PROVABLE RE-IDENTIFICATION PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

In applications involving sensitive data, such as finance and healthcare, the necessity for preserving data privacy can be a significant barrier to machine learning model development. Differential privacy (DP) has emerged as one canonical standard for provable privacy. However, DP's strong theoretical guarantees often come at the cost of a large drop in its utility for machine learning; and DP guarantees themselves can be difficult to interpret. As a result, standard DP has encountered deployment challenges in practice. In this work, we propose a different privacy notion, re-identification privacy (RIP), to address these challenges. RIP guarantees are easily interpretable in terms of the success rate of membership inference attacks. We give a precise characterization of the relationship between RIP and DP, and show that RIP can be achieved using less randomness compared to the amount required for guaranteeing DP, leading to smaller drop in utility. Our theoretical results also give rise to a simple algorithm for guaranteeing RIP which can be used as a wrapper around any algorithm with a continuous output, including parametric model training.

1 INTRODUCTION

As the popularity and efficacy of machine learning (ML) have increased, the number of domains in which ML is applied has also expanded greatly. Some of these domains, such as finance or healthcare, are based on machine learning on sensitive data which cannot be publicly shared due to regulatory or ethical concerns (Assefa et al., 2020; Office for Civil Rights, 2002). In these instances, maintaining data privacy is of paramount importance and must be considered at every stage of the machine learning process, from model development to deployment. In development, even sharing data in-house while retaining the appropriate level of privacy can be a barrier to model development (Assefa et al., 2020). After deployment, the trained model itself can leak information about the training data if appropriate precautions are not taken (Shokri et al., 2017; Carlini et al., 2021a).

Differential privacy (DP) (Dwork et al., 2014) has emerged as the gold standard for provable privacy in the academic literature. Training methods for DP use randomized algorithms applied on databases of points, and DP stipulates that the algorithm's random output cannot change much depending on the presence or absence of one individual point in the database. These guarantees in turn give information theoretic protection against the maximum amount of information that an adversary can obtain about any particular sample in the database, regardless of that adversary's prior knowledge or computational power, making DP an attractive method for guaranteeing privacy. However, DP's strong theoretical guarantees often come at the cost of a large drop in utility for many algorithms. In addition, DP guarantees themselves are difficult to interpret by non-experts. For instance, there is a precise definition for what it means for an algorithm to satisfy DP with $\varepsilon = 10$, but it is not a priori clear what this definition guarantees in terms of practical questions that a user could have, the most basic of which might be to ask whether or not an attacker can determine whether or not that user's information was included in the algorithm's input. These issues hinder the widespread adoption of DP in practice.

In this paper, we propose a novel privacy notion, re-identification privacy (RIP), to address these challenges. RIP is based on re-identification, also called membership inference. Re-identification measures privacy via a game played between the algorithm designer and an adversary or attacker. The adversary is presented with the algorithm's output and a "target" sample $\mathbf{x}^*$, which may or may not have been included in the algorithm's input set. The adversary's goal is to determine whether or not the target sample was included in the algorithm's input. If the adversary can succeed with

probability much higher than random guessing, then the algorithm must be leaking information about its input. This measure of privacy is one of the simplest for the attacker; thus, provably protecting against it is a strong privacy guarantee. Furthermore, RIP is easily interpretable, as it is measured with respect to a simple quantity–namely, the maximum success rate of an attacker. In summary, **our contributions** are as follows:

- We propose a novel privacy notion, which we dub re-identification privacy (RIP).
- We characterize the relationship between RIP and differential privacy (DP).
- We introduce algorithms for generating RIP synthetic data.
- We demonstrate that certifying RIP can allow for much higher utility than certifying DP, and never results in worse utility.

2 Related Work

Privacy attacks in ML The study of privacy attacks has recently gained popularity in the machine learning community as the importance of data privacy has become more apparent. In a *membership inference* or *re-identification* attack (Shokri et al., 2017), an attacker is presented with a particular sample and the output of the algorithm to be attacked. The attacker's goal is to determine whether or not the presented sample was included in the training data or not. If the attacker can determine the membership of the sample with a probability significantly greater than random guessing, this indicates that the algorithm is leaking information about its training data. Obscuring whether or not a given individual belongs to the private dataset is the core promise of private data sharing, and the main reason that we focus on membership inference as the privacy measure. Membership inference attacks against predictive models have been studied extensively (Shokri et al., 2017; Baluta et al., 2022; Hu et al., 2022; Liu et al., 2022; He et al., 2022; Carlini et al., 2021a), and recent work has also developed membership inference attacks against synthetic data (Stadler et al., 2022; Chen et al., 2020).

In a reconstruction attack, the attacker is not presented with a real sample to classify as belonging to the training set or not, but rather has to *create* samples belonging to the training set based only on the algorithm's output. Reconstruction attacks have been successfully conducted against large language models (Carlini et al., 2021b). At present, these attacks require the attacker to have a great deal of auxiliary information to succeed. For our purposes, we are interested in privacy attacks to measure the privacy of an algorithm, and such a granular task may place too high burden on the attacker to accurately detect "small" amounts of privacy leakage.

