1-\alpha-\delta}{\beta}, \frac{1-\beta-\delta}{\alpha}\right)$$

Second, the system also implies an upper bound on the accuracy of a binary classification problem where

$$\mathrm{acc}(\delta) \leq \frac{e^{\varepsilon}+\delta}{1+e^{\varepsilon}}$$

Because the lower bound is derived for a particular attack, other attacks may be developed in the future that have higher accuracy and thus uncover a higher privacy leakage of $\mathcal{M}$. Moreover, as with any statistical inference, computing a lower bound for ε comes with its own sampling uncertainty, which can be estimated with Clopper-Pearson confidence intervals.

The hypothesis testing interpretation is the foundation for membership inference attacks and privacy audits, which we describe next.

Application: Privacy attacks against machine learning

Background

Privacy attacks against machine learning empirically implement the hypothesis test interpretation of differential privacy. Shokri et al. (2017) was the first membership inference attack against a predictive machine learning model. Such attacks test whether an adversary can, under some more or less realistic assumptions, make inferences about the training data. Based on this work, membership inference attacks against generative machine learning were developed; Hu et al. (2022) and Chen et al. (2020) provide reviews.

Moreover, it turned out that the empirically estimated privacy protection is often higher than what is implied by the theoretical analysis of DP-SGD (Jagielski, Ullman and Oprea, 2020). Because of the trade-off between privacy and statistical utility, the important question arose whether this is because the theoretical analysis of DP-SGD is not tight, ie, that training DP-SGD with ε_{th} yields a model that has a higher privacy guarantee than what can be theoretically proven. This has been an active area of research, but recent audits can nearly recover the theoretical privacy guarantee of DP-SGD with strong adversaries and realistic data sets (Nasr et al., 2023) or strong adversaries and worst-case data sets (Nasr et al., 2021).

We will first introduce the underlying principles of empirical privacy testing. Research on auditing generative algorithms is less developed; but there are many conceptual overlaps between predictive and generative models, so we distinguish them only when necessary. We will then discuss which methods could be useful for practitioners.

Building blocks of attacks

Privacy attacks work by simulating an adversary that tries to break the promise of privacy of algorithm $\mathcal{M}$. We leave the output of $\mathcal{M}$ unspecified; but typically for predictive models it is a vector of predicted probabilities or a scalar of the predicted class. For generative models, it is for instance a new dataset. This simulation requires certain assumptions on the adversary that may not be realistic in all contexts.

A privacy attack needs to define the following ingredients.

An attack target The notion of "neighboring" databases requires that there is one record that differs between two otherwise identical databases.

Knowledge about the training data. The adversary needs some data to train their shadow models. Thus, they have at least a random sample of records that are drawn from the same distribution as the training data of the original model $\mathcal{M}$.

Knowledge about the algorithm $\mathcal{M}$. In all cases, the adversary has at least query access to $\mathcal{M}$. In a predictive model, this means they can feed $\mathcal{M}$ with new data and get back an output; in a generative model, this means they can use $\mathcal{M}$ to create a new synthetic data set of any size. This is perhaps the first counter-intuitive assumption for attacks on synthetic data where, in a realistic scenario, not the trained model but only a synthetic data set is released. To our knowledge, only in Xu et al. (2022) the adversary has access not to the generator, but only to the synthetic dataset.

A second important and sometimes counter-intuitive assumption is that the adversary knows the architecture of $\mathcal{M}$, or at least can train a new model with the same architecture. The assumption is explicitly discussed in early work (Shokri et al., 2017) but not in more recent work. This assumption is stronger in some contexts than in others. For instance, in Shokri et al. (2017) the adversary attacks the prediction of an online machine-learning service, and can therefore train a new model with the same architecture by using the service. In the context of releasing synthetic personal data, however, it is less obvious that the adversary knows how the generator was trained—yielding a stronger adversary than is perhaps realistic.

The strongest adversaries also know all the gradients from training $\mathcal{M}$ and the resulting parameters. This is referred to as "white-box model access", and corresponds to the threat model implicitly used in DP-SGD.

Shadow modeling and attack logic

The adversary wants to find out whether the target record was included for training of the released algorithm $\mathcal{M}$. They do so with a decision rule $\mathcal{B}$; to learn the rule, training data are necessary.

