Differentially Private Learning Does Not Bound Membership Inference

Thomas Humphries[§]
University of Waterloo

Matthew Rafuse[§]
University of Waterloo

Lindsey Tulloch[§]
University of Waterloo

Simon Oya
University of Waterloo

Ian Goldberg
University of Waterloo

Urs Hengartner
University of Waterloo

Florian Kerschbaum
University of Waterloo

Abstract

Training machine learning models on privacy-sensitive data has become a popular practice, driving innovation in ever-expanding fields. This has opened the door to a series of new attacks, such as Membership Inference Attacks (MIAs), that exploit vulnerabilities in ML models in order to expose the privacy of individual training samples. A growing body of literature holds up Differential Privacy (DP) as an effective defense against such attacks, and companies like Google and Amazon include this privacy notion in their machine-learning-as-a-service products. However, little scrutiny has been given to how underlying correlations or bias within the datasets used for training these models can impact the privacy guarantees provided by DP. In this work, we challenge prior findings that suggest DP provides a strong defense against MIAs.

We provide theoretical and experimental evidence for cases where the theoretical bounds of DP are violated by MIAs using the same attacks described in prior work. We first show this empirically, with real-world datasets carefully split to create a distinction between member and non-member samples, and then we study the reason why the theoretical DP bounds break when members and non-members are not independent and identically distributed. Our findings suggest that certain properties of datasets, such as bias or data correlation, play a critical role in determining the effectiveness of DP as a privacy preserving mechanism against MIAs.

1 Introduction

Machine learning (ML) is increasingly used to make predictions on privacy-sensitive data. Recent years have seen large tech companies such as Google and Amazon begin to offer machine learning as a service to the general public through their cloud platforms. Though easier access to these systems can yield new and interesting insights in a variety of fields, machine learning models trained on sensitive data also present a lucrative attack surface for adversaries. In a Membership Inference Attack (MIA) [27, 30], an adversary attempts to correctly identify data that was used to train an ML model (i.e., its members). Successful MIAs violate the privacy of correctly classified samples and pose a significant threat to unprotected machine learning models [27].

Differential Privacy (DP) can be applied to an ML model during the learning process [1, 6, 8, 27, 30, 31] to limit the effect that a single data point can have on the model's output. The TensorFlow[1] open source library, for example, includes a privacy component that data scientists can freely use to apply differential privacy to their machine learning models. Utilizing DP in this way has been touted as a defense against MIAs. Intuitively, reducing the impact that any given data sample can have on the model should make inferring the presence of a particular sample more difficult—thus mitigating MIAs.

Support for this intuition has been demonstrated in previous work where security researchers have provided experimental and theoretical evidence of DP as an effective defense against MIAs [22]. Yeom et al. [30] investigate the impact of overfitting and its influence on an adversary's ability to both successfully identify members of the training set and infer attributes about them. Most importantly, they define a *membership experiment* and provide a theoretical upper bound on the success an adversary can achieve in an MIA (the so called *membership advantage*) given that the model was trained using differential privacy as a defense.

Jayaraman and Evans [18] compare the performance of different MIAs under various DP mechanisms. They show that the upper bound on the advantage provided by Yeom et al. is very loose in practice, meaning that the adversary has little success in correctly identifying membership in comparison to what is theoretically possible, even as the privacy protection offered by DP is weakened (corresponding to large values of the DP parameter ε). This suggests that DP protects members of the training set against MIAs even when the noise applied should be inconsequential. Critical of this analysis, Erlingsson et al. [11] dispute the equating of privacy parameters between

[§]Equal contribution

[1]https://github.com/tensorflow/tensorflow

differing DP mechanisms without full consideration of the context. They provide a tighter bound and argue that the theoretical bound is meaningful in practice. In follow up work, Jayaraman et al. [19] show that Yeom et al.'s bound is even looser when the a priori distribution of members and non-members is unbalanced. They note that this is a more realistic scenario than running an MIA on balanced sets. Murakonda and Shokri [25] observe that a model may be much less vulnerable to actual MIAs than these theoretical bounds suggest. They propose a tool that empirically quantifies the vulnerability level of an ML model using state-of-the-art MIAs. This tool, ML Privacy Meter, aims to optimize the privacy/utility trade off for a given model, allowing for the release of models with higher utility (at the expense of only enjoying empirical guarantees). In summary, this body of literature suggests that applying differential privacy to machine learning models, even in severely weakened privacy regimes, provides strong protection against known MIAs.

In recent work, Tschantz et al. [28] summarize and unify a large body of work that studies the impact of correlated datasets on differential privacy protection techniques. They conclude that DP makes no claim of protection against such correlations being learned and used to reveal information about the data points. However, despite the effects of correlated data being well studied in the DP community [28], the above body of literature studying MIAs only considers the specific case where members and non-members are independent and identically distributed (IID) according to an overall population distribution [11, 18, 19, 25]. Focusing exclusively on the performance of the IID case seems to overlook the difficulty of achieving truly IID data, thereby suggesting an elevated expectation of privacy that may be misleading.

In practice, ML often relies on data that is biased toward a particular subgroup, as this is the only data the practitioner has access to. Buolamwini and Gebru [7] demonstrate with real-world commercial examples (Watson Visual Recognition, Face++, MSFT) that bias is not uncommon in widely deployed and used ML models and has detrimental effects on accuracy for underrepresented subpopulations. Bagdasaryan et al. [5] show that this accuracy loss for underrepresented subgroups increases even further when differential privacy is applied to the training process. Under some circumstances, an unconsidered bias may impact the entire dataset, creating a distinct in group and out group. As an example, consider a model that has been trained on a particular hospital's inpatient triage data in order to assess the urgency of care required by incoming patients. Such a model may be fraught with underlying external correlations related to the geographic location of the hospital and social constructs that determine the makeup of its patient base. In such a scenario where members of the training set are distinct from non-members due to some underlying correlation, the work of Tschantz et al. [28] suggests that DP may be insufficient to protect against an MIA, even under very strong privacy regimes. In this work we aim to investigate this idea and challenge the belief that DP is always a successful defense against MIAs.

First, we consider the case of independent and identically distributed data samples. We analyze Yeom et al.'s membership experiment, and derive a new bound on the membership advantage that improves over previous ones [11, 30]. We observe that theoretical proofs of these privacy bounds make a crucial assumption that members and non-members are independent and identically distributed, which we call the *IID assumption*.

We then begin with an empirical investigation of the case where the IID assumption does not hold. Using off-the-shelf attacks on DP-protected models and an evaluation setup consistent with prior work [18], we evaluate the membership advantage when the dataset has a distinction between members and non-members. To do this we use two public datasets and split them such that the data used to train the model has a clear bias compared with data from the wider population. We show, on two different splits, that the adversary is able to identify members and non-members with much higher accuracy than in previous work, breaking the aforementioned privacy bounds. We argue that this scenario, where members and non-members have different distributions, could happen in practice. However, it is perhaps unfair from a theoretical standpoint, since an adversary could distinguish members and non-members without even having access to the model.

Finally, we perform a theoretical study of membership inference when the IID assumption does not hold. Motivated by our empirical evaluations, we introduce a *fairness requirement* that implies that the adversary does not have any advantage when deciding whether a sample is a member or a non-member if they have not observed the released ML model. We propose a variation of Yeom et al.'s membership experiment that accounts for data dependency while meeting this fairness requirement, and explain why current bounds do not hold in this case. To further show that it is not possible to get meaningful upper bounds in this general case, we design a synthetic population distribution and empirically show that off-the-shelf MIAs achieve significant membership inference performance (e.g., membership advantage $\mathtt{Adv} > 0.8$ for differential privacy parameter $\varepsilon = 0.5$). This performance serves as a lower bound on the worst-case membership inference scenario and proves that one cannot generally bound membership inference performance using differential privacy when the training samples are not independent.

To summarize, our contributions are the following:

- We study the membership experiment by Yeom et al. [30] and the bounds on the membership advantage that differential privacy offers. We *derive a new bound* that improves over previous results [11, 30].

- We experimentally evaluate the performance of existing membership inference attacks against DP-protected models on real datasets where we build the training set

Table 1: Notation

Notation	Description
$z = (x, y)$	Sample with feature vector x and label y
$\mathcal{Z}$	Space of all possible samples
S	Training set of size n; $S \in \mathcal{Z}^n$
$\mathcal{D}$	Distribution of data points (over $\mathcal{Z}$)
$\mathcal{A}$	Space of all possible machine learning models
A	Learning algorithm $A : \mathcal{Z}^n \to \mathcal{A}$
a	Instance of a trained model ($a \in \mathcal{A}$)
$z \sim \mathcal{D}$	z is sampled from $\mathcal{D}$
$S \sim \mathcal{D}^n$	Each $z \in S$ is sampled independently from $\mathcal{D}$
$z \sim S$	z is chosen uniformly at random from S
$i \sim [n]$	i is sampled uniformly at random from $\{1, 2, \ldots, n\}$
$\mathtt{Att}$	Membership inference attack, outputs a bit
$\mathtt{Adv}$	Membership advantage, $\mathtt{Adv} \in [0, 1]$

with certain data dependencies. Our evaluations reveal that previous bounds on the membership advantage fail when data is not independent, and in some cases the adversary's accuracy reaches dangerously high values.

