Understanding Disparate Effects of Membership Inference Attacks and Their Countermeasures

Da Zhong[†] Haipei Sun[†] Jun Xu[†] Neil Gong[‡] Wendy Hui Wang[†]
[†]Stevens Institute of Technology [‡]Duke University

ABSTRACT

Machine learning algorithms, when applied to sensitive data, can pose severe threats to privacy. A growing body of prior work has demonstrated that membership inference attack (MIA) can disclose whether specific private data samples are present in the training data to an attacker. However, most existing studies on MIA focus on *aggregated* privacy leakage for an entire population, while leaving privacy leakage across different demographic subgroups (e.g., females and males) in the population largely unexplored. This raises two important issues: (1) *privacy unfairness* (i.e., if some subgroups are more vulnerable to MIAs than the others); and (2) *defense unfairness* (i.e., if the defense mechanisms provide more protection to some particular subgroups than the others).

In this paper, we investigate both privacy unfairness and defense fairness. We formalize a new notation of *privacy-leakage disparity* (PLD), which quantifies the disparate privacy leakage of machine learning models to MIA across different subgroups. In terms of privacy unfairness, our empirical analysis of PLD on real-world datasets shows that privacy unfairness exists. The minority subgroups (i.e., the less represented subgroups) tend to have higher privacy leakage. We analyze how subgroup size and subgroup data distribution impact PLD through the lens of model memorization. In terms of defense unfairness, our empirical evaluation shows the existence of unfairness of three state-of-the-art defenses, namely differential privacy, L_2-regularizer, and Dropout, against MIA. However, defense unfairness mitigates privacy unfairness as the minority subgroups receive stronger protection than the others. We analyze how the three defense mechanisms affect subgroup data distribution disparately and thus leads to defense unfairness.

CCS CONCEPTS

• **Security and privacy**;

KEYWORDS

Membership inference attack, privacy leakage, disparity, fairness

ACM Reference Format:
Da Zhong[†] Haipei Sun[†] Jun Xu[†] Neil Gong[‡] Wendy Hui Wang[†], [†]Stevens Institute of Technology [‡]Duke University, . 2022. Understanding Disparate Effects of Membership Inference Attacks and Their Countermeasures. In *Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22), May 30–June 3, 2022, Nagasaki, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3488932.3501279

ASIA CCS '22, May 30–June 3, 2022, Nagasaki, Japan.

ACM ISBN 978-1-4503-9140-5/22/05…$15.00
https://doi.org/10.1145/3488932.3501279

1 INTRODUCTION

Advances in the field of machine learning (ML) have resulted in algorithms and technologies for improving cybersecurity by helping identify security threats and system vulnerabilities. However, ML is also vulnerable to novel and sophisticated privacy attacks that leak information about the training dataset [25, 52]. For instance, in a *membership inference attack (MIA)* [30, 50, 52], an attacker can infer whether a given data sample is in an ML model's training dataset, even if the attacker only has black-box access to the model's prediction APIs. When the training data is sensitive or proprietary user data such as electronic health records, location/contact traces, and financial information, such leakage poses severe threats to user privacy. Meanwhile, quite a few defense mechanisms (e.g., [34, 45, 49, 52]) have been proposed to mitigate the threat of MIA.

Despite the active research on MIA [46, 49, 52, 54, 56, 58] and their defense mechanisms [34, 45, 49, 52], most of these existing studies only focus on *aggregated* privacy leakage of MIA over an entire population. However, the effect of MIA may differ across various demographic groups (e.g., females vs. males). For instance, MIA may be able to correctly infer 80% of females and 60% of male members, respectively. It is even possible that the privacy risk of some subgroups can be notably high when the privacy risk of the whole population is low. Such disparate privacy leakage across different demographic groups may raise the serious concern of "**privacy unfairness**", i.e., an ML model can pose different privacy risks to users in different demographic groups. The fairness concern also holds on the defense mechanisms against MIA, as they may affect the subgroups differently (e.g., the female group receives more privacy protection than the male group). This imposes the concern of "**defense unfairness**", i.e., a defense mechanism does not provide equitable protection across different groups.

The research community has mostly considered fairness and privacy as two equally important issues and investigated them separately. Few recent works [13, 57] started looking at fairness and privacy concurrently. Chang et al. [13] explore the privacy risks of imposing fairness constraints on ML models. Their angle is orthogonal to privacy unfairness or defense unfairness. Yaghini et al. [57] indeed study privacy unfairness and defense unfairness, but to a very preliminary extent. They show the existence of disparate privacy leakage to MIA and discuss the factors affecting disparate privacy leakage. They further inspect the impact of differential privacy (DP) [19], as an MIA defense, on the disparity. However, their study neither unveils the fundamental causes of the disparity nor investigates how and why DP affects the disparity.

