2021 IEEE European Symposium on Security and Privacy (EuroS&P)

On the Privacy Risks of Algorithmic Fairness

Hongyan Chang and Reza Shokri
Department of Computer Science, National University of Singapore (NUS)
firstname@comp.nus.edu.sg

***Abstract*—Algorithmic fairness and privacy are essential pillars of trustworthy machine learning. Fair machine learning aims at minimizing discrimination against protected groups by, for example, imposing a constraint on models to equalize their behavior across different groups. This can subsequently change the *influence* of training data points on the fair model, in a disproportionate way. We study how this can change the information leakage of the model about its training data. We analyze the privacy risks of group fairness (e.g., equalized odds) through the lens of *membership inference attacks*: inferring whether a data point is used for training a model. We show that fairness comes at the cost of privacy, and this cost is not distributed equally: the information leakage of fair models increases significantly on the unprivileged subgroups, which are the ones for whom we need fair learning. We show that the more biased the training data is, the higher the privacy cost of achieving fairness for the unprivileged subgroups will be. We provide comprehensive empirical analysis for general machine learning algorithms.**

***Index Terms*—Trustworthy Machine Learning, Group Fairness, Data Privacy, Membership Inference Attacks**

2021 IEEE European Symposium on Security and Privacy (EuroS&P) | 978-1-6654-1491-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/EUROSP51992.2021.00028

1. Introduction

Machine learning algorithms can be discriminatory against different groups, due to their training algorithm or the bias in their training data. This is shown in various applications from computer vision [1] to word embedding [2]. Fair machine learning aims at addressing this issue [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. At the same time, machine learning models are shown to leak significant amount of information about their training data, which can be exploited by inference attacks [13, 14, 15, 16, 17, 18, 19]. Privacy-preserving machine learning aims at alleviating this problem [14, 20, 21, 22, 23].

Privacy and fairness, as two societal concerns about machine learning, do not exist in isolation. For instance, in the recidivism prediction application, demographic groups (e.g., black defendants and white defendants) should experience similar treatments, namely similar prediction accuracy. Simultaneously, participation in the training data means that the individual had once committed a crime, which is very sensitive and needs to be kept private. Thus, fairness and privacy are both needed for an ethical use of machine learning. It is, therefore, imperative to understand the interactions between them.

Training a model with privacy guarantees can lead to disparate accuracy across different groups in the population. Notably, differentially private models (e.g., DP-SGD [20]) impose a larger accuracy reduction on "under-represented" subgroups [24]. In other words, privacy can come at the cost of fairness. In this paper, we ask the related yet complementary question: *Is there a privacy cost for achieving group fairness?* We study if enforcing fairness constraints on the learning algorithm can impact its privacy risk with respect to the training data.

One way to address this question is through analyzing models which are trained with differential privacy and fairness constraints [21, 25, 26], and evaluating the compatibility of the two measures. In this paper, we choose a complementary adversarial approach. We formalize privacy risk as the success of *membership inference attacks* against machine learning models. This reflects the information leakage of a model about the *individual* data points in its training set. We use this to quantify how fair and unconstrained models differ in their information leakage about different groups in their training data.

We assume the adversary observes the model's predictions and aims to find a "distinguisher" that identifies members of the training set from non-members. Finding a single global "distinguisher" for all data points (e.g., a threshold on the loss of the model on its inputs) is the prevalent method in the existing membership inference attacks [15, 17, 18]. However, this approach is evidently sub-optimal. In practice, data samples from different groups can have different underlying distributions. Thus, the machine learning models might learn different patterns on each group, for the same task. This is what leads to performance disparity of the model, which motivates using fair algorithms. In other words, the way that a model's predictions (loss) on training versus test data differs, can be distinct from one group to another. This can be exploited by the membership inference attacks. We, thus, propose an effective attack strategy where the adversary finds a "distinguisher" per sub-group. We empirically show that this simple modification of existing membership inference attacks results in a higher attack accuracy, thus leads to a more accurate estimation of the privacy risk. We focus on the information leakage of models through their predictions: *black box* setting, in which the adversary cannot observe the internal state of the model (e.g., its parameters and gradients).

We show that fairness comes at the cost of privacy. Based on our attack strategy, we empirically show that the fairness-aware learning has a disparate impact on the privacy risk of subgroups, and in particular, it increases the privacy risk of the unprivileged subgroup. Furthermore, there is a trade-off between fairness and privacy. When the underlying data and the corresponding unconstrained model are more "unfair", the trained fair models leak more information about the unprivileged subgroups. Addition-

DOI 10.1109/EuroSP51992.2021.00028

ally, the more fair a model is, the higher the privacy risk of the model on the unprivileged subgroups will be.

Fairness constraints force models to perform equally on all the subgroups. Yet, when the size of an unprivileged subgroup is small, or due to its complexities and variance, it is hard to fit a model on it, fair models memorize the training data from the unprivileged subgroups (instead of learning a general pattern on them). This memorization gives rise to a high privacy risk, and it becomes easier for the adversary to infer membership of the training data. Hence, the unprivileged subgroups have a higher privacy risk on fair models.

We perform extensive empirical analysis on machine learning models. In the evaluation, we use synthetic data to analyze how, when, and why fair models leak more information about their training data. We also conduct experiments on multiple real-world datasets, including the Law School dataset [27], Bank Marketing dataset [28], and COMPAS datasets [29].

2. Background

In this section, we present the definitions for group fairness and data privacy.

2.1. Machine Learning

We consider supervised machine learning with the focus on classification tasks. Let $M : \mathcal{X} \to \mathcal{Y}$ be a machine learning model that maps the input (feature) space $\mathcal{X}$ to the output (label) space $\mathcal{Y}$. Let ℓ be the loss function, and S be the training set of size n. We use M_S to denote the model trained on S. We assume each data point in S is sampled i.i.d. from a data distribution $\mathcal{D}$. We refer to the models obtained by the standard learning algorithm (without fairness requirements) as *unconstrained models*.

2.2. Fairness

The central problem in fair machine learning is to ensure that the machine learning model does not discriminate against individuals with specific values in their protected attributes (e.g., race, gender). We represent each data point as $z = (x, g, y) \in \mathcal{X} \times \mathcal{G} \times \mathcal{Y}$, where $x \in \mathcal{X}$ is the set of model's input features, $g \in \mathcal{G}$ is the *protected attribute*, and $y \in \mathcal{Y}$ is the label. The protected attributes partitions the population. Let $\mathcal{D}_g^y$ denote the distribution of data with protected attribute $G = g$ and label $Y = y$. We use G_g to represent the group $\{(X, G, Y)|G = g\}$. We use G_g^y to represent the subgroup with $\{(X, G, Y)|G = g, Y = y\}$.

Without loss of generality, we consider the binary classification setting, where $y \in \mathcal{Y} = \{-, +\}$. We use X, Y, and G to denote the random variables associated with the feature vector, the label, and the protected attribute, respectively. Input features X might include G. For instance, each data point can correspond to an applicant in a loan approval system, and X could be the demographics, income level, and loan amount, and G could be the applicant's race. Input features X might include G, or otherwise include other features such as zip code, which is often correlated with race.