In an attribute inference attack (Bun et al., 2021; Stadler et al., 2022), the attacker tries to infer a sensitive attribute from a particular sample, based on its non-sensitive attributes and the attacked algorithm output. It has been argued that attribute inference is really the entire goal of statistical learning, and therefore should not be considered a privacy violation (Bun et al., 2021; Jayaraman & Evans, 2022).

Differential privacy (DP) DP (Dwork et al., 2014) and its variants (Mironov, 2017; Dwork & Rothblum, 2016) offer strong, information-theoretic privacy guarantees. A DP (probabilistic) algorithm is one in which the probability law of its output does not change much if one sample in its input is changed. That is, if D and D' are two datasets (collections of n bounds) which differ in exactly one element, then the algorithm $\mathcal{A}$ is ε-DP if

$$\mathbb{P}(\mathcal{A}(D) \in S) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(D') \in S)$$

for any subset S of the output space. DP has many desirable properties, such as the ability to compose DP methods or post-process the output without losing guarantees. Many simple "wrapper" methods are also available for certifying DP. Among the simplest, the Laplace mechanism, adds Laplace noise to the algorithm output. The noise level must generally depend on the *sensitivity* of the base algorithm, which measures how much a single input sample can change the algorithm's output. The method we propose in this work is very similar to the Laplace mechanism, but we show that the amount of noise needed can be reduced drastically. Abadi et al. (2016) introduced DP-SGD, a powerful tool enabling DP to be combined with deep learning methods with only a small modification to the standard gradient descent training procedure. However, as previously mentioned, enforcing DP does

not come without a cost. Enforcing DP with high levels of privacy (small ε) often comes with sharp decreases in algorithm utility (Tao et al., 2021; Stadler et al., 2022). DP is also difficult to audit; it must be proven mathematically for a given algorithm implementation. Checking it empirically is generally computationally intractable (Gilbert & McMillan, 2018). The difficulty of checking DP has led to widespread implementation bugs (and even errors due to finite machine precision), which invalidate the guarantees of DP (Jagielski et al., 2020).

The independent work of Thudi et al. (2022) specifically applies DP to bound re-identification rates, and our results in Section 3.4 complement theirs on the relationship between re-identification and DP. However, our results show that DP is not *required* to prevent re-identification; it is merely one option, and we give alternative methods for defending against membership inference.

Auditing methods and metrics Another important component of synthetic data is privacy and utility *auditing*. This is especially crucial in regulated environments where users may be required to prove compliance of their tools with privacy regulations. Recent works (Alaa et al., 2022; Meehan et al., 2020) have proposed heuristics for measuring both synthetic data privacy and utility. Utility metrics are often based on statistical measures of similarity between the synthesized and real data (Yoon et al., 2020). Privacy metrics try to capture the notion of whether or not a generative model has "memorized" its training data, typically by looking at distances of the synthetic data to training data vs. some held out data. Most of the proposed distance-based heuristics fall victim to simple counter examples in which the proposed synthetic data scores perfectly on the privacy metric, but clearly does not preserve the privacy of the training data. On the other hand, RIP lends itself to useful empirical measurement, as the success rate of any existing membership inference attack method gives a lower bound on the best achievable privacy.

3 Re-Identification Privacy (RIP)

3.1 Notation

We make use of the following notation. We will always use $\mathcal{D}$ to refer to our entire dataset, which we assume consists of n samples all of which must remain private. We will use $\mathbf{x} \in \mathcal{D}$ or $\mathbf{x}^* \in \mathcal{D}$ to refer to a particular sample. $\mathcal{D}_{\text{train}} \subseteq \mathcal{D}$ refers to a size-k subset of our private data. We will assume is selected randomly, so $\mathcal{D}_{\text{train}}$ is a random variable. The remaining data $\mathcal{D} \setminus \mathcal{D}_{\text{train}}$ will be referred to as the holdout data. We denote by $\mathbb{D}$ the set of all size-k subsets of $\mathcal{D}$ (i.e., all possible training sets), and we will typically use $D \in \mathbb{D}$ to refer to a particular realization of the random variable $\mathcal{D}_{\text{train}}$. Finally, given a particular sample $\mathbf{x}^* \in \mathcal{D}$, $\mathbb{D}^{\text{in}}$ (resp. $\mathbb{D}^{\text{out}}$) will refer to those sets $D \in \mathbb{D}$ for which $\mathbf{x}^* \in D$ (resp. $\mathbf{x}^* \notin D$).

3.2 Theoretical Motivation

The implicit assumption behind the public release of any statistical algorithm–be it a generative or predictive ML model, or even the release of simple population statistics–is that it is acceptable for *statistical information about the modelled data* to be released publicly. In the context of membership inference, this poses a potential problem: if the population we are modeling is significantly different from the "larger" population, then if our algorithm's output contains any useful information whatsoever, it *should* be possible for an attacker to infer whether or not a given record could have plausibly come from our training data or not.