- We theoretically study why the bounds do not hold with data dependence and propose a new membership experiment that accounts for these dependencies.
- Following our new membership experiment, we create a synthetic population distribution to empirically lower bound the membership advantage when the data is dependent.

In summary, we show that we are able to achieve high MIA accuracy on real-world and artificial datasets even when DP is applied to ML models. We discuss why this violation has privacy implications in practice, and conclude that DP is not an all-purpose solution and can fail to provide protection against MIA even under strong privacy regimes.

2 Preliminaries

In this section, we summarize the concepts related to membership inference attacks and differentially private machine learning that are most relevant to our work. For reference, our notation is summarized in Table 1.

We use $z = (x, y)$ to denote an element or data sample, where x is its feature vector and y is its class or label. Let $\mathcal{Z}$ be the element space, i.e., $z \in \mathcal{Z}$. We use $S \in \mathcal{Z}^n$ to denote a training set that contains n elements $z \in \mathcal{Z}$. A training algorithm A is a (possibly randomized) function that takes a training set $S \in \mathcal{Z}^n$ and outputs a trained model $a \in \mathcal{A}$, where $\mathcal{A}$ is the space of trained models. We use $a = A(S)$ to denote that a is the model that results from applying the training algorithm A to the training set S. The goal of the trained model a is to solve a classification task; i.e., assign a label y to a feature vector x. In this work, we focus on neural networks, which are a popular model choice for solving classification problems in machine learning. However, our theoretical findings are generic and apply to other models as well.

Algorithm 1: $\mathtt{Exp}(\mathtt{Att}, A, n, \mathcal{D})$ [30]

Sample $S \sim \mathcal{D}^n$, train $a = A(S)$;
Choose $b \sim \{0, 1\}$ uniformly at random;
Draw $z \sim S$ if $b = 0$, or $z \sim \mathcal{D}$ if $b = 1$;
$\mathtt{Exp}(\mathtt{Att}, A, n, \mathcal{D}) = 1$ if $\mathtt{Att}(z, a, n, A, \mathcal{D}) = b$; else 0.

2.1 Membership Inference Attacks

Though useful for solving classification tasks, machine learning models are subject to various privacy attacks. One such attack, and the focus of this work, is the Membership Inference Attack (MIA). The goal of an MIA is to determine whether or not a specific data point z was included in the training set of a target model a. This attack is particularly dangerous when the model is trained on sensitive data, where an individual's inclusion or exclusion in the dataset could reveal sensitive or compromising information to an attacker.

Yeom et al. [30] provide a formal definition of the membership inference problem, which we formalize in Algorithm 1. In this experiment, denoted as $\mathtt{Exp}$, the training set S consists of n elements sampled independently from a distribution $\mathcal{D}$ over the space of elements $\mathcal{Z}$, and a is the model trained from S. We choose a bit $b \in \{0, 1\}$ uniformly at random. If $b = 0$, we draw a random element z from S (i.e., a *member* of the training set). Otherwise, $b = 1$, meaning we draw a random element z from the distribution $\mathcal{D}$ (i.e., a *non-member*). The adversary receives the element z and the trained model a (depending on the assumptions, this could be white-box or black-box access to the model), and knows the training set size n, the training algorithm A, and the element distribution $\mathcal{D}$.[2] With this information, the adversary carries out an attack $\mathtt{Att}(z, a, n, A, \mathcal{D})$ that outputs a bit b, indicating the membership. The adversary succeeds (denoted $\mathtt{Exp}(\mathtt{Att}, A, n, \mathcal{D}) = 1$) if the attack correctly infers the membership or non-membership of z.

Yeom et al. define the *membership advantage* as

$$\mathtt{Adv}(\mathtt{Att}, A, n, \mathcal{D}) \doteq 2 \cdot \Pr(\mathtt{Exp}(\mathtt{Att}, A, n, \mathcal{D}) = 1) - 1, \quad (1)$$

where the probabilities are taken over the coin flips of $\mathtt{Att}$, the random choices of S, b, and z, and the (possibly randomized) model training $a = A(S)$. Note that the membership

[2] Yeom et al. [30] do not explicitly consider the training algorithm A as an input to the attack $\mathtt{Att}$, perhaps because the particular attack they develop does not use this information. However, it is reasonable to assume that the adversary knows this information and could use it to try to gather information of the training data S from the model $a = A(S)$.

advantage is just a linear mapping of the adversary's success probability, which goes from 0.5 (random guess) to 1 (perfect attack success), onto the interval $[0,1]$.

The membership advantage can equivalently be defined as the difference between true and false positive rates:

$$\texttt{Adv} = \Pr(\texttt{Att}=0|b=0) - \Pr(\texttt{Att}=0|b=1) \quad (2)$$
$$= \Pr(\texttt{Att}=1|b=1) - \Pr(\texttt{Att}=1|b=0)\,. \quad (3)$$

Yeom et al. [30] point out that, in this experiment, the non-members are drawn from $\mathcal{D}$ and therefore there is a non-zero probability that $z \in S$ even though $b=1$. Other works [22,27] specify that non-members must be distinct from members, i.e., $z \sim \mathcal{D} \setminus S$ when $b=1$. We use Yeom et al.'s formulation in our work, since the probability that a non-member is identical to a member is typically negligible (e.g., when the probability of sampling any particular element z according to $\mathcal{D}$ is negligible, which might happen when the space of all possible elements $\mathcal{Z}$ is large). We also note that the membership advantage, as defined in (1), uses $\texttt{Exp}$ which assumes a uniform membership prior probability (i.e., $\Pr(b=0)=\Pr(b=1)=0.5$). Jayaraman et al. [19] consider alternative privacy metrics for the unbalanced case $\Pr(b=0) \neq \Pr(b=1)$, since there are typically significantly more non-members than members available to the adversary ($\Pr(b=0) \ll \Pr(b=1)$). We only consider the balanced case in this work, since this is what the membership advantage metric inherently assumes (e.g., Algorithm 1); however, our findings can be extended to the unbalanced case.

Due to the complexity of the relation between the training set S and the trained model a, optimal membership inference attacks $\texttt{Att}$, i.e., those that maximize (1), do not currently exist. Two of the most well-known membership inference attacks that have been evaluated in the literature are the proposals by Shokri et al. [27] and Yeom et al. [30].

The shadow model attack by Shokri et al. [27] assumes black-box access to the target model, in the sense that we allow the adversary to query the model and obtain confidence values on the classification. The adversary is assumed to be in possession of a large collection of public data that is similarly distributed to the target and either already labeled with the same classes as the target or else labeled through use of the target model. The adversary uses this data to train a set of shadow models designed to mimic the target model's functionality. The attack itself is then carried out using additional models, (one for each class) called the attack models, trained to identify the membership status of a sample using outputs from the shadow models. Since the models were trained on similarly distributed data, the intuition is that the shadow models should react similarly to the target model on points that they were trained on due to underlying structural similarities.

The logloss attack by Yeom et al. [30] assumes that the adversary has access to the loss function of the target model as well as the distribution of the loss on the private training data. The assumption that the adversary knows the entire distribution can be relaxed by assuming that the model's generalization error (i.e., the classification error of a random sample $z \sim \mathcal{D}$) is normally distributed. Given $z=(x,y)$ and a, the attack operates by simply querying the model a to obtain the loss of z. If this loss is less than the expected loss for the training set, then the attack labels it as a member (i.e., it outputs the bit $\texttt{Att}=0$), and otherwise decides that z is a non-member ($\texttt{Att}=1$).

2.2 Differential Privacy in Machine Learning

Differential Privacy (DP), a privacy notion introduced by Dwork et al. [9], has become the gold standard in database privacy. A key component of this notion posits that the presence of an individual in a dataset should have little impact on the output of a differentially private algorithm. DP has been applied to machine learning and it is intuitively useful to provide protection against membership inference attacks.

Definition 2.1 ((ε,δ)-DP)**.** A training algorithm A provides (ε,δ)-DP iff, for any two *adjacent* datasets S and S' (i.e., differing by a single entry), and all possible subsets of the space of trained models $\mathcal{R} \subseteq \mathcal{A}$,

$$\Pr(A(S) \in \mathcal{R}) \leq \Pr(A(S') \in \mathcal{R}) \cdot e^{\varepsilon} + \delta\,. \quad (4)$$

The parameter ε captures the degree of leakage of the mechanism A [9]. Small values of ε indicate that a model trained with S is indistinguishable from a model trained with S' which, intuitively, makes it difficult to infer whether or not an element z is in the training set. The parameter δ makes it easier to satisfy the DP constraint by allowing a small chance of failure in the privacy guarantee. It is typical to choose $\delta < 1/n$, where n is the number of elements in the dataset [10].