Group fairness measures require that different protected groups, on average, are treated similarly by the model. In this paper, we mainly focus on *equalized odds*, which is a widely-used definition for group fairness [6]. We also evaluate other group fairness notions, notably equal opportunity [6] and false-positive parity [6], which are explained in Section 4.

A model is fair with respect to equalized odds if the model error on different groups is the same. In other words, given the true label for a data point, a fair model's prediction on a data point and its protected attribute should be conditionally independent. Following prior work [3, 30], we use a relaxed notion of equalized odds:

Definition 1** **(δ - Equalized Odds Fairness). A classifier M satisfies δ-Equalized Odds with respective to the protected attribute $\mathcal{G}$, if for all $g, g' \in \mathcal{G}$, the false positive rate and false negative rate of the classifier for group $\{G = g\}$ and $\{G = g'\}$ are within δ range of one another.

$$\Delta(M, \mathcal{D}) \triangleq \max_{\substack{y \in \{-,+\} \\ g, g' \in \mathcal{G}}} \Big| \Pr_{\mathcal{D}}[M(X) \neq y | S = g, Y = y] - \Pr_{\mathcal{D}}[M(X) \neq y | S = g', Y = y] \Big| \leq \delta, \tag{1}$$

where the probabilities are computed over the data distribution $\mathcal{D}$. We refer to Δ as the model's **fairness gap** under equalized odds. A model satisfies exact fairness under equalized odds when $\delta = 0$.

In practice, the data distribution $\mathcal{D}$ is unknown. A fair model is thus obtained by ensuring δ-fairness empirically on the training set S. We can minimize the model's empirical loss under δ-fairness as a constraint, or through post-processing [6]. We refer to δ as the *enforced fairness level*.

2.3. Membership Privacy

We follow the privacy notion underlying *differential privacy* [31]: privacy is preserved if the output distributions of an algorithm on two neighboring datasets are almost indistinguishable. In other words, given an observation from a privacy-preserving algorithm, an adversary is unable to tell whether the record of a participant is in the input dataset or not.

Thus, we measure the privacy risk of an individual as the success of an adversary whose goal is to infer whether the individual's record is part of the input dataset. Such attacks are called *membership inference attacks* which are used as a *tool* to measure information leakage of different machine learning algorithms, including deep learning algorithms [17], adversarially robust learning algorithms [16], learning algorithms for explanations models [14], learning algorithms for embedding models [13], and reinforcement learning algorithms [32].

3. Privacy Analysis

In this section, we present our generic approach for the analysis of privacy risks. In Section 4, we use this framework to perform our empirical analysis.

3.1. Definition of Privacy Risk

We compute privacy risk as the accuracy of membership inference attacks (i.e., the probability that an adversary can correctly infer if a data point is part of the input dataset). We use the following game between an adversary and a challenger, to formalize membership inference attacks.

Attack Game 1 (Membership Inference).

1) Adversary chooses a data point z, and sends it to the challenger.
2) Challenger chooses a secret bit $b \leftarrow \{0,1\}$ uniformly at random, and samples dataset $S \sim \mathcal{D}^n$. If $b = 1$, the challenger overwrites a random element in S with z.
3) Challenger runs algorithm A on S and sends its outputs A_S to the adversary.
4) Adversary runs an inference attack $\mathcal{A}$, and tries to infer the secret bit as $\hat{b} \in \{0,1\}$.
5) The game outputs 1 (indicating that adversary wins) if $\hat{b} = b$, and 0 otherwise.

We define privacy risk of algorithm A with respect to an individual data point z as the probability that the most powerful adversary wins the attack game.

Definition 2 (Individual Privacy Risk). Given an algorithm A and data distribution $\mathcal{D}$, the privacy risk of A with respect to data point z is

$$\text{PR}(z, A, \mathcal{D}) \triangleq \max_{\mathcal{A}} \Pr[\text{Attack Game outputs 1}],$$

where the probability is taken over all the randomness in Attack Game 1.

The individual privacy risk is equivalent to the average true positive and true negative rates $\frac{1}{2}(\Pr[\hat{b} = 1|b = 1] + \Pr[\hat{b} = 0|b = 0])$ of the adversary.

Definition 3 (Subgroup Privacy Risk). We define the privacy risk of algorithm A with respect to subgroup G_g^y (i.e., data points with label y and protected attribute g) as

$$\text{PR}(G_g^y, A, \mathcal{D}) \triangleq \mathbb{E}_{z \sim \mathcal{D}_g^y}[\text{PR}(z, A, \mathcal{D})], \qquad (2)$$

which is the expectation of the privacy risk of individual data points in G_g^y.

3.2. Quantifying Privacy Risk

We assume the adversary has black-box access to the model, and can compute the loss of the model on any input data. To run a membership inference attack in this setting, the adversary can use a single attack model on all input data [17]. A simple attack model is to compare the model's loss on an input with a threshold. The attack outputs "member" if the loss is below the threshold, and "non-member" otherwise [15, 18]. This attack is based on a single distinguisher between members and non-members, which makes the attack sub-optimal. In practice, the training data might be composed of samples from (slightly) different distributions. This means the model might learn (and memorize) distinct patterns from different parts of the training set. Thus, the membership inference attack that is adapted to each sub-population should potentially perform better than a fixed attack model (based on a single loss threshold). Specifically, in the case of fair models, where the algorithm explicitly treats different subgroups differently in order to equalize its error across them, the adversary can design a separate membership inference attack for each subgroup.

Based on this, we propose to use different loss thresholds for different subgroups.

Attack 1. Let $\tau^{(g,y)}$ be the loss threshold for distinguishing training data members from non-members in subgroup G_g^y. On an input $z = (x, g, y)$, and machine learning model A_S, the adversary proceeds as follows:

1) Query the model to obtain $\ell(A_S, z)$.
2) Output 1 ("member") if $\ell(A_S, z) < \tau^{(g,y)}$ and 0 ("non-member") otherwise.

In practice, the adversary can compute the loss threshold based on the knowledge about the population [15, 18] or through using shadow models [17]. In the evaluation, we find a loss threshold that best separates the members and non-members of each subgroup. We measure the individual privacy risk and subgroup privacy risk based on Adversary 1. Note that the attacker can only obtain the best threshold when he knows the training set. Thus, measuring the privacy risks using the best loss threshold gives us a closer estimation of privacy risk under the strongest attack.

4. Empirical Analysis

We use the reductions approach for training fair machine learning models [3][1]. The algorithm produces a randomized classifier, and we compute its expected accuracy in our analysis.

We analyze the success of adversary in Attack Game 1 to quantify privacy risk for individual data points and subgroups. We assume Adversary 1 has black-box access to models, and can observe the model's loss on each query. We measure the **privacy cost** of fair algorithms as the difference between the privacy risk of fair and unconstrained models.