We illustrate this concept with an example. Suppose we wish to publish a model which predicts a patient's blood pressure from several biomarkers, specifically for patients who suffer from a particular chronic disease. To do this, we collect a dataset of individuals with confirmed cases of the disease, and use this data to train a linear regression model with coefficients $\hat{\theta}$. Formally, we let $\mathbf{x} \in \mathbb{R}^d$ denote the features (e.g. biomarker values), $z \in \mathbb{R}$ denote the patient's blood pressure, and $y = \mathbb{1}\{\text{patient has the chronic disease in question}\}$. In this case, the private dataset $\mathcal{D}_{\text{train}}$ contains only the patients with $y = 1$. Assume that in the general populace, patient features are drawn from a mixture model:

$$y \sim \text{Bernoulli}(p), \qquad \mathbf{x} \sim \mathcal{N}(0, I), \qquad z|\mathbf{x}, y \sim \theta_y^\top \mathbf{x}, \qquad \theta_0 \neq \theta_1.$$

In the re-identification attack scenario, an adversary observes a data point $(\mathbf{x}^*, z^*)$ and the model $\hat{\theta}$, and tries to determine whether or not $(\mathbf{x}^*, z^*) \in \mathcal{D}_{\text{train}}$. If θ_0 and θ_1 are well-separated, then an adversary can train an effective classifier to determine the corresponding label $\mathbb{1}\{(\mathbf{x}^*, z^*) \in \mathcal{D}_{\text{train}}\}$ for $(\mathbf{x}^*, z^*)$ by checking whether or not $z^* \approx \hat{\theta}^\top \mathbf{x}^*$. Since only data with $y = 1$ belong to $\mathcal{D}_{\text{train}}$, this provides a signal to the adversary as to whether or not $\mathbf{x}^*$ could have belonged to $\mathcal{D}_{\text{train}}$ or not. The point is that in this setting, this outcome is unavoidable if $\hat{\theta}$ is to provide any utility whatsoever. In other words:

In order to preserve utility, re-identification privacy must be measured with respect to the distribution from which the private data are drawn.

The example above motivates the following theoretical ideal for our synthetic data. Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ be the private dataset and suppose that $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}$ for some probability distribution $\mathcal{P}$. (Note: Here, $\mathbf{x}^*$ corresponds to the complete datapoint $(\mathbf{x}^*, z^*)$ in the example above.) Let $\mathcal{A}$ be our (randomized) algorithm, and denote its output by $\theta = \mathcal{A}(\mathcal{D})$. We generate a test point based on:

$$y^* \sim \text{Bernoulli}\,(1/2)\,, \qquad \mathbf{x}^* | y^* \sim y^* \text{Unif}(\mathcal{D}_{\text{train}}) + (1 - y^*)\mathcal{P},$$

i.e. $\mathbf{x}^*$ is a fresh draw from $\mathcal{P}$ or a random element of the private training data with equal probability. Let $\mathcal{I}$ denote any re-identification algorithm which takes as input $\mathbf{x}^*$ and the algorithm's output θ. The notion of privacy we wish to enforce is that $\mathcal{I}$ cannot do much better to ascertain the membership of $\mathbf{x}^*$ than guessing randomly:

$$\mathbb{P}_{\mathcal{A}, \mathcal{D}_{\text{train}}}(\mathcal{I}(\mathbf{x}^*, \mathcal{D}_{\text{synth}}) = y^*) \leq 1/2 + \eta, \quad \eta \ll 1/2. \tag{1}$$

3.3 Practical Definition

In reality, we do not have access to the underlying distribution $\mathcal{P}$. Instead, we propose to use a bootstrap sampling approach to approximate fresh draws from $\mathcal{P}$.

Definition 1 (Re-Identification Privacy (RIP))**.** *Fix $k \leq n$ and let $\mathcal{D}_{\text{train}} \subseteq \mathcal{D}$ be a size-k subset chosen uniformly at random from the elements in $\mathcal{D}$. For $\mathbf{x}^* \in \mathcal{D}$, let $y^* = \mathbb{1}\{\mathbf{x}^* \in \mathcal{D}_{\text{train}}\}$. An algorithm $\mathcal{A}$ is η-RIP with respect to $\mathcal{D}$ if for any identification algorithm $\mathcal{I}$ and for every $\mathbf{x}^* \in \mathcal{D}$, we have*

$$\mathbb{P}(\mathcal{I}(\mathbf{x}^*, \mathcal{A}(\mathcal{D}_{\text{train}})) = y^*) \leq \max\left\{\frac{k}{n}, 1 - \frac{k}{n}\right\} + \eta.$$

Here, the probability is taken over the uniformly random size-k subset $\mathcal{D}_{\text{train}} \subseteq \mathcal{D}$, as well as any randomness in $\mathcal{A}$ and $\mathcal{I}$.