Therefore, the adversary generates a set of tuples $\{(b_t, \mathcal{M}_t)\}_{t=1}^{T}$. To train $\mathcal{M}_t$, the adversary samples a training data set from the data distribution. The target record is not in this training data set. The adversary also draws the random variable $b_t \in \{0, 1\}$. Then, only if $b_t = 1$, the adversary adds the target record to the training set. The training set is then use to train $\mathcal{M}_t$.[3]

The adversary now uses the tuples to learn the decision rule $\mathcal{B}$. The more the inclusion of the target record x_0 impacts the gradients during training of $\mathcal{M}$, the more different the learned model $\mathcal{M}_t$ and therefore the easier to distinguish are the models with $b_t = 0$ from models with $b_t = 1$. There are different ways to obtain $\mathcal{B}$, and they differ between generative and predictive models.

To learn $\mathcal{B}$ for predictive models, the idea is that $\mathcal{M}$ performs differently on samples

[3] Ye et al. (2022) show that different attacks against predictive models use different sources of randomness, making attacks difficult to compare.

that were used for training than on samples that were not used for training, and this is measured through the record's loss. For a review of attacks against predictive and generative models, we refer the reader to Hu et al. (2022). We instead focus on insights from Carlini et al. (2022) and Ye et al. (2022).

First, earlier attacks derived one decision rule $\mathcal{B}$ for multiple target records.[4] Such attacks inform about the average privacy risk of a record in the training data, but are not informative about the privacy risk of an individual data record. For the latter, one has to train an attack against that particular record—in other words, the target record x_0 needs to be fixed across shadow models $(\mathcal{M}_1, ..., \mathcal{M}_T)$.

Second, even when holding the target record fixed, the adversary can face different sources of uncertainty (Ye et al., 2022): There is not only randomness from the starting parameters of $\mathcal{M}_t$ (through the model's seed), but also from the training data besides the target record. In other words, each shadow model may or may not be trained on the same training data (up to the target record). The more uncertainty, the weaker the adversary and therefore the higher appears privacy as measured through the effective ε.

Third, recent attacks appear to converge towards using a likelihood ratio (LR) test for membership inference (Carlini et al., 2022). LR tests provide the highest probability of correctly rejecting the null hypothesis in (1) at fixed error rates, and thus the strongest possible adversary given the threat model. But they require estimating a distribution for the losses of target models when the target record is included in training and when it is not included in training. Zarifzadeh, Liu and Shokri (2023) make some progress on saving computation.

Membership inference attacks against tabular data synthesizers

Membership inference attacks against generative models for tabular data also use shadow modeling to create a set of synthetic shadow datasets, labeled by whether the target record was included in the training data or not. The decision rule $\mathcal{B}$ then needs to distinguish between two groups of datasets. $\mathcal{B}$ is usually trained on statistics derived from

[4] Chen et al. (2020) and Hayes et al. (2017) train a single threshold to attack multiple training records in a generative model.

the synthetic shadow datasets. The python library of Houssiau et al. (2022*a*) contains attacks against tabular data synthesizers and allows to measure individual privacy risks.

The simplest attack compares the distance between the target record and the closest record in the synthetic shadow datasets. Given a distance metric, the attacks learn a threshold, below which the target record is predicted to be included in the training data. Stadler, Oprisanu and Troncoso (2022) propose a stronger attack by projecting the synthetic dataset onto a lower-dimensional feature space—in the simplest form, summary statistics and histograms of marginal distributions. $\mathcal{B}$ is then trained on these lower-dimensional features.

Special case: privacy audits

Privacy audits (Jagielski, Ullman and Oprea, 2020) are a special case of the hypothesis testing framework in that they design distinguishing experiments that maximize the ability to distinguish between the two hypotheses with the lowest possible type I and type II errors.

Privacy audits have two motivations. First, empirical tests of privacy leakage tend to find values for ε that are lower than the theoretical guarantee derived from DP-SGD. This could be either because the assumed adversary is too weak, or because the accounting of privacy loss in DP-SGD is too conservative, and DP-SGD effectively protects privacy better than what can be proven theoretically. But if this was the case, there would be a free lunch: one could train DP-SGD with a higher formal epsilon, without increasing effective epsilon beyond the acceptable level, and achieve higher statistical utility. In contrast, finding an adversary for whose attack the effective epsilon coincides with the theoretical upper bound would prove that the privacy guarantee derived for DP-SGD is tight and no such free lunch is available.