4.1. Experimental Results on Synthetic Data

***Data and Models*.** We generate synthetic datasets, of size $2,500$ records, similar to the prior work on analyzing fairness [33]. Specifically, we generate binary sensitive attributes, for each record, from a Bernoulli distribution $P_g = \Pr[G = g]$, for $g \in \{0,1\}$. We generate binary labels from a Bernoulli distributions $P_g^y = \Pr[Y = y|G = g]$, for $y \in \{-,+\}$ and for all $g \in \{0,1\}$. We generate a 2-dimensional feature vector from four different Gaussian distributions:

$$\begin{aligned}
&\text{For subgroup } G_0^- : \; X \sim \mathcal{N}([0,-1],[7,1;1,7]) \\
&\text{For subgroup } G_1^- : \; X \sim \mathcal{N}([-5,0],[5,1;1,5]) \\
&\text{For subgroup } G_0^+ : \; X \sim \mathcal{N}([1,2],[5,2;2,5]) \\
&\text{For subgroup } G_1^+ : \; X \sim \mathcal{N}([2,3],[10,1;1,4]) \qquad (3)
\end{aligned}$$

where, $\mathcal{N}(\mu, \Sigma)$ represents a Gaussian distribution with mean vector μ and covariance matrix Σ.

We set $P_0 = 0.2$, $P_0^- = 0.1$ and $P_1^- = 0.5$. Accordingly, $P_1 = 0.8$, $P_0^+ = 0.9$ and $P_1^+ = 0.5$. Group G_0,

1. https://github.com/fairlearn/fairlearn

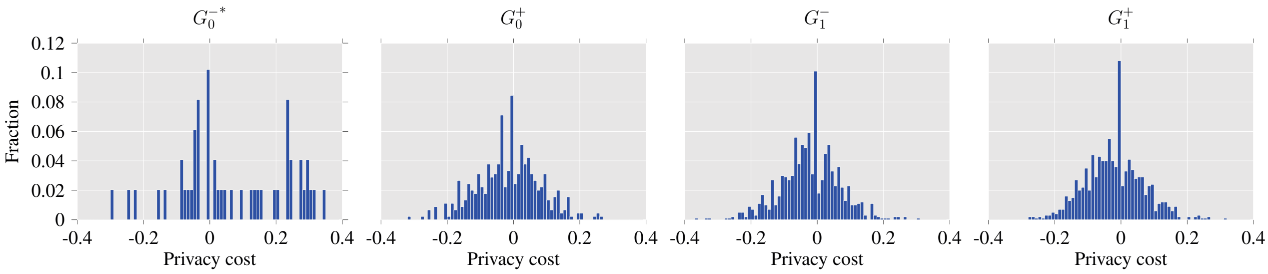

Figure 1: Histogram for individual privacy cost, across different subgroups, on models trained on synthetic data. The x-axis is the privacy cost, which is the difference between individual privacy risk on fair models and unconstrained models. The average value of the individual privacy cost is 0.069, −0.015, −0.02, −0.02 for subgroups G_0^-, G_0^+, G_1^-, and G_1^+ respectively.

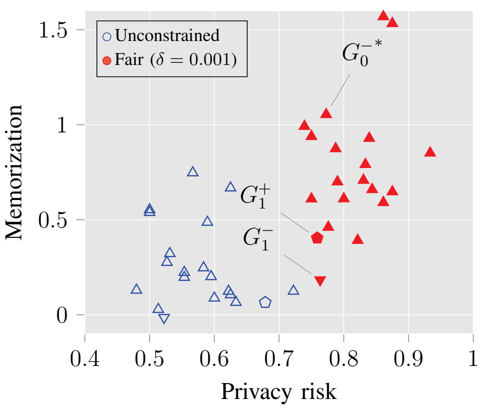

Figure 2: The most vulnerable points on fair models (trained on synthetic data). We find the top 20 vulnerable points that have the highest privacy risk on fair models. For each vulnerable point, we show its privacy risk and the memorization of models before (blue color) and after (red color) imposing fairness constraints. The marker shows which subgroup a point belongs to.

TABLE 1: Accuracy and fairness gap of unconstrained and fair models (with three different fairness constraints δ) on the synthetic datasets. The "Train Δ" and "Test Δ" columns show the fairness gap (as defined in (1)) on the training and test data.

Model	Train acc	Test acc	Train Δ	Test Δ
Unconstrained	86.2%	85.5%	0.373	0.430
Fair ($\delta = 0.1$)	89.1%	87.8%	0.105	0.332
Fair ($\delta = 0.01$)	85.6%	84.0%	0.014	0.283
Fair ($\delta = 0.001$)	84.5%	83.8%	0.001	0.275

TABLE 2: Accuracy of membership inference attacks with a fixed loss threshold for all data points, versus using multiple thresholds one for each subgroup - Synthetic dataset.

Model	Attack	G_0^-	G_1^-	G_0^+	G_1^+
Unconstrained	Single	52.9%	51.2%	51.8%	51.2%
	Group-based	61.8%	52.8%	52.4%	52.2%
Fair ($\delta = 0.001$)	Single	60.8%	51.9%	51.6%	50.8%
	Group-based	69.2%	53.4%	52.5%	51.6%

thus, is the *minority group* with a smaller number of samples. The labels in this group are also unbalanced. Subgroup G_0^- is the smallest subgroup.

We use 50% of the generated data for training and the rest for the testing. We repeat this process (of splitting the data) 30 times, and train 30 unconstrained and fair models for each training set. Each data point in our synthetic dataset, on average, appears in the training set of the model in 15 experiments. We report the average privacy cost over the 30 experiments.

We train fully connected neural network (NN) models with 3 hidden layers with size $\{32, 16, 8\}$, for both unconstrained and fair models. We use Adam Optimizer with learning rate 0.001.

Stronger inference attacks: adapting membership inference attacks to each group. Membership inference attacks use a threshold on the loss of the model on a data point to infer whether the data point belongs to the training set. It is however easier to distinguish members of the training set from non-members *within* a subgroup. Given that the adversary already has the information about the label and the group membership of a data point, he can take advantage of it to construct a stronger inference attack (compared with the prior work [15, 17, 18]). Table 2 shows the privacy risk for all subgroups when we use a single loss threshold (as in the prior work) versus using a separate attack (with a different loss threshold) for each subgroup. For each experiment, we report the average of 30 runs. The results show that using a separate attack threshold for each subgroup effectively increases the attack accuracy, which results in a better estimation of the privacy risk.

Privacy cost of fairness: unprivileged subgroups experience the largest cost. Table 1 shows the performance of unconstrained and fair models with different fairness constraints. Table 4 shows the prediction accuracy of unconstrained models for each subgroup.

We refer to G_0^- as the *unprivileged subgroup*, as it has the worst accuracy (41.6%). Compared with the samples from G_1^- (84.6%). We use asterisk on the unprivileged subgroup G_1^{-*} to distinguish it from other subgroups.

Figure 1 shows the histogram of individual privacy cost, across all subgroups. The plots illustrate the imposed privacy risk due to fairness constraints, notably for the unprivileged subgroup. We observe that G_0^-, unlike other subgroups, has a larger fraction of samples with positive privacy cost.