Definition 1 states that given the output of $\mathcal{A}$, an adversary cannot determine whether a given point was in the holdout set or training set with probability more than η better than always guessing the a priori more likely outcome. In the remainder of the paper, we will set $k = n/2$, so that $\mathcal{A}$ is η-RIP if an attacker cannot have average accuracy greater than $(1/2 + \eta)$. This gives the largest a priori entropy for the attacker's classification task, which creates the highest ceiling on how much of an advantage an attacker can possibly gain from the algorithm's output, and consequently the most accurate measurement of privacy leakage. The choice $k = n/2$ also keeps us as close as possible to the theoretical motivation in the previous subsection. We note that analogues of all of our results apply for general k.

The definition of RIP is phrased with respect to *any* classifier (whose randomness is independent of the randomness in $\mathcal{A}$; if the adversary knows our algorithm and our random seed, we are doomed). While this definition is compelling in that it shows a bound on what any attacker can hope to accomplish, the need to consider all possible attack algorithms makes it difficult to work with technically. The following proposition shows that RIP is equivalent to a simpler definition which does not need to simultaneously consider all identification algorithms $\mathcal{I}$.

Proposition 2. *Let $\mathbb{A} = \text{Range}(\mathcal{A})$ and let μ denote the probability law of $\mathcal{A}(\mathcal{D}_{\text{train}})$. Then $\mathcal{A}$ is η-RIP if and only if*

$$\int_{\mathbb{A}} \max\left\{\mathbb{P}(\mathbf{x}^* \in \mathcal{D}_{\text{train}} \mid \mathcal{A}(\mathcal{D}_{\text{train}}) = A), \mathbb{P}(\mathbf{x}^* \notin \mathcal{D}_{\text{train}} \mid \mathcal{A}(\mathcal{D}_{\text{train}}) = A)\right\} d\mu(A) \leq \frac{1}{2} + \eta.$$

Furthermore, the optimal adversary is given by

$$\mathcal{I}(\mathbf{x}^*, A) = \mathbb{1}\{\mathbb{P}(\mathbf{x}^* \in \mathcal{D}_{\text{train}} \mid \mathcal{A}(\mathcal{D}_{\text{train}}) = A) \geq 1/2\}.$$

Proposition 2 makes precise the intuition that the optimal attacker should guess the more likely of $\mathbf{x}^* \in \mathcal{D}_{\text{train}}$ or $\mathbf{x}^* \notin \mathcal{D}_{\text{train}}$ conditional on the output of $\mathcal{A}$. The optimal attacker's overall accuracy is then computed by marginalizing this conditional statement.

Finally, RIP also satisfies a post-processing inequality similar to the classical result in DP (Dwork et al., 2014). This states that any local functions of a RIP algorithm's output cannot degrade the privacy guarantee.

Theorem 3. *Suppose that $\mathcal{A}$ is η-RIP, and let f be any (potentially randomized, with randomness independent of $\mathcal{D}_{\text{train}}$) function. Then $f \circ \mathcal{A}$ is also η-RIP.*

Proof. Let $\mathcal{I}_f$ be any re-identification algorithm for $f \circ \mathcal{A}$. Define $\mathcal{I}_{\mathcal{A}}(\mathbf{x}^*, \mathcal{A}(\mathcal{D}_{\text{train}})) = \mathcal{I}_f(\mathbf{x}^*, f(\mathcal{A}(\mathcal{D}_{\text{train}})))$. Since $\mathcal{A}$ is η-RIP, we have

$$\frac{1}{2} + \eta \geq \mathbb{P}(\mathcal{I}_{\mathcal{A}}(\mathbf{x}^*, \mathcal{A}(\mathcal{D}_{\text{train}})) = y^*) = \mathbb{P}(\mathcal{I}_f(\mathbf{x}^*, f(\mathcal{A}(\mathcal{D}_{\text{train}}))) = y^*).$$

Thus, $f \circ \mathcal{A}$ is η-RIP by Definition 1. □

For example, Theorem 3 is important for the application of RIP to generative model training: if we can guarantee that our generative model is η-RIP, then any output produced by it is η-RIP as well.

3.4 Relation to Differential Privacy

In this section, we make precise the relationship between RIP and the most common theoretical formulation of privacy: differential privacy (DP). We provide proof sketches for most of our results here; detailed proofs can be found in the Appendix. Our first theorem shows that DP is at least as strong as RIP.

Theorem 4. *Let $\mathcal{A}$ be ε-DP. Then $\mathcal{A}$ is η-RIP with $\eta = \frac{1}{1+e^{-\varepsilon}} - \frac{1}{2}$. Furthermore, this bound is tight, i.e. for any $\varepsilon > 0$, there exists an ε-DP algorithm against which the optimal attacker has accuracy $\frac{1}{1+e^{-\varepsilon}}$.*

Proof sketch. Let $p = \mathbb{P}(\mathbf{x}^* \in \mathcal{D}_{\text{train}} \mid \mathcal{A}(\mathcal{D}_{\text{train}}))$ and $q = \mathbb{P}(\mathbf{x}^* \notin \mathcal{D}_{\text{train}} \mid \mathcal{A}(\mathcal{D}_{\text{train}}))$ and suppose WLOG that $q \geq p$. We have $p + q = 1$ and by Proposition 6 below, $q/p \leq e^{\varepsilon}$. This implies that $q \leq \frac{1}{1+e^{-\varepsilon}}$, and applying Proposition 2 gives the desired result.