The second motivation for privacy audits is to directly check whether DP-SGD is implemented correctly: Finding an effective epsilon above the theoretical lower bound would identify a bug in DP-SGD.

Privacy audits design the strongest adversary as being free to choose not only the neighboring datasets but also the target record: they can inject any kind of target record into $\mathcal{M}$—it does not have to be drawn from the original data. This is in line with the definition of differential privacy since it needs to hold for *any* neighboring datasets, and the differing record is not specified. Some privacy audits refer to this record as a *canary*.

Nasr et al. (2021) show that the privacy guarantee in DP-SGD is tight: for a worst-case adversary, their audit recovers the theoretical upper bound on ε. However, their adversary works with a pathological data set that is the empty set.

Nasr et al. (2023) improve on this result and show that even an adversary with a realistic data set can extract only slightly less information than what is theoretically the upper bound. Because DP-SGD is composed of applying *Gaussian* differential privacy multiple times, they exploit the results from Dong, Roth and Su (2019) to audit the learning mechanism with Gaussian DP. This allows them to audit the algorithm in a computationally more efficient way than previous work.

Practical issues and ideas for future research

We now summarize the insights from our experience and from the literature review.

Membership inference attacks with realistic adversaries do not scale to worst-case analysis

Differential privacy needs to hold for all potential target records in the training data—thus, a credible test ought to report privacy guarantees for each record. Performing such a test has a time-complexity of

$$N \times (T \times O_{\mathcal{M}}(N) + O_{\mathcal{B}}(T))$$

where N is the size of the raw data, T is the number of shadow models trained per record, and $O_{\mathcal{M}}(N)$ and $O_{\mathcal{B}}(T)$ denote the time complexity of the generator and the attack algorithms, respectively. Even with linear-time training algorithms, this implies at least a quadratic complexity in the size of the raw data. Thus, in many circumstances, such an assessment is not computationally feasible—not only for real-time deployment (Cummings et al., 2023), but also for benchmarking different algorithms on datasets with millions of records, as is common in registry data. While Steinke, Nasr and Jagielski (2023) propose a method to attack multiple records in a single training run, their attack does not use the likelihood-ratio test. It is thus unclear how informative this attack is in practice.

One may consider running the attack only on a few selected target records. Selecting outliers based on their loss of a model trained on the full data (Houssiau et al., 2022*a*) is not informative because what matters is not the level of the loss when the target is included, but rather the difference in the loss on a model when the target record is included versus not. In other words, a record may be hard to fit, in which case the loss on that record is high irrespective of whether the target was included in the training data or not (Carlini et al., 2022).

Alternatively, one may make a conscious choice about which records should be protected under a more realistic threat model, train the model accordingly with a lower theoretical privacy guarantee, and run the privacy attacks only on those selected records. But then privacy may be leaked on other records whose vulnerability is not tested: Because non-parametric models such as generative adversarial networks (Stadler, Oprisanu and Troncoso, 2022) learn an implicit density function, it becomes unpredictable ex-ante which dimension of the data will be represented by the model. This means that the vulnerability of a record that is an outlier on a particular dimension may go unnoticed by this privacy attack.

Lastly, even if there was a method that allows to cheaply select the training records with the highest risk of privacy leakage after a model has been trained, the method will need to be doing so in a mathematically rigorous and provable way.

Privacy audits of DP-SGD could be adopted at the level of the algorithm

In contrast, we believe privacy audits with strong adversaries are useful to guarantee that a particular algorithm implements DP-SGD correctly. The audit in Nasr et al.

(2023) requires two training runs—reaching much better time complexity than the black-box membership inference attacks, although they may still be unfeasible for the largest models.

These audits test whether DP-SGD is correctly implemented, and Nasr et al. (2023) discuss various reasons what can go wrong when implementing DP-SGD. First, at the level of each gradient step, it is possible that (i) gradients are not clipped at the level of the individual sample, (ii) the added noise is not generated randomly, and (iii) the scale of the added noise is not proportional to the batch size—for instance, when gradients are computed across multiple machines.

Second, to audit the full training across all batches and epochs, one can approximate the trade-off function—a function tracing out the adversary's lowest achievable type II error for all type I error rates (Dong, Roth and Su, 2019)—generated by DP-SGD with the method from Koskela, Jälkö and Honkela (2019). However, because of minibatch subsampling during training, this approximation may overstate the actual privacy (leading to a lower bound on ε that not tight) of the algorithm, particularly at higher false positive rates—and therefore not find all implementation bugs from the end-to-end perspective. This suggests that such audits should use a threshold with a low false positive rate.