In this paper, we aim to study both privacy unfairness and defense unfairness, centering around three research questions:

- **RQ1.** *Does privacy unfairness exist and how significant it can be?*
- **RQ2.** *What are the factors leading to privacy unfairness?*
- **RQ3.** *Can defense mechanisms present disparate privacy leakage of different groups and why?*

Defining metrics: There have been no systematic schemes to measure privacy unfairness in the literature. To this end, we first formalize a new notation called *privacy-leakage disparity* (PLD) to quantify privacy unfairness in the context of MIA. PLD is adapted from a well-accepted fairness notation named *accuracy parity* [6, 8, 15]. At a high level, PLD measures the gap between the MIA accuracy, which reflects the amount of privacy leakage, of different demographic groups.

Studying privacy unfairness: Leveraging PLD, we run an empirical study to measure privacy unfairness, on three real-world datasets widely adopted by both fairness and privacy communities (see Table 2). The results show the prevalent existence of PLD, and interestingly, the minority subgroup (i.e., the less represented subgroup in the data) tends to have higher privacy leakage. Using model memorization, which measures how much the model memorizes data samples, as a tool, we further gain an understanding of two major causes of PLD: (i) *subgroup size* - the model memorizes more about the subgroup with a smaller size and (ii) *subgroup data distribution* - the two subgroups that have more deviated distribution have higher PLD.

Studying defense unfairness: Aiming to understand defense unfairness, we finally inspect the impacts of MIA defenses on PLD. We consider three state-of-the-art MIA defenses, including differential privacy [20], L_2-regularizer [52] , and Dropout [49]. It turns out that these MIA defenses affect PLD in a "positive" manner, namely they tend to reduce the extent of PLD. That concurrently unveils the existence of "defense unfairness": the MIA defenses are inclined to provide stronger protection for the minority subgroup. We show the reason behind "defense unfairness" is that MIA defenses have a disparate impact on reducing the distribution deviation between different subgroups and the entire population. In particular, they reduce the distribution deviation between the minority subgroup and the entire population much quicker than the majority subgroup, which leads to stronger defense against privacy leakage.

In summary, the paper makes the following contributions.

- We formalize the new notation of privacy-leakage disparity to quantify privacy unfairness.
- We show the existence of privacy-leakage disparity in MIA and unveil the main underlying causes.
- We demonstrate that existing MIA defenses have unfair protection, leading to a reduction of privacy-leakage disparity.

2 MEMBERSHIP INFERENCE ATTACKS

This paper aims to investigate the disparity of privacy leakage of ML models. An essential component is a proper method to measure the privacy leakage. We consider Membership Inference Attack (MIA) [52] as the method because of its wide acceptance [28, 30, 42, 44, 46, 49–51, 54, 56] and practical influence [33]. The rest of this section covers the background of MIA. Notations to facilitate our description are summarized in Table 1.

Table 1: Common notations used in the paper

Notation	Description
$\mathcal{T}$	Target model
$\mathcal{M}$	Membership inference attack (MIA) model
$y \in \{+,-\}$	Label of the target model $\mathcal{T}$
$b \in \{+,-\}$	Label of the attack model $\mathcal{M}$. $b = +/-$: member/non-member
$D^{\mathcal{T}}_{train}$, $D^{\mathcal{T}}_{test}$	Training and testing datasets of target model $\mathcal{T}$
$D^{\mathcal{M}}_{train}$, $D^{\mathcal{M}}_{test}$	Training and testing datasets of attack model $\mathcal{M}$
A	Protected attribute. $A = a$ ($a \in \{0, 1\}$) specifies the value of the protected attribute.
G_a	Group G_1/G_0
G^y_a	Set of data points with protected attribute a and target label y

2.1 Target Model

Given an ML model $\mathcal{T}$, which we call *target model*, MIA aims to infer the *membership* of a given data sample, i.e., whether the data sample is in the model's training dataset $D^{\mathcal{T}}_{train}$. Similar to previous research [52], we consider target models that are classification models with the following configurations. Given a training dataset $D^{\mathcal{T}}_{train}$ of domain $X \times Y$, where X denotes the input features and Y denotes the output label, the classification model $\mathcal{T}$ is trained on $D^{\mathcal{T}}_{train}$ based on the ground truth of Y. For a testing dataset $D^{\mathcal{T}}_{test}$ which follows the same distribution of $D^{\mathcal{T}}_{train}$, $\mathcal{T}$ outputs a *confidence score vector* (CSV) for each data sample in $D^{\mathcal{T}}_{test}$. The CSV is a probability distribution over the class labels of Y, and the label of the highest CSV is deemed the prediction. For simplicity, we only consider binary classification models (i.e., $y \in \{-,+\}$), but the methodologies and principles of our study apply to all types of classification models.

2.2 Attack Model

In an MIA attack, the attacker trains another ML model $\mathcal{M}$ called the *attack model*. $\mathcal{M}$ takes some input features X^{MIA} produced by $\mathcal{T}$ as an input, and predicts the membership for any given data sample. Formally, $\mathcal{M}$ can be considered as a binary classification model which predicts the label b for any data sample (x, y), where $b = +$ if $\mathcal{M}$ predicts $(x, y) \in D^{\mathcal{T}}_{train}$, otherwise $b = -$.