For the tightness result, there is a simple construction on subsets of size 1 of $\mathcal{D} = \{0, 1\}$. Let $p = \frac{1}{1+e^{-\varepsilon}}$ and $q = 1 - p$. The algorithm $\mathcal{A}(D)$ which outputs D with probability p and $\mathcal{D} \setminus D$ with probability q is ε-DP, and the optimal attacker has exactly the accuracy given in the theorem. □

To help interpret this result, we remark that for $\varepsilon \approx 0$, we have $\frac{1}{1+e^{-\varepsilon}} - \frac{1}{2} \approx \varepsilon/4$. Thus in the regime where strong privacy guarantees are required ($\eta \approx 0$), $\eta \approx \varepsilon/4$.

In fact, it is the case that DP is *strictly* stronger than RIP, which we make precise with the following theorem.

Theorem 5. *For any $\eta > 0$, there exists an algorithm $\mathcal{A}$ which is η-RIP but not ε-DP for any $\varepsilon < \infty$.*

Proof sketch. The easiest example is an algorithm which publishes each sample in its input set with extremely low probability. Since the probability that any given sample is published is low, the probability that an attacker can do better than guess randomly is low marginally over the algorithm's output. However, adding a sample to the input dataset changes the probability of that sample's being published from 0 to a strictly positive number, so the guarantee on probability ratios required for DP is infinite. □

In order to better understand the difference between DP and RIP, let us again examine Proposition 2. Recall that this proposition showed that *marginally* over the output of $\mathcal{A}$, the conditional probability that $\mathbf{x}^* \in \mathcal{D}_{\text{train}}$ given the synthetic should not differ too much from the unconditional probability that $\mathbf{x}^* \in \mathcal{D}_{\text{train}}$. The following proposition shows that DP requires this condition to hold for *every* output of $\mathcal{A}(\mathcal{D}_{\text{train}})$.

Proposition 6. *If $\mathcal{A}$ is an ε-DP synthetic data generation algorithm, then for any $\mathbf{x}^*$, we have*

$$\frac{\mathbb{P}(\mathbf{x}^* \notin \mathcal{D}_{\text{train}} \mid \mathcal{A}(\mathcal{D}_{\text{train}}))}{\mathbb{P}(\mathbf{x}^* \in \mathcal{D}_{\text{train}} \mid \mathcal{A}(\mathcal{D}_{\text{train}}))} \leq e^{\varepsilon} \frac{\mathbb{P}(\mathbf{x}^* \notin \mathcal{D}_{\text{train}})}{\mathbb{P}(\mathbf{x}^* \in \mathcal{D}_{\text{train}})}.$$

Proposition 6 can be thought of as an extension of the Bayesian interpretation of DP explained by Jordon et al. (2022). Namely, the definition of DP immediately implies that, for any two adjacent sets D and D',

$$\frac{\mathbb{P}(\mathcal{D}_{\text{train}} = D \mid \mathcal{A}(\mathcal{D}_{\text{train}}))}{\mathbb{P}(\mathcal{D}_{\text{train}} = D' \mid \mathcal{A}(\mathcal{D}_{\text{train}}))} \leq e^{\varepsilon} \frac{\mathbb{P}(\mathcal{D}_{\text{train}} = D)}{\mathbb{P}(\mathcal{D}_{\text{train}} = D')}.$$

4 Guaranteeing RIP via Noise Addition

There are a number of mechanisms for guaranteeing DP which operate via simple noise addition (the Laplace mechanism) or sampling (the exponential mechanism) (Dwork et al., 2014). More recently, Abadi et al. (2016) showed how to make a small modification to the standard deep neural network training procedure to guarantee DP. In this section, we show that a small modification to standard training procedures can be used to guarantee RIP as well.

Suppose that $\mathcal{A}$ takes as input a data set D and produces output $\theta \in \mathbb{R}^d$. For instance, $\mathcal{A}$ may compute a simple statistical query on D, such as mean estimation, but our results apply equally well in the case that e.g. $\mathcal{A}(D)$ are the weights of a neural network trained on D. If θ are the weights of a generative model, then if we can guarantee RIP for θ, then by the data processing inequality (Theorem 3), this guarantees privacy for any output of the generative model.