To further study the privacy cost for individual data points, we identify 20 vulnerable points with the largest privacy risk on fair models. Figure 2 shows the privacy

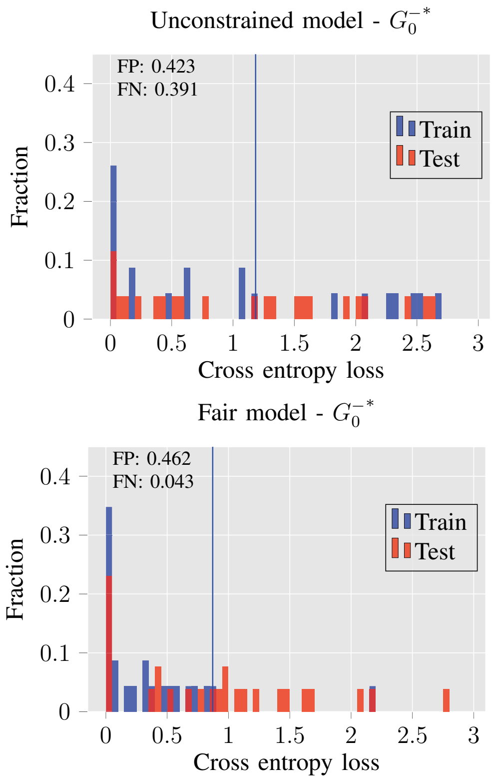

Figure 3: Loss distribution of an unconstrained model and a fair model on subgroup G_0^-. The vertical blue line shows the loss threshold used in the membership inference attack. FN is the false negative, and FP is the false positive rate for the attack.

risk for these data points before and after imposing fairness constraints. We observe that these points are mainly from the unprivileged subgroup G_0^-. In addition, the fairness constraints increase the privacy risk of these data points significantly. For example, the privacy risk of a data point increases from 0.63 to 0.93 after imposing fairness constraints.

Figure 3 compares the loss distribution of fair and unconstrained models on their training and test data from subgroup G_0^-. We observe that *members of training set are more distinguishable from non-members on fair models* compared with unconstrained models. Thus, on this unprivileged subgroup, the adversary achieves a very low false-negative rate on fair model.

Training data memorization on fair models. To further study why fairness constraints increase the privacy risk, we take a closer look at the memorization of individual training data by unconstrained versus fair models.

Figure 4(a) compares the average training accuracy of unconstrained and fair models for all subgroups over 30 experiments. The unconstrained models have a low accuracy on subgroup G_0^- (which, in comparison with accuracy on G_1^-, shows the unfairness of the model according to the equalized odds measure). After imposing the fairness constraints, the training accuracy on G_0^- increases from 50.1% to 78.8%. We study if this improvement in accuracy is due to the training data memorization.

We quantify the memorization of training data points, as the difference in the model's loss when the data point

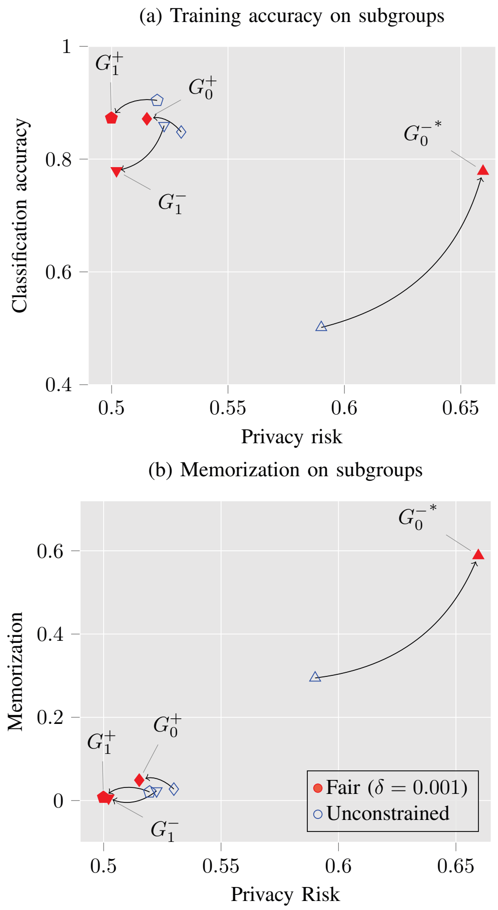

Figure 4: Memorization and training accuracy of fair and unconstrained models on all subgroups - trained on synthetic data. The blue/red color show the results on the unconstrained/fair models. The markers represent different subgroups.

is in the training set versus the case where the data is not in the training set [34].

Figure 4(b) shows that the memorization of fair models on subgroup G_0^- is 0.58, which is two times larger than that of unconstrained models. It shows that fair models memorize the points from G_0^- instead of learning a general pattern about them. On the contrary, fairness constraints only barely change the privacy risk and the memorization for other subgroups. Accordingly, the privacy cost on other subgroups is low. Overall, these results show that fair constraints impose "privacy unfairness".

Accuracy-privacy trade-off on training data. Figure 5 shows the effect of fairness constraints on the individual privacy cost and accuracy gain of training data in G_0^-.

We measure the gain in training accuracy for each data point as the difference in prediction accuracy between fair models and unconstrained models on that data point. Thus, a positive accuracy gain implies that fairness constraints improve accuracy. As we are analyzing the performance change on the training dataset, we only report the true positive rate of the adversary. Recall that the true positive rate of the adversary reflects the probability of correctly predicting the membership of a data point when it *is* a member of the training dataset. The figure shows that there is a clear correlation between the accuracy gain and the

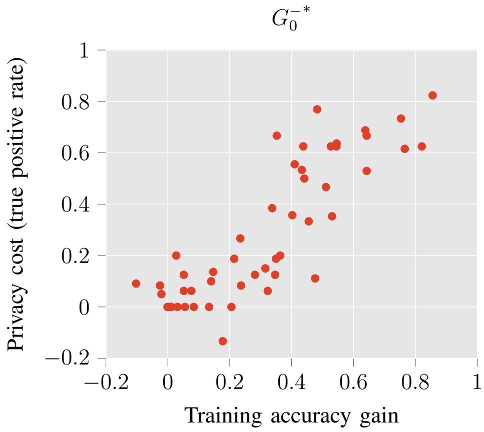

Figure 5: Accuracy gain versus privacy cost on the unprivileged subgroup G_0^-. Each point in the plot represents a data point in the training dataset. The training accuracy gain is the difference in the training accuracy between fair and unconstrained models. The y-axis is the difference in the attacker's true positive rate between fair and unconstrained models on each training point.

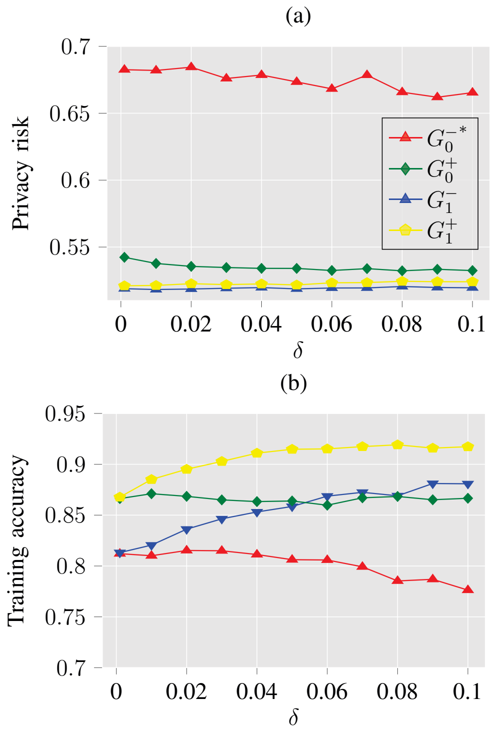

Figure 6: (a) The effect of the enforced fairness level δ on the privacy risk of fair models for different subgroups. (b) The effect of enforced fairness level on the classification accuracy of fair models for different subgroups.

privacy cost on training data in the unprivileged group.