There are different approaches to developing a differentially private training algorithm A. One can add noise to the training set directly before training, to each step of the training algorithm, or to the model parameters directly [2,8]. However, doing so incurs a utility loss, in the sense that the classification accuracy of the model will decrease. In the case of neural networks, the differentially private stochastic gradient descent technique by Abadi et al. [1] is widely used. This technique involves clipping the gradients used for updating the network's weights during training time, and adding Gaussian noise to the average of different gradients.

Due to the complexity and iterative nature of training algorithms in neural networks, it is hard to keep track of the differential privacy budget (captured by the parameter ε) these algorithms provide. The so-called naïve composition technique [10] is the simplest technique to compute the overall ε of a training mechanism, but it yields values of ε that are unrealistically large. More refined techniques to account for the differential privacy budget are the moments accounting scheme of Abadi et al. [1] and Rényi Differential Privacy

(RDP) [24]. Using an RDP-based privacy accountant, one can prove that the leakage level ε is considerably lower than reported by naïve composition techniques. In our evaluation we use RDP to account for the privacy level of the model, as it allows for tighter privacy analysis than other accounting techniques [18].

2.3 Theoretical Guarantees of Differential Privacy against Membership Inference

Differential privacy provides worst-case privacy guarantees, in the sense that $\Pr(A(S) \in \mathcal{R})$ and $\Pr(A(S') \in \mathcal{R})$ have to be close for any two adjacent training sets S and S' and any set of output models $\mathcal{R}$. This makes it amenable to various theoretical analyses. Yeom et al. provide an upper bound on the membership advantage when the learning algorithm is $(\varepsilon, 0)$-DP:

Theorem 2.1 (Yeom et al.'s bound [30])**.** *Let A be an $(\varepsilon, 0)$-DP learning algorithm. Then, for all attacks* `Att`*, training set sizes n, and data point distributions $\mathcal{D}$, the membership advantage in* `Exp` *satisfies*

$$\mathtt{Adv}(\mathtt{Att}, A, n, \mathcal{D}) \leq e^{\varepsilon} - 1\,. \tag{5}$$

Note that the advantage is also upper bounded by 1, so this bound is loose for $\varepsilon > \ln 2$. More recently, Erlingsson et al. [11] provide a tighter bound using results from Hall et al. [15] for the more generic notion of (ε, δ)-DP:

Theorem 2.2 (Erlingsson et al.'s bound [11])**.** *Let A be an (ε,δ)-DP learning algorithm. Then, for all attacks* `Att`*, training set sizes n, and data point distributions $\mathcal{D}$, the membership advantage in* `Exp` *satisfies*

$$\mathtt{Adv}(\mathtt{Att}, A, n, \mathcal{D}) \leq 1 - e^{-\varepsilon}(1 - \delta)\,. \tag{6}$$

We can see that this bound is less than or equal to 1. Also, if we particularize this bound for the strict differential privacy notion ($\delta = 0$) we see that it is tighter than (5).

3 Tighter Bounds on Membership Inference under the IID Assumption

In this section, we derive a new bound on the membership advantage that improves over previous bounds by Yeom et al. (5) and Erlingsson et al. (6). These bounds hold when the advantage is measured following Yeom et al.'s membership experiment [30] (`Exp` in Algorithm 1), which is an accepted approach to measure the success of a membership inference attack. Recall that this membership experiment assumes that every element in the training set S and the possible non-member z are *independent and identically distributed*, following $\mathcal{D}$. We call this the *IID assumption*.

In order to derive our bound, we first consider an alternative formulation of `Exp`, that we denote by $\mathtt{Exp}'$, and define it in

Algorithm 2: $\mathtt{Exp}'(\mathtt{Att}, A, n, \mathcal{D})$

Sample $\tilde{S} \sim \mathcal{D}^{n-1}$, $z \sim \mathcal{D}$, $z' \sim \mathcal{D}$;
Choose $b \sim \{0,1\}$ uniformly at random;
Build $S = \tilde{S} \cup \{z\}$ if $b = 0$, or $S = \tilde{S} \cup \{z'\}$ if $b = 1$;
Train $a = A(S)$;
$\mathtt{Exp}'(\mathtt{Att}, A, n, \mathcal{D}) = 1$ if $\mathtt{Att}(z, a, n, A, \mathcal{D}) = b$; else 0.

Algorithm 2. In this experiment, we first sample $n-1$ points independently from the data distribution $\mathcal{D}$, as well as two additional points z and z'. Then, we flip a coin b to choose which of $\{z, z'\}$ we include in the training data S, and train the model with S following algorithm A. The adversary receives the trained model a as well as the point z and input parameters n, A, and $\mathcal{D}$, and has to decide whether or not $z \in S$.

Lemma 3.1. *For any attack* `Att`*, training algorithm A, positive integer n, and distribution over data samples $\mathcal{D}$, experiments* `Exp` *and* $\mathtt{Exp}'$ *are statistically equivalent; i.e.,*

$$\mathtt{Adv}(\mathtt{Att}, A, n, \mathcal{D}) = \mathtt{Adv}'(\mathtt{Att}, A, n, \mathcal{D})\,. \tag{7}$$

Proof. It is enough to prove that the joint distribution of (b, a, z) given A, n, and $\mathcal{D}$ is identical in `Exp` and $\mathtt{Exp}'$ since, in that case, the probability that a particular attack `Att` guesses the bit b, $\Pr(\mathtt{Att}(z, a, n, A, \mathcal{D}) = b)$, would be the same in both experiments.

In both experiments, b is chosen uniformly from $\{0,1\}$. When $b = 0$, in `Exp` the adversary receives a model a trained from a training set sampled from $\mathcal{D}^n$, and receives a random sample from this training set $z \sim S$. This is also true for $\mathtt{Exp}'$ since, when $b = 0$, we have $\tilde{S} \sim \mathcal{D}^{n-1}$, $z \sim \mathcal{D}$, and $S = \tilde{S} \cup \{z\}$ (the order of elements in S is irrelevant; e.g., we assume they are shuffled before training). When $b = 1$, in both experiments the training set S is composed of IID samples of $\mathcal{D}$ and the non-member z is independent from S and also sampled from $\mathcal{D}$. □

We now state and prove our bound.

Theorem 3.1 (Tighter Bound)**.** *Let A be an (ε,δ)-DP learning algorithm. Then, for all attacks* `Att`*, training set sizes n, and data point distributions $\mathcal{D}$, the membership advantage in* `Exp` *satisfies*

$$\mathtt{Adv}(\mathtt{Att}, A, n, \mathcal{D}) \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}\,. \tag{8}$$

Proof. We prove the bound using $\mathtt{Exp}'$, since $\mathtt{Adv}' = \mathtt{Adv}$, as stated in Lemma 3.1. In this experiment, consider a *more informed* adversary that, besides a and z, also knows the values of $\tilde{S}$ and z'. We denote an attack that this adversary can carry out by $\mathtt{Att}^*(z, z', \tilde{S}, a, n, A, \mathcal{D})$ and use $\mathtt{Adv}^*$ to denote the advantage of $\mathtt{Att}^*$. We can characterize any attack $\mathtt{Att}(z, a, n, A, \mathcal{D})$ in $\mathtt{Exp}'$ as an attack $\mathtt{Att}^*$ (the attack can just

ignore the variables z' and $\tilde{S}$). In other words, for every attack $\texttt{Att}$ there exists another attack $\texttt{Att}^*$ such that $\texttt{Adv} \leq \texttt{Adv}^*$. Therefore, an upper bound for $\texttt{Adv}^*$ (the advantage of the informed adversary) is also an upper bound for $\texttt{Adv}$ (the advantage of the standard adversary in $\texttt{Exp}$ and $\texttt{Exp}'$).