The distribution over training data (in our case, the uniform distribution over size $n/2$ subsets of our complete dataset $\mathcal{D}$) induces a distribution over the output θ. The idea is the following: What is the smallest amount of noise we can add to θ which will guarantee RIP? If we add noise on the order of $\max_{D \sim D' \subseteq \mathcal{D}} \|\mathcal{A}(D) - \mathcal{A}(D')\|$, then we can adapt the standard proof for guaranteeing DP in terms of algorithm sensitivity to show that a restricted version of DP (only with respect subsets of $\mathcal{D}$) holds in this case, which in turn guarantees RIP. On the other hand, it seems possible that we should be able to reduce the amount of noise even further. Recall that by Propositions 2 and 6, RIP is only asking for a *marginal* guarantee on the change in the posterior probability of D given A, whereas DP is asking for a *conditional* guarantee on the posterior. So while $\max$ seems necessary for a conditional guarantee, the *moments* of θ should be sufficient for a marginal guarantee. Theorem 7 shows that this intuition is correct.

Theorem 7. *Let $\|\cdot\|$ be any norm, and let $\sigma^M \geq \mathbb{E}\|\theta - \mathbb{E}\theta\|^M$ be an upper bound on the M-th central moment of θ with respect to this norm over the randomness in $\mathcal{D}_{\text{train}}$ and $\mathcal{A}$. Let X be a random variable with density proportional to $\exp(-\frac{1}{c\sigma}\|X\|)$ with $c = (7.5/\eta)^{1+\frac{2}{M}}$. Finally, let $\hat{\theta} = \theta + X$. Then $\hat{\theta}$ is η-RIP, i.e., for any adversary $\mathcal{I}$,*

$$\mathbb{P}(\mathcal{I}(\mathbf{x}^*, \hat{\theta}) = y^*) \leq 1/2 + \eta.$$

Proof sketch. The proof proceeds by bounding the posterior likelihood ratio $\frac{\mathbb{P}(\mathbf{x}^* \notin \mathcal{D}_{\text{train}} \mid \hat{\theta})}{\mathbb{P}(\mathbf{x}^* \in \mathcal{D}_{\text{train}} \mid \hat{\theta})}$ from above and below for all $\hat{\theta}$ in a large $\|\cdot\|$-ball. This in turn yields an upper bound on the max in the integrand in Proposition 2 with high probability over $\mathcal{A}(\mathcal{D}_{\text{train}})$. The central moment σ allows us to apply a generalized Chebyshev inequality to establish these bounds. The full proof is computationally intensive and the complete details can be found in the Appendix. □

At first glance, Theorem 7 may appear to be adding noise of equal magnitude to all of the coordinates of θ, regardless of how much each contributes to the central moment σ. However, by carefully selecting the norm $\|\cdot\|$, we can add non-isotropic noise to θ such that the marginal noise level reflects the variability of each specific coordinate of θ. This is the content of Corollary 8.

Corollary 8. *Let $\sigma_i^2 \geq \mathbb{E}|\theta_i - \mathbb{E}\theta_i|^2$, and define $\|x\|_{\sigma,2} = \left(\sum_{i=1}^d \frac{|x_i|^2}{d\sigma_i^2}\right)^{1/2}$. Generate $Y_i \sim \mathcal{N}(0, \sigma_i^2)$, set $U = Y/\|Y\|_{\sigma,2}$, and draw $r \sim \text{Laplace}\left(\left(\frac{6.16}{\eta}\right)^2\right)$. Finally, set $X = rU$ and return $\hat{\theta} = \theta + X$. Then $\hat{\theta}$ is η-RIP.*

Proof sketch. Let $\|\cdot\| = \|\cdot\|_{\sigma,2}$. It can be shown that the density of X has the proper form. Furthermore, by definition of the σ_i, we have $\mathbb{E}\|\theta - \mathbb{E}\theta\|^2 \le 1$. The corollary follows directly from Theorem 7 with $M = 2$. The improvement in the numerical constant (from 7.5 to 6.16) comes from numerically optimizing some of the bounds in Theorem 7. □

Algorithm 1 RIP via noise addition

Require: Private dataset $\mathcal{D}$, σ estimation budget B, RIP parameter η
$\mathcal{D}_{\text{train}} \leftarrow \text{RANDOMSPLIT}(\mathcal{D}, 1/2)$

 # Estimate σ if an a priori bound is not known
$i \leftarrow 1$
for $i = 1, \dots, B$ **do**
$\mathcal{D}^{(i)}_{\text{train}} \leftarrow \text{RANDOMSPLIT}(\mathcal{D}_{\text{train}}, 1/2)$
$\theta^{(i)} \leftarrow \mathcal{A}(\mathcal{D}^{(i)}_{\text{train}})$
end for
$\bar{\theta} \leftarrow \frac{1}{B}\sum_{i=1}^{B} \theta^{(i)}$
$\sigma^2 \leftarrow \frac{1}{B-1}\sum_{i=1}^{B} \|\theta^{(i)} - \bar{\theta}\|^2$