In the black-box setting [52], an attacker has access to the CSVs output by $\mathcal{T}$ for each data sample, through channels like prediction APIs of the target model. When the attack has access to the ground truth of some members and non-members of $D^{\mathcal{T}}_{train}$. The attacker generates the training dataset $D^{\mathcal{M}}_{train}$ of the attack model $\mathcal{M}$, which includes the CSVs of all members/non-members in the ground truth as features X^{MIA} and their member/non-member status as labels. The attacker then trains $\mathcal{M}$ on $D^{\mathcal{M}}_{train}$. When the attacker has no access to the ground truth of members/non-members in $D^{\mathcal{T}}_{train}$, the attack model can be learnt using the following shadow model approach [52].

Technically, the attacker first synthesizes data samples to mirror the training samples of $\mathcal{T}$. One way is to initialize a random sample and gradually improve its quality using output of the target model [52]. Based on the synthesized data, the attacker creates a

group of shadow models to approximate the target model. Each shadow model is trained using some synthesized data samples and the outputs of those samples predicted by the target model. The attacker eventually deems the shadow models as the target model and trains the attack model as described above.

Other works also considered the white-box setting [39, 46], where the attacker is assumed to have access to the parameters of $\mathcal{T}$. In this setting, the attack model can further use features extracted from the model parameters (e.g., the gradients of the prediction with respect to a data sample [46]).

In this work, we focus on the black-box setting for a consideration of its generality. Since our goal is to observe the disparity of privacy leakage, acquiring a more evident level of privacy leakage is obviously beneficial. To this end, we take two actions. First, we assume some ground truth members/non-members are available to train the target model, avoiding using the the shadow models that can be noisy. Second, we follow [52] to train a separate attack model for each label of $\mathcal{T}$.

2.3 Quantifying Privacy Leakage by Membership Inference Attacks

MIA effectiveness is measured by standard metrics [49, 52] including *accuracy*, *precision*, and *recall*. Given an *attack testing dataset* $D^{\mathcal{M}}_{test}$, which consists of members and non-members of the target model's training dataset. The attacker uses the attack model $\mathcal{M}$ to predict memberships of data samples in $D^{\mathcal{M}}_{test}$. Specifically, *accuracy* is the fraction of the data correctly predicted as member or non-member, *precision* is the fraction of the predicted members that are true members of $D^{\mathcal{T}}_{train}$, and *recall* is the fraction of the true members in $D^{\mathcal{M}}_{test}$ that are predicted as members.

3 DEFINING PRIVACY-LEAKAGE DISPARITY

Using MIA, we have a method to measure the privacy leakage experienced by the entire population, but we still lack schemes to assess the disparity of the privacy leakage. In fact, privacy-leakage disparity (PLD) has not been systematically defined in the literature. In this paper, we aim to take the initial step towards establishing the definition of PLD, with a focus at the group level (i.e., how to quantify the difference in the privacy leakage of different groups). Our insights are borrowed from Algorithm Fairness in ML, an area attracting tremendous attention in recent years. In the rest of this section, we first briefly introduce algorithm fairness in ML and then explain how we adapt that to PLD.

3.1 Algorithmic Fairness in Machine Learning

Roughly speaking, given a dataset D of domain $X \times Y$ where X denotes the input features and Y denotes the output label, each sample in D is associated with a set of *protected attributes* $A \subseteq X$ (e.g., gender, race). For simplicity, we consider only one protected attribute in this paper. Depending on the value of the protected attribute, the data samples in D are divided into two groups: *protected group* (denoted by $A = 1$) and *unprotected group* $A = 0$. For instance, consider *gender* as the protected attribute. The whole population can be grouped into the female and male groups, where the female group is considered as the protected group.

Based on the definition of the protected attribute and the corresponding groups, an ML system satisfies *group fairness* if its predicted outcomes are similar across different groups. The fairness community has proposed many mathematical notations to formalize the similar treatment [10, 17, 27, 37]. For example, *equal opportunity* [27] requires the same true positive rate across different groups, while *equalized odds* [27] requires the same true and false positive rates across different groups [27]. More recent research also brings up the concept of *accuracy parity* [6, 8, 15], observing that ML systems often exhibit substantial accuracy disparities among different demographic groups. Formally, accuracy parity is defined as follows (formalism of other fairness metrics is omitted because we adapt accuracy parity to define PLD):

Definition 1 (Accuracy Parity [6, 8, 15]). *Given a prediction model h, and a pre-defined accuracy metric ACC that measures the accuracy of the prediction output made by h, let $ACC_a = ACC(h, G_a)$ be the prediction accuracy of the group G_a ($a \in \{0, 1\}$). Then h satisfies* ***accuracy parity*** *if $ACC_0 = ACC_1$.*