We consider a deterministic attack; i.e., an attack $\texttt{Att}^*$ that always performs the same decision given a particular observation. One can prove that adversaries that randomize their decision cannot achieve a greater advantage than deterministic adversaries [20]. Consider a particular realization of $\tilde{S}$, z, and z'. For this realization, let $\mathcal{R}$ be the region for which, if $a \in \mathcal{R}$, then the attack outputs $\texttt{Att}^*(z,z',\tilde{S},a,n,A,\mathcal{D}) = 1$. Then, using the membership advantage formulation in (3),

$$\begin{aligned}\texttt{Adv}^* =& \Pr(\texttt{Att}^* = 1 | b = 1, \tilde{S}, z, z') \\ &- \Pr(\texttt{Att}^* = 1 | b = 0, \tilde{S}, z, z') \\ =& \Pr(A(S) \in \mathcal{R} | S = \tilde{S} \cup \{z'\}) \\ &- \Pr(A(S) \in \mathcal{R} | S = \tilde{S} \cup \{z\}) .\end{aligned} \tag{9}$$

We define $S_0 = \tilde{S} \cup \{z\}$ and $S_1 = \tilde{S} \cup \{z'\}$. Then, the equation above is simply

$$\texttt{Adv}^* = \Pr(A(S_1) \in \mathcal{R}) - \Pr(A(S_0) \in \mathcal{R}) \tag{10}$$

$$= 1 - [\Pr(A(S_1) \notin \mathcal{R}) + \Pr(A(S_0) \in \mathcal{R})] . \tag{11}$$

Since S_0 and S_1 are adjacent datasets and $A(S)$ is the output of an (ε,δ)-DP mechanism with input S, we have that

$$\Pr(A(S_1) \in \mathcal{R}) \leq \Pr(A(S_0) \in \mathcal{R}) \cdot e^{\varepsilon} + \delta , \tag{12}$$

$$\Pr(A(S_0) \notin \mathcal{R}) \leq \Pr(A(S_1) \notin \mathcal{R}) \cdot e^{\varepsilon} + \delta . \tag{13}$$

We can rewrite these equations as:

$$1 - \delta \leq \Pr(A(S_0) \in \mathcal{R}) \cdot e^{\varepsilon} + \Pr(A(S_1) \notin \mathcal{R}) , \tag{14}$$

$$1 - \delta \leq \Pr(A(S_1) \notin \mathcal{R}) \cdot e^{\varepsilon} + \Pr(A(S_0) \in \mathcal{R}) , \tag{15}$$

and adding them yields

$$[\Pr(A(S_1) \notin \mathcal{R}) + \Pr(A(S_0) \in \mathcal{R})](e^{\varepsilon} + 1) \geq 2(1 - \delta) . \tag{16}$$

Using this in (11) results in:

$$\texttt{Adv}^* \leq 1 - \frac{2(1-\delta)}{e^{\varepsilon}+1} = \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon}+1} . \tag{17}$$

Since an upper bound on the advantage of the more informed adversary $\texttt{Adv}^*$ is also an upper bound for the advantage of the standard adversary $\texttt{Adv}$, this concludes the proof. □

This bound is tighter than (less than or equal to) (5) and (6) for all $\varepsilon \geq 0$ and $\delta \leq 1$.

Figure 1 shows a comparison between the existing membership advantage upper bounds (Yeom et al. (5) and Erlingsson et al. (6)) and our bound (8). We used a value of $\delta = 10^{-5}$ for (6) and (8), since this is the value of δ we use in our experiments later. Our bound is clearly the tightest, improving Erlingsson et al.'s by almost 0.2 in advantage for relevant privacy values $0.1 \leq \varepsilon \leq 2$. These bounds suggest that high privacy settings $0.01 \leq \varepsilon \leq 0.1$ completely thwart membership inference attacks, values of $\varepsilon \approx 1$ achieve intermediate privacy levels, and large values of $\varepsilon \geq 10$ provide no worst-case privacy guarantee on the membership advantage.

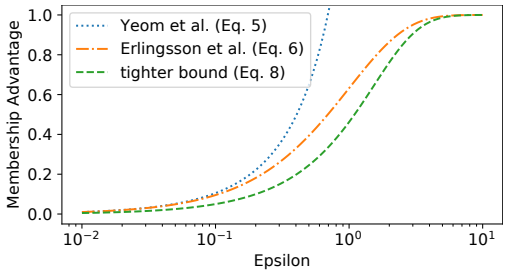

Figure 1: Membership advantage upper bounds

4 Evaluation of Membership Inference Attacks with Biased Datasets

In the previous sections, we have seen that differentially private training algorithms provide upper bounds on the adversary's membership advantage. A crucial assumption for these bounds to hold is that training set samples (i.e., members), as well as non-members, are independent and identically distributed. Thus, a natural question to ask is: do these bounds still hold when the IID assumption does not hold? Furthermore, if the bounds do not hold, how much does the membership advantage increase in this case?

In this section, we investigate these questions empirically. We consider the case where members and non-members are independent, but they come from *different distributions* (instead of being samples from the same distribution; e.g., $\mathcal{D}$ in Algorithm 1). In order to achieve this empirically, we use real datasets and split them into two statistically different subsets that we use as member and non-member sets. We test different splits, and evaluate the performance of existing MIAs [27, 30] in these scenarios. Our results reveal that not only do the bounds for the IID case not hold, but the membership advantage can reach dangerously high values even for high privacy settings (e.g., $\texttt{Adv} \approx 0.7$ for $\varepsilon = 0.1$). We first explain our evaluation setup, and then delve into each of the splits that we evaluate.

4.1 Evaluation Setup

We construct a series of evaluations to simulate the scenario where the data owner has only trained their model on a specific subset of the population. That is, the data is *biased* in

Table 2: Evaluation Setup Summary

Evaluation (Section)	Dataset and split	Owner set		Non-owner set	Attacks	
		train (members)	test	(non-members)	shadow model attack [27]	logloss attack [30]
4.2	`adult` education	10 000	2 500	10 000	✓	✓
4.3	`adult` cluster	10 000	2 500	10 000	✓	✓
	`compas` cluster	2059	515	2059	-	✓

some way, whether having an interpretable bias such as excluding a certain protected group of people or a more subtle *geometric* (spatial) bias in the feature space. To do this we create two sets: an *owner set*, containing the data used by the data owner to train and test their model, and a *non-owner set*, containing instances the data owner excluded. We split off 80% of the owner's dataset and use it to train the model (n elements), and use the remaining 20% as a testing set ($n/4$ elements). This testing set is used only to evaluate the model's classification accuracy. Having trained and tested our model using only the owner set, we then evaluate the success of the membership inference attack by running it for every element in the training set (from the owner set) as well as every element from the non-owner set (n elements in each set). We compute the membership advantage from the *average* number of times the adversary correctly identified the membership status of an element. This mimics the procedure in the theoretical experiments we defined in the previous section, where the choice of member/non-member (the bit b) is unbiased. However, in practice an adversary may have a more or less extensive non-owner set in their possession as pointed out by Jayaraman et al. in recent work [19]. We consider the case of a balanced prior probability as this is implicitly assumed by the membership advantage metric [30], but our findings also apply to the unbalanced scenario [19]. Table 2 summarizes our evaluation setup.

Datasets. We use two publicly available datasets considered in prior work [27, 29]:

1. The *Adult Dataset* (`adult`) [21] is an extraction from the 1994 US census database available on the UCI Machine Learning Repository.[3] The dataset contains 48 842 data samples, where the feature vector x contains 14 attributes such as age, level of education, and race, and the label y is a binary attribute indicating if the individual earns more or less than $50k a year. We replaced all missing values with the mean or mode value for that attribute and used a one-hot encoding for all categorical attributes.

2. The *Compas dataset* (`compas`) [3] is a dataset extracted from ProPublica's investigation into racial bias in Machine Learning. The data contains information about the recidivism of criminal defendants in the state of Florida and the COMPAS recidivism risk scores. Specifically, we use the two-year COMPAS scores data file and follow the same preprocessing carried out by ProPublica in their Jupyter Notebook.[4] This preprocessing removes rows with missing information and narrows down the samples using factors such as committing crimes eligible for jail time. Additionally, similarly to Yaghini et al. [29] we remove the risk score and apply a one-hot encoding on categorical attributes. This leaves 6 172 samples, where the feature vector x contains 15 attributes such as age, sex, and number of prior offenses, and the label y is a binary attribute indicating if the individual re-offended within 2 years.

Model Architecture. Following the work of Jayaraman et al. [18], and several other works in the field [1, 27], we use a ReLU network with 2 hidden layers, each with 256 neurons trained for 100 epochs as the target model in all evaluations. The models use the DP ADAM Gaussian Optimizer from TensorFlow Privacy[5] with ℓ_2 regularization to avoid overfitting. The ℓ_2 regularization coefficient, learning rate, and batch size hyper-parameters were optimized using a 5-fold cross validated grid search on the non-private model. However, we found that the default parameters used by Jayaraman et al. gave better accuracy under differential privacy in all evaluations. Therefore we used 10^{-5} as the ℓ_2 regularization coefficient, 10^{-2} as the learning rate, and 200 as the batch size, following the previous work [18]. For all evaluations, we vary ε over the values considered in prior work [18] and fix δ to equal 10^{-5}. This follows the general recommendation that δ should be less than the inverse of the dataset size [10].