Figure 6 shows the effect of the enforced fairness gap δ on the privacy risk and training accuracy of subgroups. Recall that smaller δ makes the models less discriminatory on the training dataset. We observe that, as δ decreases, accuracy as well as privacy risk for data points in G_0^- increases.

We also analyze the effect of δ on memorization. Figure 7 shows that, when a fair model is less discriminatory

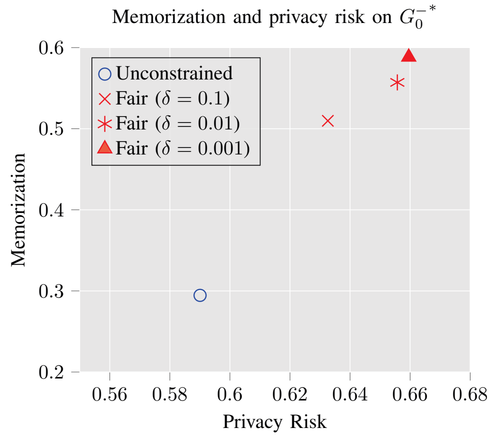

Figure 7: The effect of the enforced fairness gap δ on the privacy risk and memorization of the model - Synthetic dataset.

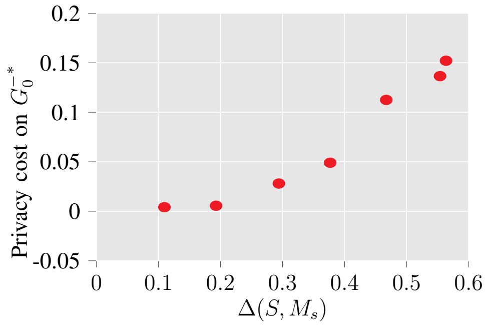

Figure 8: The effect of the unconstrained model's fairness gap (which captures the unfairness that needs to be removed by the fair algorithm) on the privacy cost of G_0^-. The x-axis is the fairness gap of unconstrained models on the training dataset.

on its training data (i.e., δ is smaller), the memorization of data points in subgroup G_0^- is larger (thus, the privacy risk is larger).

Effect of underlying unfairness on privacy cost. We generate multiple synthetic datasets, by varying the mean of distribution (3) for data in subgroup G_0^-, to control the underlying unfairness that this change causes on the unconstrained model. Varying mean results in different separability between positive and negative samples, which influences the difficulty of learning an accurate model that distinguishes between G_0^- and G_0^+ samples. This influences the fairness gap on the unconstrained models. In this setting, a large fairness gap means that subgroup G_0^- has a worse accuracy compared with subgroup G_1^-.

Figure 8 shows the correlation between the fairness gap $\Delta(S, M_S)$ (in the unconstrained model) and the subgroup privacy cost for G_0^-. We observe that a large fairness gap in the underlying unconstrained model results in a large privacy cost (of imposing fairness constraints). In other words, when the unconstrained model is more discriminatory, and there is more need for a fairness mechanism, *reducing accuracy disparity results in a significant privacy disparity for unprivileged groups*.

Effect of dataset size on privacy cost. We generate synthetic data by varying the fraction of samples in G_0

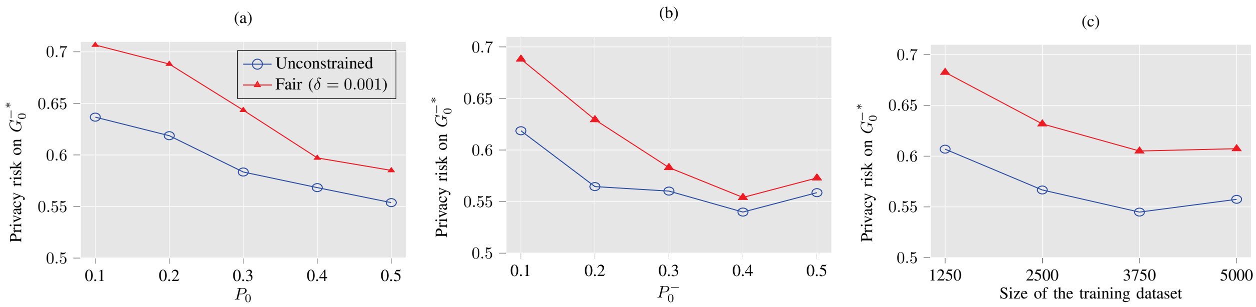

Figure 9: Privacy risk on the unprivileged subgroup G_0^-, for various faction of data in group G_0, various fraction of data in subgroup G_0^-, and for various training set size.

TABLE 3: Prediction accuracy and privacy risk for unconstrained models, and fair models trained using reduction approach [3] and post-processing (PP) approach [6].

Model		G_0^{-*}	G_1^-	G_0^+	G_1^+
Unconstrained	Train acc	52.2%	85.6%	85.3%	89.8%
	Test acc	39.1%	84.7%	86.2%	89.5%
	Privacy risk	0.639	0.521	0.518	0.523
Fair (reduction) ($\delta = 0.001$)	Train acc	80.1%	80.2%	88.4%	88.5%
	Test acc	49.2%	79.2%	88.1%	88.3%
	Privacy risk	0.688	0.521	0.542	0.518
Fair (PP) ($\delta = 0$)	Train acc	48.4%	48.4%	90.7%	90.7%
	Test acc	41%	47.9%	89.4%	90.5%
	Privacy risk	0.55	0.507	0.509	0.504

(by controlling $P_0 = \Pr[G = 0]$) and the underprivileged subgroup G_0^- (by controlling $P_0^- = \Pr[y = -|G = 0]$). Figure 9 shows the effect of the size of groups, and the size of the training set, on the privacy risk on the unprivileged subgroup. As we expect, we observe that a smaller number of samples in the unprivileged subgroup G_0^- or its group G_0 results in a higher privacy cost for the unprivileged group. When the unprivileged subgroup G_0^- is relatively small with respect to subgroup G_0^+, or the whole group G_0 is small, the underlying fairness gap in an unconstrained model increases. This leads to further memorization of the unprivileged group, thus higher privacy risk, when fairness constraints are enforced. Figure 9(c) shows that the training set size has a similar effect on both unconstrained and fair models, as long as the fraction of data in each subgroup remains the same. The results show that, as expected, the privacy risk (with respect to G_0^-) decreases as we increase the size of the training set.

Effect of fair algorithms. Table 3 compares the prediction accuracy and privacy risk of fair models using an in-processing approach [3] (which we analyze in all our experiments) and a post-processing approach [6]. The post-processing algorithm does not improve the training accuracy of the unprivileged subgroup G_0^-. Instead, it decreases the accuracy for another subgroup G_1^- from 85.6% to 48.4% to satisfy fairness constraints. This means that, in the case of using post-processing approach, the fair model does not fit the unprivileged subgroup better than the unconstrained models. Thus, due to underfitting, the privacy risk of fair models is less than that of unconstrained models.