The violation of accuracy parity is known as *disparate mistreatment* [59]. To measure the level of disparate mistreatment, we can use *accuracy gap* defined as follows:

Definition 2 (Accuracy gap). *Given a prediction model h and two groups G_0 and G_1, the* accuracy gap *of h on these two groups is $\Delta := |ACC_0 - ACC_1|$. By definition, if $\Delta(h)$ satisfies accuracy parity, $\Delta(h)$ will be zero.*

3.2 From Algorithmic Fairness to Privacy-Leakage Disparity

To align with the working principles of MIAs, we mainly consider *subgroups*, instead of protected/unprotected groups. In particular, since we follow the state-of-the-art MIA model [52] to train a separate attack model for each target label, we further split each group (G_0 or G_1) into ℓ subgroups, where ℓ is the number of unique labels of the target model. For example, consider a binary classification task that predicts whether an individual has annual income greater than \$50K based on their demographic information, the group G_1= *Female* is split into two subgroups: G_1^+ for females who are labeled with annual income greater than \$50K, and G_1^- for the remaining females. In the following discussions, we use G_a to indicate the group with $A = a$, where $a \in \{0, 1\}$. We use G_a^y to present the subgroup of G_a with label y, where $y \in \{-, +\}$. Given two subgroups G_a^y and $G_{a'}^y$ that have the same target label y, we call G_a^y the *minority subgroup* if $|G_a^y < G_{a'}^y|$, and $G_{a'}^y$ the *majority subgroup*.

We adapt accuracy parity to define PLD. But instead of looking at the accuracy of the ML model (i.e., target model), we consider the accuracy of MIA (i.e., the attack model): the probability that an adversary can correctly infer if a data point is a member/non-member in $D^{\mathcal{T}}_{train}$. Based on MIA accuracy, the privacy leakage of a subgroup can be measured as follows:

Definition 3 (Subgroup Privacy Leakage). *Given a dataset D, a target model $\mathcal{T}$, and the MIA model $\mathcal{M}$ that predicts the membership label b, we define the privacy leakage of the target model $\mathcal{T}$ with respect to the subgroup G_a^y (i.e., data points in D with the protected attribute a ($a \in \{0, 1\}$) and label y ($y \in \{-, +\}$) against $\mathcal{M}$ as:*

$$PL(G_a^y) = P(\hat{b} = b | Y = y, A = a).$$

Table 2: Datasets used in our study

Dataset	Size	Attributes (#)	Target labels (#)
Adult	45k	14	2
Broward	7.2k	8	2
Hospital	52.7k	16	2

Based on the definition of subgroup privacy leakage, we then define PLD to quantify the difference in the privacy leakage between two subgroups (e.g., females v.s. males). We require the two subgroups to have the same label since the existing fairness literature typically do not compare groups with different labels. Formally:

Definition 4 (Privacy-leakage Disparity). *For any two subgroups G_a^y and $G_{a'}^y$ that have the same label y, the privacy-leakage disparity between these two subgroups is measured as:*

$$PLD(G_a^y, G_{a'}^y) = |PL(G_a^y) - PL(G_{a'}^y)|.$$

In essence, PLD defined above quantifies the "accuracy gap" of $\mathcal{M}$ between two subgroups, which indicates the difference in their privacy-risk levels. Specifically, the subgroup with higher MIA accuracy has a higher risk than the other subgroup.

We note that we may also adapt other fairness metrics (e.g., equalized odds [27] and equal opportunity [27]) to define PLD. We omit doing so because our follow-up studies are less dependent on which fairness metrics we use to define PLD.

4 MEASURING PRIVACY-LEAKAGE DISPARITY

To answer research question **RQ1** (§1), we perform an empirical study to measure PLD of MIA, aiming to understand its existence and extent. Our code and datasets are shared at https://github.com/dzhong2/MIA_disparity.

4.1 Experimental Setup

4.1.1 Datasets. We use three real-world datasets that are widely adopted by both fairness and MIA communities, as listed in Table 2. We elaborate on them as follows.

Adult dataset [1] includes 45,222 instances and 14 attributes (such as age, gender, education, marital status, occupation, working hours, and native country) that describe the information about individuals from the 1994 U.S. census. The prediction task is to determine whether a person makes over $50k annually.

Broward [2] contains criminal history, jail and prison time, demographics and COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions) risk scores for defendants from Broward County, Florida. The prediction task is to infer whether a criminal defendant will be a recidivist (i.e., a criminal who re-offend) within two years.

Hospital dataset [47] is released by the Texas Department of State Health Services. It contains records of inpatient stays in some health facilities. The features include types of external causes of injury, diagnosis, the procedures the patient underwent, and demographic information such as the gender, age, and race of the patients. The classification task is to predict the patient's main procedure. We categorize the main procedures into two groups (corresponding to the prediction labels): cardiology and pulmonology.

Table 3: Performance of the target models and MIA. Acc: accuracy; Prec: precision; Rec: recall.