Membership Inference Attacks. We consider two membership inference attacks evaluated in prior work [18] that we summarized in Section 2.1. The *shadow model attack* [27] requires a large pool of publicly available data in order to train shadow models that mimic the target model. Therefore we only run this attack on the `adult` dataset and omit the `compas` dataset. Following prior work [18, 27], we train five shadow models, each with the same architecture as the target model. The attack model architecture also follows prior work using the same architecture as the attack model, with 64 neurons in the hidden layer. The *logloss attack* [30] does

[3] http://archive.ics.uci.edu/ml

[4] https://github.com/propublica/compas-analysis

[5] https://github.com/tensorflow/privacy

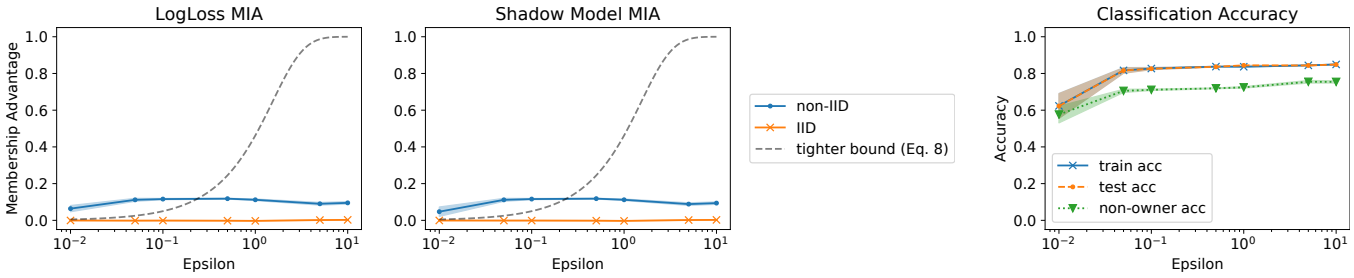

Figure 2: `adult` dataset with attribute-based (education) split

not require additional data, and instead assumes access to the average training loss, so we use it as the main attack in all of our evaluations.

Evaluation Metrics. We evaluate the performance of these attacks using the membership advantage metric from prior work [18, 30] as defined in (1). For each evaluation we plot the average advantage against our bound from Theorem 3.1 (tighter bound).

Implementation. We used Python 3.7 for our implementation, building upon the source code made available by Jayaraman et al. [18].[6] We utilize Jayaraman et al.'s implementation of both the logloss attack and shadow model attack, where the latter is based on the original source code from Shokri et al. [27]. The source code leverages TensorFlow[7] version 1.13.1 and its corresponding privacy library, TensorFlow Privacy[5] version 0.2.0. To implement RDP, we use the RDP accountant in TensorFlow Privacy.

4.2 Attribute-based Split

When all samples in the training set come from a particular subpopulation that is not representative of the global population, dependencies or bias between training samples could reasonably exist. This can happen in practice when the data owner builds the training set using individuals from a certain socioeconomic class, a certain geographic location, etc. In this evaluation, we consider a particular case of this bias, where we build the member set by selecting individuals based on a particular implicit attribute.

We take the `adult` dataset described above and split it based on the education attribute by choosing individuals that are high school graduates (15 784 samples) versus those with any other value (33 058 samples). Once we have split the dataset, we remove the education attribute used to make the split, as it will be of no use in the classification when the entire set has the same value. After performing the splits, we sample 12 500 samples from the high school graduate set to be the *owner set*, and 10 000 samples from the other set to become the *non-owner set*. We select 10 000 samples from the owner set to use as the *training set* (i.e., members) and use the remaining 2 500 samples as the *testing set*. The 10 000 samples in the non-owner set are the non-members. We use the remaining 26 342 data samples (leftovers from both sets) to train the shadow models for the shadow model attack.

We train the model with the training set, and run the membership inference attack on the 10 000 members and 10 000 non-members, using the average number of correct guesses to compute the membership advantage. We use the 2 500 samples in the testing set to measure the classification accuracy. We repeat the training and evaluation process 40 times to account for the randomness of the training algorithm. Figure 2 shows the average accuracy of the logloss attack and the shadow model attack (shaded areas are the 95% confidence intervals for the mean), as well as the average classification accuracy over the training, testing, and non-member sets. For the sake of comparison, we also show the performance of the attacks when the owner and non-owner sets are the result of a random split of the original database (this is the empirical equivalent of the IID case).

We see that the performance of both membership inference attacks is modest (always below 0.15 advantage), but the accuracy of these attacks is significantly larger when the members (and non-members) are biased compared to the IID scenario. More importantly, we can see that the differential privacy bounds do not hold: for high privacy regimes $\varepsilon < 0.1$, the empirical membership advantage is above the bounds. It is worth noting that the confidence intervals are very small in this evaluation, which means that the empirical membership advantage is close to the true advantage, and therefore increasing the number of runs would not change the conclusions of this evaluation. This evaluation highlights the weaknesses of models trained on biased data against membership inference attacks.

4.3 Cluster Split

The previous evaluation already reveals that training set bias can improve the performance of MIAs even beyond the theoretical bounds that hold in the IID case. However, the over-

[6] https://github.com/bargavj/EvaluatingDPML
[7] https://github.com/tensorflow/tensorflow

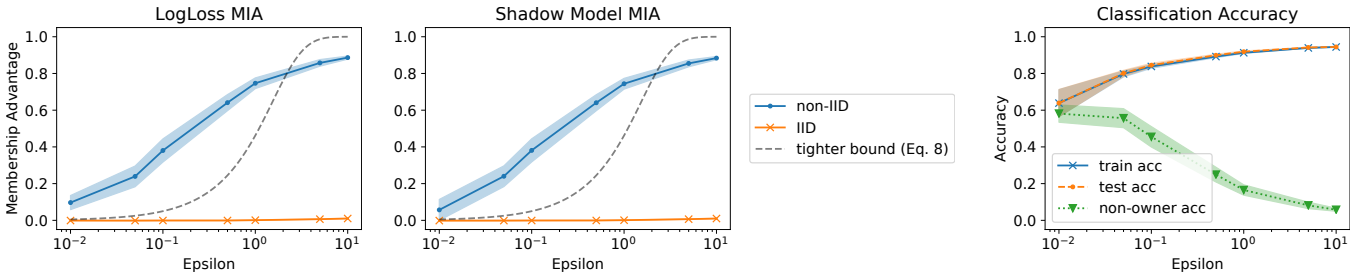

Figure 3: `adult` dataset with cluster split

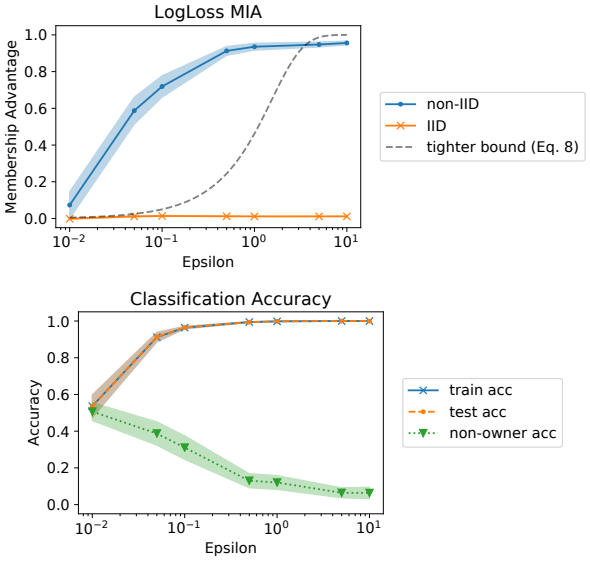

Figure 4: `compas` dataset with cluster split

all advantage score was still modest (below 0.15). In the following evaluation, we aim at designing a data split that further increases the performance of these attacks. Particularly, we design this split with the idea of benefiting the logloss attack [30], but we will observe that these techniques also generalize to the shadow model attack.

The logloss attack is particularly effective when the classification loss is low for members and high for non-members. Therefore, we design a split such that the samples of each class in the member set are geometrically different from the samples of the same class in the non-member set. We utilize k-means clustering to achieve this, as follows: for each class, we take all samples from that class and use k-means to separate them into two clusters. Then, we randomly assign one of the clusters to the owner set (i.e., the members, used for training, as well as the testing set — see Table 2), and the other to the non-owner set (i.e., non-members). This ensures that members of a certain class are geometrically different from non-members of the same class. That is, the characteristics used to distinguish between the classes will likely be significantly different between the member and non-member set.

For this evaluation, we use the Adult dataset (`adult`) and the Compas dataset (`compas`). We perform no additional modifications to this data other than the basic prepossessing mentioned in Section 4.1. After performing the cluster split on the `adult` dataset we get two sets, one of size 21 189 and another of size 27 653. We sample 12 500 data points from the first set to make the owner set and 10 000 data points from the second to make the non-owner set. We use the remaining 26 342 data samples (leftovers from both clusters) as the shadow data; i.e., to train the shadow models for the shadow model attack. For the `compas` dataset we do not run the shadow model attack and simply take the smaller cluster (size 2 574) and make it the owner set. We then randomly sample n (2 059) samples from the other cluster for the non-owner set.