The audits operate at the level of a training algorithm, and are "largely independent" (Nasr et al., 2023) from the training data. For this reason, they can be seen as analogous to an integration test in software engineering, which are essential to ensure the correct working of software.

How to make black-box membership inference attacks against generative models workable?

We do believe that black-box models are useful and complementary to the white-box threat models and privacy audits—precisely because the former depend on context, they are easier to communicate to data owners and give more realistic estimates of privacy leakage, promising to deliver higher statistical utility than a reliance on unnecessarily high privacy guarantees from white-box threat models. To overcome the challenges described above, we suggest three areas where future research and exploration could be worthwhile.

Use and develop threat models specific to context. Each membership inference attack will need to be tailored to the specific use-case, with the required assumptions on the adversary. As a result, it appears impossible to have a general benchmark that compares different generators. Rather, before a generative algorithm is deployed, it must be chosen according to the specific use-case. If no attacks for the given use-case exist, they need to be developed.

For a given threat model, use the strongest adversary possible. In particular, this means that the attacks with likelihood ratio tests proposed for predictive models should also be evaluated for attacks against generative models. And this should be possible, at least as long as the generator has a unit-level loss function as is the case for generative adversarial networks. Hayes et al. (2017) is an early example of using per-example loss of the discriminator in a GAN as data for an attack, and this could be merged with the attack in Carlini et al. (2022).

Find ways to overcome the scaling problem with current non-parametric models. Finally, to be able to empirically test the privacy leakage of generators with black-box threat models, the scalability problem of current attacks needs to be overcome. This problem could be tackled by solving two of its underlying sources. First, generators with more predictable behavior than non-parametric models may be easier to privacy-proove because it is clear which dimension of the data will be vulnerable for privacy leakage. Houssiau et al. (2022*b*) is some work in this direction. Second, new training approaches and architectures that incur the large fixed-cost of training only once could be promising. One example here is using public data to train a foundation model, and then using the private data only to fine-tune the generator to the specific use-case. This approach has been proposed for language models (Tramèr, Kamath and Carlini, 2022), but for population and health data it is unclear if existing publicly available data—such as census tables at geographic disaggregations—contain enough information to do this.

Conclusion

We have synthesized the literature on empirical privacy testing and discussed some practical challenges with using these tests in practice. Our first conclusion is that there is a need to develop workflows and software that make privacy audits easy to use in the learning pipeline. This could foster the adoption of such audits in the machine learning community, and is an important first step towards a wider use of DP-SGD in applications.

Our second conclusion is a need for dialogue and research about whether and how the privacy audits and privacy attacks should be used in practice. Empirically quantifying privacy leakage is an important aspect to consider before deploying privacy-protective machine learning systems, possibly alongside other aspects (Cummings and Sarathy, 2023; Cummings et al., 2023). But relying only on the results from privacy audits makes such systems less useful in practice because the strong adversarial assumption implies low statistical accuracy. An important question for future research is therefore whether and which sacrifices on privacy leakage are acceptable to make DP-SGD suitable for opening data and models trained on sensitive data from statistical offices and from the health care system.

Despite this, black-box privacy attacks can be useful because they provide context. Solving the scaling problem for these attacks is an important area for future research.

References

Abadi, Martin, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. "Deep learning with differential privacy." 308–318.

Carlini, Nicholas, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. "Membership Inference Attacks From First Principles."

Chen, Dingfan, Ning Yu, Yang Zhang, and Mario Fritz. 2020. "GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models." 343–362.

Choi, Edward, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. "Generating multi-label discrete patient records using generative adversarial networks." 286–305, PMLR.

Cummings, Rachel, and Jayshree Sarathy. 2023. "Centering Policy and Practice: Research Gaps Around Usable Differential Privacy."

Cummings, Rachel, Damien Desfontaines, David Evans, Roxana Geambasu, Matthew Jagielski, Yangsibo Huang, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii, Yu-Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, and Wanrong Zhang. 2023. "Challenges towards the Next Frontier in Privacy."

Dong, Jinshuo, Aaron Roth, and Weijie J. Su. 2019. "Gaussian Differential Privacy."