Dataset	Train Acc	Test Acc	MIA Acc	MIA Prec	MIA Rec
Adult	90.4%	82.5%	54.1%	52.4%	89.7%
Broward	71.3%	67.3%	51.8%	51.5%	64.4%
Hospital	91.8%	64.9%	65.1%	62.1%	77.5%

The datasets are pre-processed to remove all the samples with missing values. Further, all attributes are converted into numeric values by using one hot encoding.

4.1.2 Setup of groups/subgroups. For all datasets, we follow the literature of fairness research and consider *gender* and *race* as the *protected attributes*. For each protected attribute, we setup two groups G_0 and G_1. The final grouping results can be found in Appendix A (Table 8). Since all datasets have binary target labels, we further each group into two subgroups, based on their target labels. Thus, each dataset is split into four subgroups G_1^+, G_1^-, G_0^+, and G_0^-.

4.1.3 Training Target and Attack Models. In this paper, we consider a Neural Network + Neural Network (**NN+NN**) setting, where both the target and attack models are neural networks. Different architectures are used for the two models, as the attacker does not know the architecture of the target model in black-box MIA. More details of the two neural networks are described below.

Training the target model: For each dataset, we randomly select 50% of the samples for training and the remaining for testing. We train the neural network with Keras toolkit [3]. The neural network consists of 3 hidden layers with $\{512, 256, 128\}$ neurons at each layer and a softmax output layer. The model is trained for 200 epochs with learning rate 0.01 and batch size 800.

Training the attack model: As described in §2.2, we consider an attacker with access to some ground-truth of members/non-members of the target model's training dataset. Specifically, we pick 50% of a target model's training samples and 50% of its testing samples as the ground-truth members and non-members. We then obtain the CSVs of the ground-truth members/non-members via querying the target model and uses the CSVs to train the attack model. The attack model is a neural network also trained by Keras. The neural network consists of two hidden fully-connected layers respectively with $\{256, 128\}$ neurons and a sigmoid output layer. We use the learning rate 0.01 and batch size 64 for 400 epochs in our experiments.

Evaluation metrics of target and attack models: The training and testing accuracy of the target model $\mathcal{T}$ is defined by the probability that $\mathcal{T}$ correctly predicts the true label.

$$acc = P(\mathcal{T}(x_i) = y_i)$$

where x_i is a data sample and y_i is the true label of x_i.

In contrast, MIA accuracy is defined by the probability that the attack model $\mathcal{M}$ correctly predicts the membership label:

$$\text{MIA } acc = P(\mathcal{M}(\mathcal{T}(x_i)) = b_i)$$

where b_i (+/-) indicates whether the data point is included in the training dataset of the target model.

Similarly, MIA precision is measured by the fraction of true members that the attack model predicts to be members:

$$\text{MIA } prec = P(\mathcal{M}(\mathcal{T}(x_i)) = b_i | \mathcal{M}(\mathcal{T}(x_i)) = +)$$

Table 4: Privacy-leakage disparity and average of influence scores (AIS) of subgroups. Between two subgroups of the same label, the minority subgroup is marked with green, the higher MIA accuracy with orange, and the higher AIS with pink.

Dataset	Label y	Subgroup	Gender				Race			
			Size	MIA Accuracy	AIS	PLD	Size	MIA Accuracy	AIS	PLD
Adult	+	G_1^+	6512	51.9%	0.0015	1.3%	2654	52.9%	0.0018	0.3%
		G_0^+	10493	53.2%	0.0221		14351	52.6%	0.0005	
	–	G_1^-	833	65.4%	0.0181	7.8%	495	69.8%	0.0193	12.5%
		G_0^-	4773	57.2%	0.0017		5111	57.3%	0.0128	
Broward	+	G_1^+	453	52.6%	0.0893	1.1%	1084	51.0%	0.0202	1.4%
		G_0^+	1530	51.5%	0.0467		899	52.4%	0.1	
	–	G_1^-	249	54.0%	0.1608	2.5%	678	52.7%	0.245	1.2%
		G_0^-	1375	51.5%	0.0319		946	51.5%	0.0294	
Hospital	+	G_1^+	4849	71.2%	0.0296	2.3%	3022	72.7%	0.0437	1.3%
		G_0^+	4264	73.5%	0.0308		6092	71.4%	0.0235	
	–	G_1^-	8253	62.1%	0.014	1.5%	5051	62.6%	0.0175	1.9%
		G_0^-	9022	60.6%	0.0122		12225	60.7%	0.0025	

MIA recall is measured by the fraction of all true members identified by the attack model:

$$\text{MIA } recall = P(\mathcal{M}(\mathcal{T}(x_i)) = b_i | b_i = +)$$

Performance of target and attack models: Both training accuracy and testing accuracy of the target models are measured, respectively using the training samples and the testing samples (50% v.s. 50%). To measure MIA performance, we respectively select 20% of the target model's training samples and 20% of its testing samples as an *attack testing dataset*. This attack testing dataset do not overlap with the ground-truth members/non-members to train the attack models, and we use it to compute the accuracy, precision, and recall of MIA. Each MIA attack is repeated 25 times and the average results are gathered in Table 3.