TABLE 4: Prediction accuracy and privacy risk for unconstrained models, versus fair models under different notions of fairness. We use the reduction approach [3] to train fair models and set $\delta = 0.001$.

Model		G_0^{-*}	G_1^-	G_0^+	G_1^+
Unconstrained	Train acc	47.8%	85.1%	85.8%	89.2%
	Test acc	41.6%	84.6%	85.1%	89%
	Privacy risk	0.607	0.521	0.529	0.522
Fair (EO)	Train acc	81.2%	81.3%	86.6%	86.8%
	Test acc	53.8%	80.9%	83.8%	86.4%
	Privacy risk	0.683	0.519	0.542	0.521
Fair (EOPP)	Train acc	47.4%	91%	91.6%	91.7%
	Test acc	39.1%	90.1%	89.7%	91.3%
	Privacy risk	0.605	0.52	0.534	0.522
Fair (FPP)	Train acc	83.1%	83.2%	85.5%	91.8%
	Test acc	54.5%	82.9%	84%	90.9%
	Privacy risk	0.679	0.518	0.535	0.523

Effect of fairness notions. We evaluate the privacy risk of fairness for two other notions of fairness: equal opportunity (EOPP), and false-positive parity (FPP).

Definition 4** (δ **- Equal Opportunity Fairness [6]). A classifier M satisfies δ-Equal Opportunity condition with respect to the protected attribute $\mathcal{G}$, if for all $g, g' \in \mathcal{G}$, the false negative rate of the classifier in the group $\{G = g\}$ and $\{G = g'\}$ are within δ range of one another:

$$\Delta(M, \mathcal{D}) \triangleq \max_{\substack{y=+\\ g,g' \in \mathcal{G}}} \Big| \Pr_{\mathcal{D}}[M(X) \neq y | G = g, Y = y] - \Pr_{\mathcal{D}}[M(X) \neq y | G = g', Y = y] \Big| \leq \delta. \tag{4}$$

By setting $y = -$ instead of $y = +$ in Equation (4), we get the definition of false-positive parity.

In Table 4, we compare the prediction accuracy and privacy risk of unconstrained and fair models that satisfy different fairness notions. Overall, we observe similar patterns across different group-fairness metrics.

TABLE 5: The data partitioning based on the protected attributes, and the percentage of data in different subgroups for the real-world datasets.

Name	G_0^-	G_1^-	G_0^+	G_1^+
Bank (age)	2.2%	85.2%	0.7%	12.0%
COMPAS (race)	28.3%	24.4%	31.7%	15.6%
COMPAS (gender)	12.8%	39.9%	7.3%	40.0%
Law (race)	2.3%	2.7%	13.5%	81.5%
Law (gender)	2.5%	2.5%	41.4%	53.6%

TABLE 6: Accuracy and fairness gap Δ of unconstrained models and fair models (with different enforced fairness level δ) on the training and test dataset – Decision tree model with max depth 10.

Dataset	Model	Train acc	Test acc	Train Δ	Test Δ
Bank (age)	Unconstrained	92.5%	87.8%	0.063	0.074
	Fair ($\delta = 0.1$)	94.7%	89.3%	0.057	0.078
	Fair ($\delta = 0.01$)	94.6%	89.3%	0.011	0.063
	Fair ($\delta = 0.001$)	94.6%	89.3%	0.001	0.066
COMPAS (race)	Unconstrained	72.4%	60.1%	0.097	0.133
	Fair ($\delta = 0.1$)	79.4%	64.3%	0.108	0.131
	Fair ($\delta = 0.01$)	79%	64%	0.017	0.073
	Fair ($\delta = 0.001$)	78.7%	64%	0.002	0.067
COMPAS (gender)	Unconstrained	72.4%	60.2%	0.117	0.107
	Fair ($\delta = 0.1$)	79.3%	64.4%	0.081	0.1
	Fair ($\delta = 0.01$)	78.6%	64.4%	0.013	0.083
	Fair ($\delta = 0.001$)	78.5%	64.3%	0.001	0.08
Law (race)	Unconstrained	95.9%	92.2%	0.236	0.165
	Fair ($\delta = 0.1$)	97.6%	93.6%	0.148	0.156
	Fair ($\delta = 0.01$)	97.5%	93.4%	0.018	0.12
	Fair ($\delta = 0.001$)	97.5%	93.5%	0.002	0.112
Law (gender)	Unconstrained	95.9%	92.2%	0.035	0.039
	Fair ($\delta = 0.1$)	97.6%	93.6%	0.097	0.039
	Fair ($\delta = 0.01$)	97.6%	93.6%	0.019	0.04
	Fair ($\delta = 0.001$)	97.6%	93.6%	0.002	0.033

4.2. Experimental Results on Real Data

Data and models. We conduct experiments on the Law School dataset (Law) [27][2] (with 19 features and $16,672$ data points), Bank Marketing dataset (Bank) [28] (with 58 features and $24,391$ data points), and COMPAS dataset [29] (with 11 features and $4,302$ data points). We use the same preprocessing on these datasets as in IBM's AI Fairness 360 [35]. For COMPAS and Law datasets, we consider two versions for each dataset, one where the protected attribute is "race" (white versus non-white) and the other where the protected attribute is "gender" (male versus female). For Bank dataset, the protected attribute is "age" (age ≥ 25 versus age < 25). Table 5 shows the distribution of data points across different subgroups. For all datasets, we use 50% of the available data for training and the remaining 50% for test.

We train decision tree (DT) models, which are commonly evaluated in the fairness literature. We train fair models using the reductions approach [3]. For each experiment, we report the average results over 20 runs. Table 6 shows the performance of the unconstrained and fair models. We defer readers to Table 10 in Appendix A for the detailed results for all subgroups.

2. Downloaded from https://github.com/jjgold012/lab-project-fairness (Bechavod and Ligett, 2017)

TABLE 7: Privacy risk of unconstrained and fair models (with $\delta = 0.001$) across different subgroups – Decision tree models with max depth 10. We indicate the protected attribute for each dataset. Unprivileged subgroups are identified by asterisks.

Dataset	Model	G_0^-	G_1^-	G_0^+	G_1^+
Bank (age)	Unconstrained	0.545	0.516	0.645	0.611*
	Fair	0.574	0.521	0.707	0.644
COMPAS (race)	Unconstrained	0.582	0.565	0.579	0.611*
	Fair	0.599	0.589	0.601	0.648
COMPAS (gender)	Unconstrained	0.583	0.569*	0.643	0.576
	Fair	0.572	0.591	0.643	0.599
Law (race)	Unconstrained	0.726	0.711*	0.541	0.510
	Fair	0.745	0.818	0.555	0.515
Law (gender)	Unconstrained	0.721	0.724*	0.514	0.514
	Fair	0.774	0.788	0.521	0.519

Privacy cost of fairness. Table 7 compares the privacy risk on fair and unconstrained models for different subgroups. After imposing the fairness constraint, the gap between the privacy risk across different subgroups widens. For instance, in the experiment on the Bank dataset, the difference between the privacy risk across subgroups G_1^- and G_0^+ increases from 0.131 (on the unconstrained model) to 0.186 (on the fair model). This shows the privacy risk disparity due to fairness, in the real data, similar to what we observe on synthetic data.