 # Add appropriate noise to the base algorithm's output
$U \leftarrow \text{Unif}(\{u \in \mathbb{R}^d \,:\, \|u\| = 1\})$
$r \leftarrow \text{Laplace}\left(\left(\frac{7.5}{\eta}\right)^2 \sigma\right)$
$X \leftarrow rU$
return $\mathcal{A}(\mathcal{D}_{\text{train}}) + X$

When does RIP improve over DP? By Theorem 4, any DP algorithm gives rise to a RIP algorithm, so we *never* need to add more noise than the amount required to guarantee DP, in order to guarantee RIP. However, Theorem 7 shows that RIP affords an advantage over DP when the variance of our algorithm's output (over subsets of size $n/2$) is much smaller than its sensitivity Δ, which is defined as the maximum change in the algorithm's output when evaluated on two datasets which differ in only one element. For instance, applying the Laplace mechanism from DP requires a noise which scales like Δ/ϵ to guarantee ε-DP. It is easy to construct examples where the variance is much smaller than the sensitivity if the output of our "algorithm" is allowed to be completely arbitrary as a function of the input. However, it is more interesting to ask if there are any *natural* settings in which this occurs. Proposition 9 answers this question in the affirmative.

Proposition 9. *For any finite $D \subseteq \mathbb{R}$, define $\mathcal{A}(D) = \frac{1}{\sum_{x \in D} x}$. Given a dataset $\mathcal{D}$ of size n, define $\mathbb{D} = \{D \subseteq \mathcal{D} \,:\, |D| = \lfloor n/2 \rfloor\}$, and define*

$$\sigma^2 = \text{Var}(\mathcal{A}(D)), \qquad \Delta = \max_{D \sim D' \in \mathbb{D}} |\mathcal{A}(D) - \mathcal{A}(D')|.$$

Here the variance is taken over $D \sim \text{Unif}(\mathbb{D})$. Then for all n, there exists a dataset $|\mathcal{D}| = n$ such that $\sigma^2 = O(1)$ but $\Delta = \Omega(2^{n/3})$.

Proof sketch. Assume n is even for simplicity. Let $p = \binom{n}{n/2}^{-1}$ and $A = \sqrt{p} - \sum_{i=0}^{\frac{n}{2}-2} 2^i$. Take

$$\mathcal{D} = \{2^i \,:\, i = 0, \dots, n-2\} \cup \{A\}.$$

When $D = \{2^0, \dots, 2^{\frac{n}{2}-2}, A\}$, then $\mathcal{A}(D) = p^{-1/2}$, and this occurs with probability p. For all other subsets D', $0 \le \mathcal{A}(D') \le 1$. □

We remark that similar results should hold for e.g. subset precision matrix queries, perhaps even without such a carefully constructed $\mathcal{D}$ if the size of the subset is comparable to the dimension of the data.

5 SIMULATION RESULTS

To illustrate our theoretical results, we plot the noise level needed to guarantee RIP vs. the corresponding level of DP (with the correspondence given by Theorem 4) for the example in Proposition 9.

Refer to Fig. 1. Dotted lines refer to DP, while the solid line is for RIP. The x-axis gives the best possible bound on the attacker's improvement in accuracy over random guessing–i.e., the parameter η for an η-RIP method–according to that method's guarantees. For DP, the value along the x-axis is given by the (tight) correspondence in Theorem 4, namely $\eta = \frac{1}{1+e^{-\varepsilon}} - \frac{1}{2}$. $\eta = 0$ corresponds to perfect privacy (the attacker cannot do any better than random guessing), while $\eta = \frac{1}{2}$ corresponds to no privacy (the attacker can determine membership with perfect accuracy). The y-axis denotes the amount of noise that must be added to the non-private algorithm's output, as measured by the scale parameter of the Laplace noise that must be added. For RIP, by Theorem 7, this is $(6.16/\eta)^2\sigma$ where σ is an upper bound on the variance of the base algorithm over random subsets, and for DP this is $\frac{\Delta}{\log \frac{1+2\eta}{1+2\eta}}$. (This comes from solving $\eta = \frac{1}{1+e^{-\varepsilon}}$ for ε, then using the fact that $\text{Laplace}(\Delta/\varepsilon)$ noise must be added to guarantee ε-DP.) For DP, the amount of noise necessary changes with the size n of the private dataset. For RIP, the amount of noise does not change, so there is only one line.

The results show that for even small datasets ($n \geq 36$) and for $\eta \geq 0.01$, direct noise accounting for RIP gives a large advantage over guaranteeing RIP via DP. In practice, such small datasets are uncommon. As n increases above even this modest range, the advantage in terms of noise reduction for RIP vs. DP quickly becomes many orders of magnitude and is not visible on the plot. (Refer to Proposition 9. The noise required for DP grows exponentially in n, while it remains constant in n for RIP.)