Overall, MIA presents varied performance against different datasets. On Adult and Hospital, MIA accuracy reaches 89.7% and 77.5 %, respectively. This creates an environment of high MIA accuracy to study PLD. In contrast, MIA on the Broward dataset has an accuracy of 51.8%, which gets closer to random guess. Such a result, consistent with what is reported by the literature [52], shows that MIA has a weaker effectiveness with this dataset. We envision this does not hurt our study, but instead, benefits our study since it provides an extreme context — low MIA accuracy — of understanding PLD.

Our results are also consistent with the common understanding that MIA is attributable to over-fitting of the target model. Considering the gap between training accuracy and testing accuracy as the over-fitting metric, all our target models present a certain level of over-fitting, matching the effectiveness of MIA. Moreover, larger over-fitting leads to more effective MIA. For instance, Hospital dataset presents the largest gap between training accuracy and testing accuracy, and thus, shows the highest MIA accuracy.

4.2 Results of Privacy-Leakage Disparity

For each dataset, we measure the size and MIA accuracy of all four subgroups (G_1^+, G_1^-, G_0^+, G_0^-). Table 4 shows the results, and we summarize the major observations as follows.

PLD exists: Although all subgroups show privacy leakage to MIA to a certain extent (MIA accuracy changes from 51% to 73.5%) in each setting, different subgroups experience different amounts of privacy leakage. Consider Adult dataset with Race as the protected attribute as an example. MIA accuracy varies from 52.9% to 69.8% across the four subgroups. Such MIA accuracy disparity is observed across all the three datasets and their subgroups. We also observe such disparity in MIA precision and recall across different subgroups. The results can be found in Appendix (Table 10).

Second, MIA accuracy of subgroups can be significantly different from that of the whole population. Indeed, even when MIA accuracy over the whole population is low (close to 50%), MIA accuracy of subgroups still can be notably high. For example, MIA accuracy of the whole population of Adult dataset is 54.1% (Table 3), while MIA accuracy of the subgroup G_1^- is as high as 69.8% (Table 4).

Furthermore, *the amounts of PLD (Def. 4) vary across different settings*. For example, the PLD between the two subgroups G_1^+ and G_0^+ is around 1.2% on Broward dataset with Race as the protected attribute, but it jumps to 7.8% between the same two subgroups for Adult dataset with Gender as the protected attribute. Even for the same dataset and same protected attribute, the PLD can vary significantly for subgroups with different labels. In Adult dataset with Race as the protected attribute, for example, the PLD between G_1^+ and G_0^+ is 0.3%, which increases to 12.5% between G_1^- and G_0^-. More details about the factors influencing PLD are discussed in §5.

The minority subgroup has higher privacy leakage: Another important observation is that, out of the two subgroups of the same label, the minority one has larger MIA accuracy in most of the settings. Furthermore, the group size is not inversely proportional to MIA accuracy. Consider Broward dataset with Gender as the protected attribute as an example, the size of G_1^+ is almost one third of that of G_0^+. However, MIA accuracy of G_1^+ is close to that of G_0^+. The only exception happens to G_1^+ and G_0^+ on Adult dataset with Gender as the protected attribute. The minority subgroup G_1^+ has smaller MIA accuracy than that of G_0^+, although its size is smaller than G_0^+. This observation is consistent with the results reported

Table 5: **Model overfitting disparity (OD) *v.s.* PLD.**

Dataset	Label y	Gender		Race	
		OD	PLD	OD	PLD
Adult	+	3.1%	1.3%	1.0%	0.3%
	-	15.4%	7.8%	24.8%	12.5%
Broward	+	3.7%	1.1%	4.4%	1.4%
	-	5.0%	2.5%	0.7%	1.2%
Hospital	+	4.9%	2.3%	3.2%	1.3%
	-	3.1%	1.5%	3.9%	1.9%

by prior works [13, 57]. Briefly, the major reason is that the target model "memorizes" more of the minority subgroup than the other subgroups. A deeper analysis is given in §5.

Higher overfitting leads to higher PLD: A widely recognized reason of MIA is overfitting. As we have pointed out in §4.1, more overfitting leads to higher MIA accuracy. This brings us a hypothesis that PLD is correlated with the disparity in model overfitting of different subgroups. To verify the hypothesis, we separately measured the overfitting of each subgroup and calculated the disparity between different subgroups. In particular, the overfitting of a subgroup is measured as the difference between training and testing accuracy of the subgroup. And the *overfitting disparity* is measured as the difference between the overfitting of protected and unprotected groups. As summarized in Table 5, the disparity in overfitting exists whenever PLD appears. More importantly, higher disparity in overfitting leads to more evident PLD across all settings. This empirically validates our hypothesis.