Dwork, Cynthia, and Aaron Roth. 2013. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science*, 9(3-4): 211–407.

Fang, Mei Ling, Devendra Singh Dhami, and Kristian Kersting. 2022. "Dp-ctgan: Differentially private medical data generation using ctgans." 178–188, Springer.

Hayes, Jamie, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2017. "Logan: Membership inference attacks against generative models." *arXiv preprint arXiv:1705.07663.*

Houssiau, Florimond, James Jordon, Samuel N. Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. 2022*a*. "TAPAS: A Toolbox for Adversarial Privacy Auditing of Synthetic Data."

Houssiau, Florimond, Samuel N. Cohen, Lukasz Szpruch, Owen Daniel, Michaela G. Lawrence, Robin Mitra, Henry Wilde, and Callum Mole. 2022*b*. “A Framework for Auditable Synthetic Data Generation.” *https://arxiv.org/abs/2211.11540v1*.

Hu, Hongsheng, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. “Membership Inference Attacks on Machine Learning: A Survey.”

Jagielski, Matthew, Jonathan Ullman, and Alina Oprea. 2020. “Auditing Differentially Private Machine Learning: How Private Is Private SGD?”

Kairouz, Peter, Sewoong Oh, and Pramod Viswanath. 2015. “The Composition Theorem for Differential Privacy.”

Koskela, Antti, Joonas Jälkö, and Antti Honkela. 2019. “Computing Tight Differential Privacy Guarantees Using FFT.”

Kumar, Sasi, and Reza Shokri. 2020. “ML Privacy Meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning.”

Mukherjee, Soumya, Araktrika Mustafi, Aleksandra Slavković, and Lars Vilhuber. 2023. “Assessing Utility of Differential Privacy for RCTs.” *https://www.vilhuber.com/lars/wp-content/uploads/2023/05/CEGA_Gates_DP_RCT.pdf*.

Nasr, Milad, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. 2023. “Tight Auditing of Differentially Private Machine Learning.”

Nasr, Milad, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. 2021. “Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning.”

Park, Noseong, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. "Data synthesis based on generative adversarial networks." *arXiv preprint arXiv:1806.03384*.

Ponomareva, Natalia, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Thakurta. 2023. "How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy." *Journal of Artificial Intelligence Research*, 77: 1113–1201.

Qian, Zhaozhi, Bogdan-Constantin Cebere, and Mihaela van der Schaar. 2023. "Synthcity: Facilitating Innovative Use Cases of Synthetic Data in Different Data Modalities."

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. "Membership Inference Attacks against Machine Learning Models."

Stadler, Theresa, Bristena Oprisanu, and Carmela Troncoso. 2022. "Synthetic Data – Anonymisation Groundhog Day."

Steinke, Thomas, Milad Nasr, and Matthew Jagielski. 2023. "Privacy Auditing with One (1) Training Run."

Sun, Chang, Johan van Soest, and Michel Dumontier. 2022. "Improving Correlation Capture in Generating Imbalanced Data Using Differentially Private Conditional GANs."

Tramèr, Florian, Gautam Kamath, and Nicholas Carlini. 2022. "Considerations for Differentially Private Learning with Large-Scale Public Pretraining."

Wasserman, Larry, and Shuheng Zhou. 2009. "A Statistical Framework for Differential Privacy." , (arXiv:0811.2501).

Xie, Liyang, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. "Differentially private generative adversarial network." *arXiv preprint arXiv:1802.06739*.

Xu, Lei, and Kalyan Veeramachaneni. 2018. "Synthesizing tabular data using generative adversarial networks." *arXiv preprint arXiv:1811.11264*.

Xu, Yixi, Sumit Mukherjee, Xiyang Liu, Shruti Tople, Rahul Dodhia, and Juan Lavista Ferres. 2022. "MACE: A Flexible Framework for Membership Privacy Estimation in Generative Models."

Ye, Jiayuan, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. "Enhanced Membership Inference Attacks against Machine Learning Models."

Yoon, Jinsung, Lydia N Drumright, and Mihaela Van Der Schaar. 2020. "Anonymization through data synthesis using generative adversarial networks (ads-gan)." *IEEE journal of biomedical and health informatics*, 24(8): 2378–2388.

Zarifzadeh, Sajjad, Philippe Liu, and Reza Shokri. 2023. "Low-Cost High-Power Membership Inference by Boosting Relativity."