One interesting observation is that the privacy cost on the COMPAS (gender) dataset is relatively smaller than that of the COMPAS (race) dataset. Note that the two datasets are exactly the same. The difference is that the fairness gap of the unconstrained model with respect to gender is 0.095, which is much smaller than the fairness gap of the same model with respect to race. Thus, the fairness constraint has less impact on the model that is trained on COMPAS (gender). This results in less memorization, hence smaller privacy cost.

On the Law (gender) dataset, we observe a different phenomenon. Even though the fairness gap of the unconstrained model is small (0.031), we observe a high subgroup privacy cost (0.064) on G_1^-. We think the reason might be that the fair model is unable to learn generalizable patterns on this subgroup due to its relatively small size (as shown in Table 5); hence, it memorizes the data. Table 8 shows the prediction performance and subgroup privacy risk for the same dataset with a different protected attribute (Law (race) dataset). By increasing the enforced level, we can increase the accuracy of the unprivileged subgroup, yet at the cost of its privacy.

Effect of model complexity on privacy cost of fairness. Table 9 shows the privacy risk across different subgroups, when we change the complexity of the decision tree models (by controlling their maximum depth). When the model has a lower complexity, i.e., the max depth of the decision tree model is low, the privacy cost is negligible. This is because the fair models have a low accuracy even on the training dataset. For the decision tree models with max depth 5, after imposing fairness constraints, the test accuracy drops from 33.4% to 13.3% on subgroup G_0^-,

TABLE 8: Prediction accuracy and privacy risk of unconstrained and fair models with different enforced fairness gap δ on decision tree models with max depth 10 – Law (race) dataset.

	Model	G_0^-	G_1^{-*}	G_0^+	G_1^+
Train acc	Unconstrained	70.1%	43.4%	99.0%	99.8%
	Fair ($\delta = 0.1$)	64.7%	49.7%	99.3%	99.8%
	Fair ($\delta = 0.01$)	55.6%	53.8%	99.6%	99.8%
	Fair ($\delta = 0.001$)	54.5%	54.3%	99.6%	99.7%
Test acc	Unconstrained	28.6%	10.8%	92.2%	98.6%
	Fair ($\delta = 0.1$)	26.8%	10.9%	92.8%	98.4%
	Fair ($\delta = 0.01$)	23.7%	10.9%	93.5%	98.2%
	Fair ($\delta = 0.001$)	23.3%	11.8%	94.0%	98.1%
Privacy risk	Unconstrained	0.726	0.711	0.541	0.510
	Fair ($\delta = 0.1$)	0.744	0.777	0.550	0.513
	Fair ($\delta = 0.01$)	0.743	0.810	0.553	0.514
	Fair ($\delta = 0.001$)	0.745	0.818	0.555	0.515

TABLE 9: Privacy risk of unconstrained and fair models on Law (race) dataset (with $\delta = 0.001$) – Decision tree models. The "DT-x" row shows the results on decision tree models with max depth x.

Model type	Model	G_0^-	G_1^{-*}	G_0^+	G_1^+
DT-5	Unconstrained	0.585	0.561	0.515	0.503
	Fair	0.574	0.583	0.517	0.504
DT-10	Unconstrained	0.726	0.711	0.541	0.510
	Fair	0.745	0.818	0.555	0.515
DT -15	Unconstrained	0.815	0.874	0.557	0.516
	Fair	0.879	0.955	0.589	0.527

and increases from 9.2% only to 13.1% on G_1^-. Note that the random guessing accuracy is 50%. Therefore, when the model has a lower capacity, the model cannot fit the data well and the privacy risk is low, for both fair and unconstrained model.

5. Related Work

5.1. Algorithmic Fairness

Various notions of algorithmic fairness are studied in the literature. These include metric equality across sensitive groups [4, 6], individual fairness [5], and causality-based measures [8]. For training models that satisfy these fairness definitions, many techniques are proposed in the recent years, which include pre-processing methods [9, 12], in-processing methods [3, 7, 10, 11, 36], and post-processing methods [6]. In the pre-processing methods, the goal is to find a new representation of data to retain information of input features about the learning task while scrapping the information that can lead to bias. In-processing methods enforce fairness during the training process, by incorporating the fairness measures into the objective function or by reducing the constrained optimization problems to a sequence of cost-sensitive classification problems. Post-processing methods correct a given model's predictions to satisfy fairness criteria.

5.2. Membership Inference Attacks

In the machine learning context, the membership inference attacks aim to determine whether a given data point has been part of a model's training set [15, 17, 19, 37]. Membership inference attacks are used to measure the information leakage of machine learning algorithms about the individual data records in their training set. This approach is used on classification models [17], adversarially robust learning algorithms [16], model explanations [14], embedding algorithms [13] and reinforcement learning algorithms [32].

5.3. Privacy and Fairness

In decision-making processes where fairness is a pressing need, the training dataset typically contains sensitive information about individuals (e.g., in the case of loan approval application, health condition assessment, and the recidivism prediction which we describe in Section 1).

Shokri et al. [17] and Yaghini et al. [38] demonstrate the information leakage of classification models about their training data varies across different classes and groups. Imposing fairness constraints during the training does not eliminate the disparity of vulnerability. The follow-up work shows that the privacy risk of some data records can be notably high even when the average privacy risk is low [39]. A recent interesting work studies the effect of the balance of the dataset on the privacy risk of the standard (unconstrained) models [40]. These results are consistent with the findings we have on unconstrained models. However, in this paper, we focus mainly on the effect of fairness constraints on the privacy risk of subgroups and individuals, instead of only analyzing the disparity of privacy loss across the population.

Dwork et al. [5] explore the relationship between fairness in machine learning and differential privacy. The authors point out that differential privacy tools can be adopted for satisfying fairness constraints. Later on, Ekstrand et al. [41] pose multiple questions regarding the relation between fairness and privacy. Our results provide answers to one of the questions, and show that fairness could reduce the privacy of its subjects.

Kuppam et al. [42] show that resource allocation, based on differentially private statistics, can disproportionately affect some subgroups. In the machine learning context, Bagdasaryan et al. [24] study the impact of differential privacy on the prediction accuracy on subgroups, and demonstrate that if the original (unconstrained) model is unfair, the unfairness becomes worse once privacy-preserving algorithms (i.e., DP-SGD [20]) are applied.

5.4. Fair Learning with Differential Privacy

A rigorous framework to protect privacy is to use differentially private training algorithms to bound the model's information leakage about the members of its training set. For a given false positive rate, the true positive rate of the membership inference attacks is upper-bounded in differentially private models (as also shown in [43, Proposition-1]). Thus, the privacy risk (in Definition 2) is upper bounded by $(e^\epsilon + \delta)/2$ when the model is (ϵ, δ)-differentially private (for all the individuals regardless of their subgroups). Several works propose algorithms to learn a model that satisfies differential privacy and fairness (including demographic parity [25, 26], equality of opportunity [21], and equalized odds [44]). Recently, Tran et al.