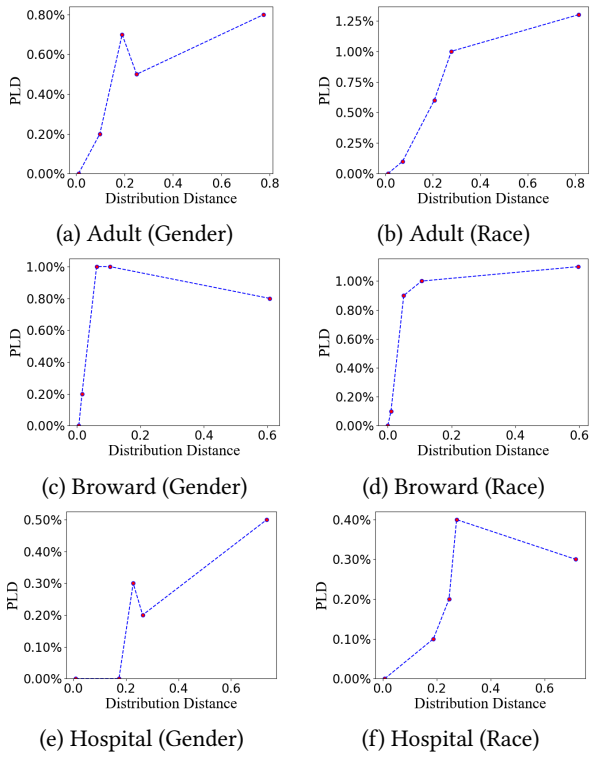
5 ANALYSIS OF PLD

Our empirical study unveils the existence of PLD, but does not give an explanation. In this section, we aim to answer research question **RQ2** (§1) and explore why PLD arises. In principle, PLD should be rooted from differences between subgroups. This gives us a direction of explaining the existence of PLD: *what types of difference lead to PLD and why*?

In a broad sense, subgroups mainly differ in their values of the protected attribute, their size, and their data distribution. We believe the values of the protected attribute are less critical. Running permutation feature importance [4] on the target models presented in §4, we obverse that the protected attribute (gender or race) is less important to the target models, as summarized in Table 9. As such, the value of the protected attribute will unlikely affect the output of the target models as well as the privacy they can leak. Thus in the rest of this section, we focus on discussing the impact of difference in subgroup size and data distribution.

5.1 Impact of Subgroup Size on PLD

As we pointed out in §4, the minority subgroup tends to have a higher privacy leakage, indicating a connection between subgroup size and PLD. To demystify the connection, we leverage the concept of *model memorization*. According to the existing literature [11, 40], a root cause of privacy leakage of an ML model to MIA is that the model *memorizes* too much information of the training data. Our hypothesis is that the model memorizes more about the smaller subgroup, which bridges the gap between the subgroup



(a) Adult (Gender) (b) Adult (Race)

(c) Broward (Gender) (d) Broward (Race)

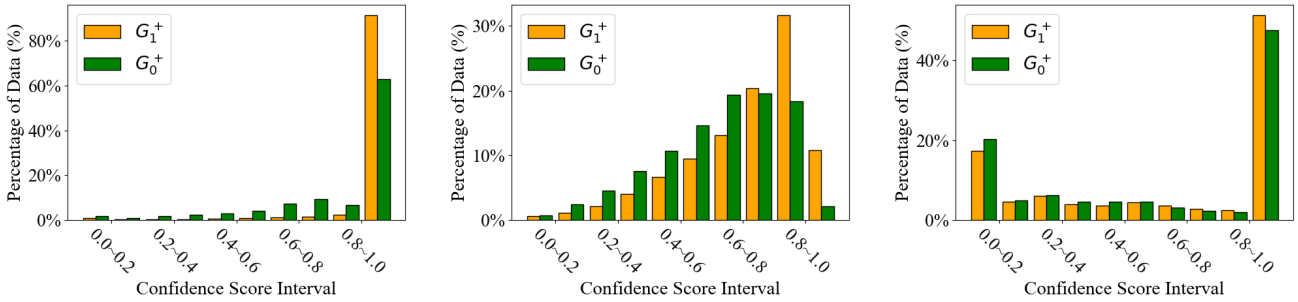
(e) Hospital (Gender) (f) Hospital (Race)

Figure 1: Impacts of data distribution on PLD. X-axis is the distance between two data distributions, and y-axis is the difference of MIA accuracy between the two subgroups.

size and PLD. To verify the hypothesis, we first propose a scheme to quantify model memorization. Then we measure how much the model memorizes subgroups of different sizes.

Quantifying model memorization via influence scores: An intuitive idea measuring whether the model memorizes a given training sample is to ask the counterfactual [12, 24]: what would happen to the model output if the model did not see the training sample? Answering this counterfactual enables to trace a model's output back to the training data through the learning algorithm.