As before, we run the training and evaluation 40 times and plot the average results as well as 95% confidence intervals. We show the results of this evaluation in `adult` and `compas` datasets in Figures 3 and 4, respectively. Once again we compare to a baseline scenario with the same dataset and model, where the owner and non-owner set is randomly split. We observe that, in both datasets, the performance of the attack when the owner/non-owner sets are spatially separated is significantly higher than when we these sets are the result of a random split. From the accuracy plots we see that there is a clear distinction between the owner and non-owner sets which closely follows the performance of the logloss attack. The cluster split introduces a higher bias in the members/non-members than the attribute-based split we evaluated in the previous section. Thus, we observe that the same attacks on the cluster split (Fig. 3) significantly outperform their attribute-based counterparts (Fig. 2). This evaluation reveals that member/non-member set bias plays a prominent role in determining the success of membership inference attacks, and that DP-based upper bounds on the membership advantage do not hold when such bias is present in the datasets.

4.4 Fairness Discussion

The evaluations in this section explore one particular instance of a non-IID scenario, namely the case where members and non-members are independent samples from different dis-

tributions. We have evaluated this empirically by designing splits that maximize the membership advantage, and we have seen that the bounds break using off-the-shelf attacks. One could argue that these dataset splits are not fair since, when the member and non-member distributions are distinct, an adversary with background information on these distributions could achieve high accuracy without even using the ML model. However, it is unsurprising that DP is insufficient as a protection against membership inference when the adversary has prior information about the membership status of a sample, since DP provides a relative privacy guarantee, not an absolute one.

Despite this, we argue that the evaluations in this section still allow us to draw reasonable conclusions about the protection of DP in non-IID scenarios. This is because we use off-the-shelf attacks that rely solely on the target model (and not background information) to draw their conclusions. We see evidence to support this claim in all of our evaluations. For example, when the model is extremely noisy ($\varepsilon = 0.01$), the attacks cannot make meaningful predictions $\texttt{Adv} < 0.1$. We acknowledge that the strong splits provided in this section are very unlikely to occur naturally in practice. However, we recall that their purpose is to show that the bounds do not always hold, even when using real datasets and off-the-shelf attacks.

We remark that we must be careful to define a membership experiment in which the membership advantage is a fair metric of the effectiveness of the private training algorithm. Motivated by this, in the next section we define a fairness requirement for membership experiments and propose an extension of Yeom et al.'s membership experiment to the non-IID scenario that meets this fairness requirement, and show that the DP bounds still break in this case.

5 Analysis of Membership Inference with Data Dependencies

Yeom et al.'s membership experiment assumes that members and non-members are independent samples of a global population distribution $\mathcal{D}$ (IID assumption). This is a strong assumption that in many practical scenarios will not hold. In the previous section, we have seen that MIAs are strong in the particular non-IID scenario where members and non-members are sampled independently from different distributions, but we have argued that this experiment is not entirely fair.

In this section, we first introduce a *fairness requirement* for membership experiments. If a membership experiment meets this requirement, then the membership advantage measured following the experiment will better represent the effectiveness of the training algorithm against MIAs. Then, we extend Yeom et al.'s membership experiment ($\texttt{Exp}$, in Algorithm 1) to a non-IID scenario ensuring the fairness requirement is still met. Finally, we explain why, in this case, the bounds on the membership advantage still do not hold, and we empirically find a lower bound for the worst-case advantage in the non-IID case.

5.1 Fairness of a Membership Experiment

Differentially private training algorithms ensure that, when ε is small, the adversary does not get useful information about whether a sample z is or is not a member of the training set by observing the released model. However, if the prior distribution of members and non-members is different, and the adversary has background information about these distributions, they could guess the membership status of z without even observing the released model. In that case, the membership advantage is not a true measure of the effectiveness of DP as a defense against MIAs. Thus, we claim that an experiment is fair if it meets the following condition:[8]

Definition 5.1 (Fairness Requirement)**.** A membership experiment $\texttt{Exp}$ is fair if and only if the (marginal) distribution of z is the same regardless of the membership status b.

In a fair experiment, the distributions of any given member (z when $b = 0$) and non-member (z when $b = 1$) are identical. This implies that, when $\Pr(b = 0) = \Pr(b = 1) = 0.5$ (the balanced prior scenario we are considering in our evaluations), before observing the model, the adversary cannot do better than randomly guessing the membership status b. In other words, when $\varepsilon = 0$, the membership advantage is $\texttt{Adv} = 0$. Any reduction of the privacy protection (increase in ε) can cause an increase in $\texttt{Adv}$, and the membership advantage will be a fair metric to quantify the effectiveness of DP as a defense against MIAs.

5.2 Membership Advantage with Data Dependencies

We define a membership experiment that does not follow the IID assumption but meets the fairness requirement in Definition 5.1. Let $\boldsymbol{\mathcal{D}}$ be a *joint distribution* over $\mathcal{Z}^n$, and let $\boldsymbol{\mathcal{D}}^{(i)}$, with $i \in [n]$, be the marginal distribution of its ith component. Here, $[n]$ denotes the set of integers from 1 to n; i.e, $[n] \equiv \{1, 2, \ldots, n\}$. Algorithm 3 defines experiment $\texttt{Exp}^{\texttt{non-IID}}$, which is a generalization of Yeom et al.'s original experiment $\texttt{Exp}$ to the non-IID scenario with data dependencies.

In this experiment, the training set S is sampled from the joint distribution $\boldsymbol{\mathcal{D}}$, which we denote by $S \sim \boldsymbol{\mathcal{D}}$. Note that this means that the ith sample in S will be distributed according to the marginal $\boldsymbol{\mathcal{D}}^{(i)}$, but any two samples from S might have statistical dependencies. When $b = 0$, we draw z uniformly from the training set ($z \sim S$). When $b = 1$, we choose a

[8] Not to be confused with the notion of fairness of machine learning models.

Algorithm 3: $\texttt{Exp}_{\texttt{non-IID}}(\texttt{Att}, A, n, \boldsymbol{\mathcal{D}})$

Sample $S \sim \boldsymbol{\mathcal{D}}$, train $a = A(S)$;
Choose $b \sim \{0,1\}$ uniformly at random;
Draw $z \sim S$ if $b = 0$, or $z \sim \boldsymbol{\mathcal{D}}^{(i)}$, for $i \sim [n]$, if $b = 1$;
$\texttt{Exp}(\texttt{Att}, A, n, \boldsymbol{\mathcal{D}}) = 1$ if $\texttt{Att}(z, a, n, A, \boldsymbol{\mathcal{D}}) = b$; else 0.

marginal uniformly at random ($\boldsymbol{\mathcal{D}}^{(i)}$ with $i \sim [n]$) and sample z from this marginal. The adversary wins if they successfully guess z's membership status, b. The experiment meets our fairness requirement, since the marginal distribution of z is the same regardless of the value of b (note that we force this in the experiment, by sampling z from a marginal of $\boldsymbol{\mathcal{D}}$ when $b = 1$). This implies that using $\varepsilon = 0$ yields $\texttt{Adv} = 0$. However, when the trained model leaks information about the training set ($\varepsilon > 0$), the membership advantage will increase.

We note that existing bounds [11, 30], as well as our new bound in (8), do not apply in this case. This type of result has also been observed in related work. For example, Liu et al. [23] show that data correlations can allow an adversary to break a DP bound that otherwise holds in the IID scenario.

The reason for this is that, when training samples have dependencies, leaking one sample also leaks information about the others [28]. In the MIA case, when generating a model a with a training set S with dependent samples, the contribution of each sample $z \in S$ to the model leaks information about the other members. In order to achieve the bound in (8) in the non-IID case, the data owner would have to add enough noise to hide the contribution of *each* of the dependent training samples to the output. For example, when all the training samples are mutually dependent, one can prove that adding the noise that provides $(\varepsilon/n, \delta/n)$-DP would ensure the bound in (8). However, achieving a reasonable amount of privacy without completely sacrificing utility is infeasible in this scenario, as n can be very large.

Even though we cannot determine meaningful upper bounds on the membership advantage when the training data is dependent, we can still obtain lower bounds on the worst-case scenario by empirical evaluation. Any empirical result is a *lower bound* on the worst-case (maximum) membership advantage one could get in the non-IID case. Therefore, we run a synthetic evaluation where we aim to maximize the adversary's performance. We carefully craft a synthetic population distribution $\boldsymbol{\mathcal{D}}$ with the goal of maximizing the accuracy of the logloss attack. As with our previous evaluations in Section 4.3, we will see from our results in Section 5.4 that this also generalizes to the shadow model attack. Our results show that, without the IID assumption, one cannot claim meaningful *worst-case* protection guarantees against MIAs by relying on DP training algorithms.

Sampling S from $\boldsymbol{\mathcal{D}}$ with $m = 5$ subpopulations.
1) Choose a subpopulation $\boldsymbol{\mathcal{D}}_j$, with $j \in [5]$, at random.
2) Sample n times independently from $\mathcal{D}_j$, as below:

Prob	Features (x)	Label (y)				
		$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
1/5	$[1,0,0,0,0]+N$	1	2	3	4	5
1/5	$[0,1,0,0,0]+N$	2	3	4	5	1
1/5	$[0,0,1,0,0]+N$	3	4	5	1	2
1/5	$[0,0,0,1,0]+N$	4	5	1	2	3
1/5	$[0,0,0,0,1]+N$	5	1	2	3	4

where $N \sim N(\mathbf{0}, \sigma\mathbf{I})$.