[45] introduce a differential privacy framework to train deep learning models that satisfy several group fairness notions, including equalized odds, accuracy parity, and demographic parity.

It is important to highlight that the differential privacy considered in [44, 45] is to protect privacy with respect to the sensitive attribute instead of the membership. Jagielski et al. [44] present two algorithms that can achieve fairness and differential privacy with respect to the protected attribute. The major observation, however, is that model accuracy significantly drops.

On the contrary, we analyze the privacy risks of individual data records and subgroups when enforcing fairness constraints. Cummings et al. [21] present the most related theoretical results by showing the impossibility of achieving pure differential privacy and exact (equality of opportunity) fairness.

6. Conclusions

In this paper, we have presented a simple yet effective framework for analyzing privacy risks of group fairness algorithms for machine learning. We have shown that fair algorithms tend to memorize data from the under-represented subgroups, while trying to equalize the model's error across groups (partitioned based on their protected attribute). This memorization leads to an increase in the model's information leakage about unprivileged groups. We have provided comprehensive evaluations (using membership inference attacks) on synthetic data, as well as real data, to show how and why fair models leak information about their training data.

Acknowledgments

The authors would like to thank Ergute Bao, Anmin Kang, Sasi Kumar Murakonda, Ta Duy Nguyen, and Martin Strobel for helpful discussions and their feedback. This work is supported in part by the Singapore Ministry of Education Academic Research Fund, R-252-000-660-133, the NUS Early Career Research Award (NUS ECRA), grant number NUS ECRAFY19 P16, the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative, and Intel within the www.private-ai.org center.

References

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, 2018, pp. 77–91.

[2] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," *arXiv preprint arXiv:1803.02453*, 2018.

[4] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 2009, pp. 13–18.

[5] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[6] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[7] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 643–650.

[8] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in neural information processing systems*, 2017, pp. 4066–4076.

[9] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," *arXiv preprint arXiv:1802.06309*, 2018.

[10] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *arXiv preprint arXiv:1507.05259*, 2015.

[11] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 962–970.

[12] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[13] C. Song and A. Raghunathan, "Information leakage in embedding models," *arXiv preprint arXiv:2004.00053*, 2020.

[14] R. Shokri, M. Strobel, and Y. Zick, "Privacy risks of explaining machine learning models," *arXiv preprint arXiv:1907.00164*, 2019.

[15] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 268–282.

[16] L. Song, R. Shokri, and P. Mittal, "Privacy risks of securing machine learning models against adversarial examples," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 241–257.

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

[18] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *International Conference on Machine Learning*, 2019, pp. 5558–5567.

[19] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 739–753.

[20] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[21] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern, "On the compatibility of privacy and fairness," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 2019, pp. 309–315.

[22] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011.

[23] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable private learning with pate," *arXiv preprint arXiv:1802.08908*, 2018.

[24] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 479–15 488.

[25] J. Ding, X. Zhang, X. Li, J. Wang, R. Yu, and M. Pan, "Differentially private and fair classification via calibrated functional mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 622–629.

[26] D. Xu, S. Yuan, and X. Wu, "Achieving differential privacy and fairness in logistic regression," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 594–599.

[27] L. F. Wightman and H. Ramsey, "Law school admission council." 1998.

[28] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[29] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "COMPAS dataset," https://github.com/propublica/compas-analysis, 2017, [COMPAS dataset (2017)].

[30] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Advances in Neural Information Processing Systems*, 2018, pp. 2791–2801.

[31] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[32] X. Pan, W. Wang, X. Zhang, B. Li, J. Yi, and D. Song, "How you act tells a lot: Privacy-leaking attack on deep reinforcement learning," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019, pp. 368–376.

[33] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification." *J. Mach. Learn. Res.*, vol. 20, no. 75, pp. 1–42, 2019.

[34] V. Feldman, "Does learning require memorization? a short tale about a long tail," in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020, pp. 954–959.

[35] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[36] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[37] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[38] M. Yaghini, B. Kulynych, and C. Troncoso, "Disparate vulnerability: On the unfairness of privacy attacks against machine learning," *arXiv preprint arXiv:1906.00389*, 2019.

[39] Y. Long, L. Wang, D. Bu, V. Bindschaedler, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "A pragmatic approach to membership inferences on machine learning models."

[40] S. M. Tonni, D. Vatsalan, F. Farokhi, D. Kaafar, Z. Lu, and G. Tangari, "Data and model dependencies of membership inference attack," *arXiv preprint arXiv:2002.06856*, 2020.

[41] M. D. Ekstrand, R. Joshaghani, and H. Mehrpouyan, "Privacy for all: Ensuring fair and equitable privacy protections," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 35–47.

[42] S. Kuppam, R. McKenna, D. Pujol, M. Hay, A. Machanavajjhala, and G. Miklau, "Fair decision making using privacy-protected data," *arXiv preprint arXiv:1905.12744*, 2019.

[43] Ú. Erlingsson, I. Mironov, A. Raghunathan, and S. Song, "That which we call private," *arXiv preprint arXiv:1908.03566*, 2019.

[44] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman, "Differentially private fair learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3000–3008.

[45] C. Tran, F. Fioretto, and P. Van Hentenryck, "Differentially private and fair deep learning: A lagrangian dual approach," *arXiv preprint arXiv:2009.12562*, 2020.

Appendix A.
Additional Results

Table 10 shows the prediction accuracy of unconstrained and fair models on all subgroups for decision tree models with max depth 10.

TABLE 10: Prediction accuracy, decision tree (max depth 10).

Dataset	Model		G_0^-	G_1^-	G_0^+	G_1^+
Bank (age)	Unconstrained	Train acc	95.7%	97.7%	79.8%	74.3%
		Test acc	88.5%	94.5%	56.1%	53.3%
	Fair ($\delta = 0.001$)	Train acc	97.3%	97.3%	75.3%	75.2%
		Test acc	89.6%	94.4%	51.1%	54.6%
COMPAS (race)	Unconstrained	Train acc	81.9%	92.3%	74.2%	61.4%
		Test acc	67.8%	81.1%	60.0%	42.2%
	Fair ($\delta = 0.001$)	Train acc	85.6%	85.7%	70.8%	70.8%
		Test acc	70.9%	73.4%	56.8%	50.3%
COMPAS (gender)	Unconstrained	Train acc	92.0%	82.5%	72.7%	71.6%
		Test acc	79.2%	70.6%	47.0%	57.8%
	Fair ($\delta = 0.001$)	Train acc	84.6%	84.4%	71.6%	71.5%
		Test acc	73.9%	70.8%	49.9%	57.2%
Law (race)	Unconstrained	Train acc	70.1%	43.4%	99.0%	99.8%
		Test acc	28.6%	10.8%	92.2%	98.6%
	Fair ($\delta = 0.001$)	Train acc	54.5%	54.3%	99.6%	99.7%
		Test acc	23.3%	11.8%	94.0%	98.1%
Law (gender)	Unconstrained	Train acc	57.3%	54.2%	99.6%	99.7%
		Test acc	20.8%	18.0%	97.4%	97.8%
	Fair ($\delta = 0.001$)	Train acc	57.2%	57.1%	99.8%	99.7%
		Test acc	20.7%	17.8%	97.4%	97.7%