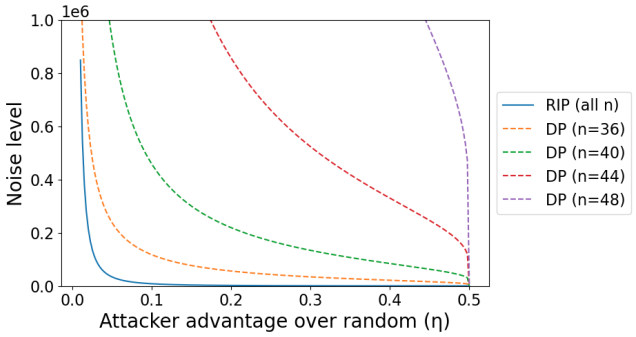

Figure 1: Noise level vs. privacy guarantee for RIP and DP. For datasets with at least $n = 36$ points and for almost all values of η, RIP allows us to add much less noise than what would be required by naively applying DP. For $n > 48$, the amount of noise required by DP is so large that it will not appear on the plot.

6 CONCLUSION

In this work, we propose a novel privacy property, re-identification privacy (RIP) and explained its properties and relationship with differential privacy (DP). The RIP property is more readily interpretable than the guarantees offered by (DP). RIP also requires a smaller amount of noise to guarantee as compared to DP, and therefore can retain greater utility in practice. We proposed a simple "wrapper" method for guaranteeing RIP, which can be implemented with a minor modification both to simple statistical queries or more complicated tasks such as the training procedure for parametric machine learning models.

Limitations As the example used to prove Theorem 5 shows, there are cases where apparently non-private algorithms can satisfy RIP. Thus, algorithms which satisfy RIP may require post-processing to ensure that the output is not one of the low-probability events in which data privacy is leaked. In addition, because RIP is determined with respect to a holdout set still drawn from $\mathcal{D}$, an adversary

may be able to determine with high probability whether or not a given sample was contained in $\mathcal{D}$, rather than just in $\mathcal{D}_{\text{train}}$, if $\mathcal{D}$ is sufficiently different from the rest of the population.

Future work Theorem 4 suggests that DP implies RIP in general. However, Theorem 7 shows that a finer-grained analysis of a standard DP mechanism (the Laplace mechanism) is possible, showing that we can guarantee RIP with less noise. It seems plausible that a similar analysis can be undertaken for other DP mechanisms. In addition to these "wrapper" type methods which can be applied on top of existing algorithms, bespoke algorithms for guaranteeing RIP in particular applications (such as synthetic data generation) are also of interest. Lastly, noise addition is a simple and effective way to enforce privacy, but other classes of mechanisms may also be possible. For instance, is it possible to directly regularize a probabilistic model using Proposition 2? Finally, the connections between RIP and other theoretical notions of privacy (Renyi DP (Mironov, 2017), concentrated DP (Dwork & Rothblum, 2016), etc.) are also of interest. Lastly, this paper focused on developing on the theoretical principles and guarantees of RIP, but systematic empirical evaluation is an important direction for future work. Practical membership inference attacks–particularly those against synthetic data and generative models rather than predictive models–still have a gap between practical efficacy and the theoretical upper bounds. It is likely that this gap can be closed through a combination of improved privacy accounting, but also through improved practical attacks. For the "shadow model" approach used by Stadler et al. (2022), improved computational efficiency is also of interest for improving membership inference attacks. These improved attacks will in turn allow model developers to better audit the empirical privacy limitations of their methods.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pp. 290–306. PMLR, 2022.

Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020.

Teodora Baluta, Shiqi Shen, S Hitarth, Shruti Tople, and Prateek Saxena. Membership inference attacks and generalization: A causal perspective. *ACM SIGSAC Conference on Computer and Communications Security*, 2022.

Mark Bun, Damien Desfontaines, Cynthia Dwork, Moni Naor, Kobbi Nissim, Aaron Roth, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Statistical inference is not a privacy violation. DifferentialPrivacy.org, 06 2021. `https://differentialprivacy.org/inference-is-not-a-privacy-violation/`.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2021a.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021b.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 343–362, 2020.

Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Anna C Gilbert and Audra McMillan. Property testing for differential privacy. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 249–258. IEEE, 2018.

Xinlei He, Zheng Li, Weilin Xu, Cory Cornelius, and Yang Zhang. Membership-doctor: Comprehensive assessment of membership inference against machine learning models. *arXiv preprint arXiv:2208.10445*, 2022.

Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. Mˆ 4i: Multi-modal models membership inference. *Advances in Neural Information Processing Systems*, 2022.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33: 22205–22216, 2020.

Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? *ACM SIGSAC Conference on Computer and Communications Security*, 2022.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory. *ACM SIGSAC Conference on Computer and Communications Security*, 2022.

Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

HHS Office for Civil Rights. Standards for privacy of individually identifiable health information. final rule. *Federal register*, 67(157):53181–53273, 2002.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Theresa Stadler, Bristen Oprisanu, and Carmela Troncoso. Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, 2022.

Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv:2112.09238*, 2021.

Anvith Thudi, Ilia Shumailov, Franziska Boenisch, and Nicolas Papernot. Bounding membership inference. *arXiv preprint arXiv:2202.12232*, 2022.

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