Figure 5: Sampling from the synthetic population distribution. Example with $m = 5$ subpopulations.

5.3 Synthetic Population Distribution

We build a joint distribution $\boldsymbol{\mathcal{D}}$ that is a *mixture distribution* with m (unweighted) mixture components, denoted by $\boldsymbol{\mathcal{D}}_j$ for $j \in [m]$. This simply means that, in order to sample S from $\boldsymbol{\mathcal{D}}$, we can simply choose a mixture $\boldsymbol{\mathcal{D}}_j$ uniformly at random (with probability $1/m$), and then sample S from that mixture component. Note that the mixture components $\{\boldsymbol{\mathcal{D}}_j, j \in [m]\}$ specify a distribution for n samples (i.e., a distribution over $\mathcal{Z}^n$) and they are not to be confused with the marginals $\{\boldsymbol{\mathcal{D}}^{(i)}, i \in [n]\}$, which specify a distribution for a single sample $z \in \mathcal{Z}$. Each mixture component considers that training set samples are independent, i.e, sampling $S \sim \boldsymbol{\mathcal{D}}_j$ is equivalent to sampling n times from a distribution over $\mathcal{Z}$ we denote by $\mathcal{D}_j$.

Each sample $z = (x, y)$ from $\mathcal{D}_j$ has a feature vector $x \in \mathbb{R}^m$ and a label $y \in [m]$ (i.e., the samples have m attributes and there are m classes in our classification problem). In order to sample from $\mathcal{D}_j$, we first choose a label uniformly at random $y \sim [m]$, and then generate the attribute vector x such that its kth component follows:

$$x_k = \begin{cases} N(1, \sigma) & \text{if } k = (y + j - 2 \bmod m) + 1\,, \\ N(0, \sigma) & \text{otherwise}\,. \end{cases} \tag{18}$$

Figure 5 shows the particular case of $m = 5$. In this case, to generate the training set S we first choose a marginal $j \in \{1,2,3,4,5\}$ uniformly at random, and then sample n times independently from $\mathcal{D}_j$. To generate a sample z from $\mathcal{D}_j$, we randomly select one of the five rows of the table in Fig. 5. The feature vector x of the sample will be the vector specified by that table row, plus a small amount of Gaussian noise added to each of its components ($\sigma = 0.01$ in our evaluations). The label y is determined by the row and the subpopulation j. For example, if we choose the second row of the table and we want to sample from $\mathcal{D}_3$, then we select $x = [0,1,0,0,0] + N$

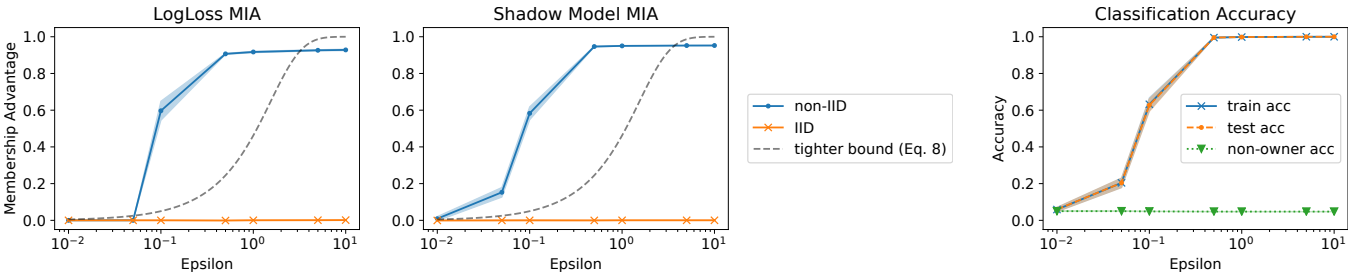

Figure 6: Synthetic population distribution evaluation

and $y = 4$, and build the sample $z = (x, y)$. We repeat this procedure n times to generate every element in S.

The rationale behind this design is the following. Given n samples from $\mathcal{D}_j$, for a fixed subpopulation j, the network will learn a trivial classification task which consists in finding the index of the largest element in x. This model will likely have a high accuracy in samples drawn from $\mathcal{D}_j$, but will most likely fail on samples drawn from $\mathcal{D}_k$ for $k \neq j$, since the classification task in each subpopulation is identical except for a relabeling. This means that the logloss of training samples (members) will likely be very low, while the logloss of non-members (which are, with probability $1 - 1/m$, sampled from a different mixture), will be large. Following the same argument that we used to design the k-means split in Section 4.3, this means that the logloss attack will have a large success probability (if m is large enough).

5.4 Evaluation Results

We run an evaluation that follows $\mathtt{Exp}_{\mathtt{non\text{-}IID}}$ with the synthetic $\boldsymbol{\mathcal{D}}$ explained above, using $m = 20$ subpopulations with $\sigma = 0.01$. As in our evaluations with `adult` dataset, we generate 10 000 training samples, 2 500 testing samples, and 10 000 non-members (these are sampled independently from a marginal of $\boldsymbol{\mathcal{D}}$). We use the same model parameters as in previous evaluations. Figure 6 shows the performance of the logloss attack and the shadow model attack, as well as the classification accuracy. As we mentioned above, there is a probability $1/m$ that a non-member is sampled from the same mixture as all the members. If the attack classifies all non-members sampled from the member's mixture as members, the membership inference attack will cap at $\mathtt{Adv} \leq 2 \cdot (1 - 1/m) - 1 = 0.9$. We observe that the logloss attack caps exactly at this value. However, the shadow model attack reaches a slightly higher value, since it is able to detect some non-members that are sampled from the same mixture as the members. Both attacks perform similarly, and they break the membership advantage bound that holds in the IID scenario.

6 Discussion

6.1 A False Sense of Security

Existing empirical evaluation of membership inference attacks [18] indicate that the attack performance an adversary can achieve is considerably below the upper bound provided by differential privacy. This might lead one to believe that differentially private training mechanisms provide sufficient protection against MIAs, even for large values of ε. However, we have shown above that this is not always true. This raises questions as to why MIAs on differentially private ML have performed so poorly in prior work.

These past evaluation results [18] use a random split of the dataset to create the member and non-member sets (these correspond to the orange lines in our membership advantage plots). This random split loosely simulates the IID scenario, where the theoretical DP bounds hold. When the assumption of IID sampling is relaxed and models are trained on dependent or biased data samples, we show that they are vulnerable to MIAs even when DP noise is applied.

The second reason why previous works exhibit low MIA performance is that current state-of-the-art MIAs are far from optimal. In the proof of Theorem 3.1, we derive our bound for a strong adversary that knows all the samples in the training set except for one ($\tilde{S}$) and must decide if the remaining sample is either z or z'. This adversary has considerably more information than in existing membership inference attacks [27, 30] that have been evaluated empirically. Therefore, it is not surprising that these attacks often fall below the bound. We want to emphasize that *this does not mean the bounds hold*, but merely that 1) current adversaries are weaker than the strong (more informed) adversary that DP implicitly assumes, and 2) the chosen datasets might not exhibit strong correlations that could possibly be exploited by current MIAs.

6.2 Correlation and MIAs

Our results support the fact that DP does not offer worst-case protection guarantees when input data is correlated. Work by Tschantz et al. [28] addresses an ongoing debate between what they refer to as a *causal* and *associative* view of differential privacy. The debate centers around the limits of

differential privacy's guarantees and the practical notion of privacy that it ought to be expected to provide. As defined by Tschantz et al. [28], the causal view is concerned with the system—a machine learning model in our case, and ensuring that the effect size of a single data point is limited so its inclusion cannot be determined. The associative view, instead, is concerned with the distribution of the data points and associations that can be learned, despite differential privacy, about individuals included in a dataset. Dwork [9] is clear that differential privacy does not wholly prevent privacy disclosures from occurring but does ensure that an individual with data present in a database will not be the cause of the breach. This suggests that the original definition of DP takes a causal view of differential privacy, limiting the impact of a single point rather than obscuring an underlying correlation exclusive to member points. This view supports our findings in that applying DP to a machine learning model may limit the impact of a single point on the model, but may not prevent an adversary from learning the underlying association that distinguishes member from non-member points.

Our results indicate that data dependencies can cause substantial increases in the vulnerability of a model. Even though the specific data dependency scenarios we have studied could be considered unrealistic, one could argue that assuming data samples are totally independent and identically distributed is also unrealistic in many cases. We posit that in practice one could expect to be somewhere in between these two cases. A dataset may be distinct from similar data samples for a multitude of reasons; e.g., time (day, year, century) when data was collected, lighting levels for image data, location of data collection, etc. There are examples of real datasets where this type of data dependency exists. For example, Buolawmini and Gebru [7] and Bagdasaryan et al. [5] find datasets that exhibit a bias towards a particular group of samples. Therefore, the IID assumption provides an over-optimistic picture of the protection that differential privacy offers to ML models, and fails to capture real world cases where MIA performance can be significantly higher. Practitioners must be aware of the extra risks posed by correlated data when crafting their datasets and be cautious not to overstate the protections that they offer.

6.3 Vulnerability Can Be Difficult to Detect in Practice

Although differential privacy is widely used as a privacy protection mechanism in machine learning, we have shown that it does not always provide the protection that it is credited to [11, 18, 19, 27]. Furthermore, the cases where current attacks break the bounds provided by differential privacy can be difficult to detect. This is especially true when the data owner may not be aware of a bias in their dataset, whether because it has not been explicitly measured or because it is not obvious. For example, in the scenario of Section 4.2, we selected members based on their education level and then removed the education attribute from the dataset. In this case the data analyst might not be able to identify that their dataset includes only individuals with high school diplomas. Furthermore, the model does not show indication of this member set bias either, since the train and test accuracy are nearly identical. This means that the model may function well within the data owner's testing environment but jeopardize the privacy of its members once released. This implies that tools such as ML Privacy Meter [25] might give the data owner a false sense of security since these tools only consider the data they are given, which may be biased. These biases might unintentionally occur in practice; e.g., when building a dataset by taking data samples from individuals of a particular socioeconomic status or geographical location. Our evaluations show that building an equitable training set, that fairly represents the "whole population" (orange "IID" lines in Figures 2–6), is critical towards achieving machine learning models resilient against membership inference attacks.

7 Related Work

7.1 Inference Attacks

Inference attacks on machine learning models have been studied extensively and are known to take several forms:

Attribute Inference Attacks. Datasets containing sensitive information may eventually be publicly released with the sensitive information redacted or removed. In an attribute inference attack, the adversary attempts to recover this missing data [13, 30]. Fredrikson et al. [12, 13] demonstrated that an ML model output can be exploited under certain conditions to leak sensitive information about the inputs to a model. In a concrete example of this attack, Fredrikson et al. [13] showed that a model trained to predict the Warfarin dosages of patients can be utilized by an adversary to complete missing genotype data in partially obscured medical records that are similar to those used to train the model. Because some sensitive information is statistically true for a population, this type of inference attack can have implications that extend beyond a single model and are difficult to mitigate. Later work by Fredrikson et al. [12] found that an adversary with access only to the outputs of a model could use the confidence scores to piece together information about training data without any prior knowledge of the training set. Yeom et al. [30] show a distinct relationship between membership and attribute inference with results suggesting that the existence of an attribute advantage implies a membership advantage.

Property Inference Attacks. The concept of a property inference attack in ML involves inferring an attribute of a target classifier's training dataset, such as the proportion of members in the dataset that are students [4]. Ganju et al. [14] developed property inference attacks against fully connected neural net-

works that leverage permutation invariance and are effective on real-world datasets. Our work suggests that membership in a dataset may itself be the result of a common property between data samples. In this sense, an MIA could be considered as a property inference attack specifically targeting membership.

Membership Inference Attacks. Aside from the examples by Shokri et al. [27], Yeom et al. [30], and Jayaraman and Evans [18] that we have explored in detail throughout the paper, there are other notable works in this space that we would be remiss to omit. Recent work by Jagielski et al. [17] improves MIA success by developing a novel data poisoning attack. While this work complements ours by raising the lower bound, it does so under the IID assumption. Early work in this space by Li et al. [22] made the connection between ε-DP and membership privacy and introduced a membership privacy framework for ensuring an adversary can not significantly increase its ability to identify members and non-members under the IID assumption. Nasr et al. [26] provide a comprehensive privacy analysis of MIAs under several white-box conditions and design an inference attack that targets vulnerabilities in the stochastic gradient descent algorithm. They find that each training sample uniquely impacts the gradients of the model's loss function—which seems like a problem differential privacy would be best suited to solve. However, they stop short of applying or discussing theoretical bounds and privacy guarantees.

7.2 Related Findings in Differentially Private Machine Learning

Yaghini et al. [29] demonstrate that membership inference attacks may be more successful for certain vulnerable data points within ML models. Following up on this work, Bagdasaryan et al. [5] find that applying Differentially Private Stochastic Gradient Descent (DP-SGD), which involves gradient clipping and random noise addition, results in a disproportionate amount of accuracy loss for vulnerable groups.

Hyland et al. [16] look at the intrinsic privacy properties of SGD and provide experimental evidence from three datasets that SGD provides intrinsic ε values between 2.8 and 7.8. They conclude that changing the seed in SGD is likely to have a far greater impact on the resulting model than including or excluding any given training example [16]. This conclusion seems to offer a reason for the poor performance of MIAs in weak privacy regimes (very high values of ε) that Jayaraman and Evans [18] demonstrate and may indeed pertain to the impact of including or excluding a single point. However, as we have shown, under certain dataset conditions, DP is an ineffective defense against MIA. Even if SGD does intrinsically provide some notion of privacy, it does not provide a sufficient level of protection to prevent highly successful MIAs in the attack scenarios we present.

8 Conclusion

Our results challenge the practical validity of *privacy guarantees* computed from analysis based on differential privacy in machine learning. We show that current guarantees hinge on a critical assumption that member and non-member data samples belong to an independent and identically distributed dataset. In this case, differential privacy ensures that the worst-case adversary performance is below certain privacy bounds. We provide a tighter expression for this bound, which shows that the protection offered by differential privacy is stronger than previously thought when all data samples are independent and follow a global population distribution. Conversely, when member data samples are sufficiently distinct from the rest of the population, or when there are statistical dependencies among the training set samples, we show that current state-of-the-art membership inference attacks pose a far greater threat than previously reported. Our work shows that *training a machine learning model on a representative sample of the population* is critical to protect against membership inference attacks. If a data analyst is unable to ensure this condition is met, they must be aware that differentially private training algorithms *do not guarantee* protection against membership inference attacks, and must be cautious not to overestimate the protection that DP provides.

Acknowledgements

We gratefully acknowledge the support of the Natural Sciences and Engineering Research Council (NSERC) for grants RGPIN-05849, RGPIN-03858, CRDPJ-531191, and IRC-537591, the Royal Bank of Canada, and the Waterloo-Huawei Joint Innovation Laboratory for funding this research. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. This work was also partially funded by the Ontario Graduate Scholarship (OGS) provided by the Ontario government and the Canada Graduate Scholarship-Master's (CGS-M) provided by NSERC. This work benefited from the use of the CrySP RIPPLE Facility at the University of Waterloo.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 308–318, 2016.

[2] Mohammad Al-Rubaie and J Morris Chang. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy*, 17(2):49–58, 2019.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[4] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

[5] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems (NIPS)*, pages 15479–15488, 2019.

[6] Brett K Beaulieu-Jones, William Yuan, Samuel G Finlayson, and Zhiwei Steven Wu. Privacy-preserving distributed deep learning for clinical data. https://arxiv.org/abs/1812.01484, 2018.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[8] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pages 265–284, 2006.

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

[11] Úlfar Erlingsson, Ilya Mironov, Ananth Raghunathan, and Shuang Song. That which we call private. https://arxiv.org/abs/1908.03566, 2019.

[12] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1322–1333, 2015.

[13] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*, pages 17–32, 2014.

[14] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 619–633, 2018.

[15] Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *Journal of Machine Learning Research*, 14(Feb):703–727, 2013.

[16] Stephanie L Hyland and Shruti Tople. On the intrinsic privacy of stochastic gradient descent. https://arxiv.org/abs/1912.02919, 2019.

[17] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? https://arxiv.org/abs/2006.07709, 2020.

[18] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, pages 1895–1912, 2019.

[19] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. https://arxiv.org/abs/2005.10881, 2020.

[20] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[21] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996.

[22] Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: a unifying framework for privacy definitions. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 889–900, 2013.

[23] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnberable: Differential privacy under dependent tuples. In *Network and Distributed System Security Symposium (NDSS)*, volume 16, pages 21–24, 2016.

[24] Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

[25] Sasi Kumar Murakonda and Reza Shokri. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. https://arxiv.org/abs/2007.09339, 2020.

[26] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning. In *IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019.

[27] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[28] Michael Carl Tschantz, Shayak Sen, and Anupam Datta. SoK: Differential Privacy as a Causal Property. In *IEEE Symposium on Security and Privacy (SP)*, pages 354–371, 2020.

[29] Mohammad Yaghini, Bogdan Kulynych, Giovanni Cherubin, and Carmela Troncoso. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. https://arxiv.org/abs/1906.00389, 2019.

[30] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018.

[31] Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. Differentially private model publishing for deep learning. In *IEEE Symposium on Security and Privacy (SP)*, pages 332–349, 2019.