To quantify the effect of the counterfactual, we leverage *influence function*, a state-of-the-art model explanation method [38]. Influence function estimates the impact of a training point on a model prediction. In our setting, to estimate the effect of a data sample $z(x, y)$ on the target model, the explanation measures the difference in the loss function of the model when it is trained with and without z. Formally, assuming $\theta_z \triangleq \mathcal{T}(D^{\mathcal{T}}_{train} \backslash \{z\})$, that is, θ_z is induced by training the target model $\mathcal{T}$ given the dataset excluding z. Then the *influence score* of z on $\mathcal{T}$ is measured as $I(z) \triangleq L(\theta_z) - L(\theta)$, where θ is a prameterization induced by the training model $\mathcal{T}$ and L is the loss function of $\mathcal{T}$. Intuitively, higher influence score indicates the model memorizes more of the data sample. Given that the influence scores quantify how much the model memorizes individual data samples, they help explain why MIA better identifies some data samples than the others: data samples have higher influence scores are more likely to be correctly predicted by MIA.



(a) Adult (distance 0.33) (b) Broward (distance 0.23) (c) Hospital (distance 0.07)

Figure 2: Distribution of CSVs of subgroups (G_1^+ v.s. G_0^+, Gender, "$y = +$"). The distribution is split into five intervals.

A main challenge is to design an influence function to efficiently estimate the influence scores. A straightforward idea is to remove a training point, retrain the target model $\mathcal{T}$ with the remaining points from scratch, and compute the new prediction results. An influence function, working in this way, however can incur extremely high computation cost. To this end, we adapt the *influence function* proposed by Koh et al. [38], which estimates the influence scores for classification models without a re-training process.

The model memorizes more of the smaller subgroup: To understand how the target model memorizes each subgroup, we compute the influence score of every data sample, followed by calculating the average influence score (AIS) for all samples in each subgroup. The results are presented in Table 4 (AIS column). Except for the case on Adult dataset with gender as the protected attribute and "+" as the label, the minority subgroup consistently has a higher AIS. This aligns with our hypothesis: *the model memorizes more of the minority subgroup than the majority one*, and thus, leaks more privacy about the minority subgroup.

Our analysis unveils disparity in AIS of different subgroups. An related and interesting question is whether influence score and MIA accuracy are correlated at the individual level. To understand this question, we performed another evaluation where we refine the granularity of the subgroups such that each subgroup approximates an individual. Specifically, we sort and split the data by their influence score into 10 bins, followed by separately calculating the average MIA accuracy of each bin. The results in Figure 3 indicate that different subgroups, when classified by influence score, present similar MIA accuracy. Thus, we anticipate no evident correlation between influence score and MIA accuracy at the individual level.

To summarize, the target model memorizes more of the minority subgroup at an aggregate level (as shown by Table 4). When considering the individual level, it does not memorize more of particular records than the others (as shown in Figure 3).

Subgroup size is not the only factor that impacts influence score: Table 4 includes an exceptional case, namely G_1^+ v.s. G_0^+ for Adult data with Gender as the protected attribute. In the case, G_1^+ is the minority subgroup but has a lower AIS and thus a lower MIA accuracy than the majority group. The existence of the case indicates that the difference in AIS across groups is not solely determined by the group size. What remains as a potential factor is the difference in data distribution of the two groups. We accordingly run a follow-up analysis in the next subsection.

5.2 Impact of Data Distribution on PLD

To understand the impact of subgroup data distribution on PLD, we design a simulation experiment where the size of different subgroups is kept identical. Specifically, we randomly pick a subset of sample S from each dataset uniformly, ensuring the distribution of S is similar to the whole dataset. Once obtaining S, we generate a same-sized subset S' by adding noise N on the distribution of S. N follows the normal distribution with mean $\mu = 0$ and standard deviation $\delta \in \{0, 0.01, 0.1, 0.2, 0.4\}$. $\delta = 0$ means the distributions of S and S' are identical. Larger δ indicates more noise and thus higher distance between the distributions of S and S'.

When a pair of S and S' is created, we consider them as two subgroups and separately measure the MIA accuracy for each of them, reusing the target and attack models trained on the entire dataset. Meanwhile, we use Kolmogorov–Smirnov (KS) test to quantify the distance between the distributions in the target model output, or CSV, for S and S'. We considered the distribution of CSV instead of the distribution of the original data for two reasons. First, the original data has too many dimensions (see Table 2), whose distribution is complex to measure and visualize. In contrast, the CSV only has two dimensions (as all the target models are binary classifiers), making distribution analysis much easier. Second, the difference in CSV distribution reflects the difference in the original data distribution. An analysis on the CSV distribution should also support our goal.

Each test above is repeated 10 times. Figure 1 shows the results of PLD (i.e., the difference between the MIA accuracy on S and S') when the distributions of S and S' have a different distance. Apparently, difference in the data distribution of subgroups also affects PLD. In general, PLD increases when two subgroups have a larger distribution distance, i.e., their distributions are more different.

Which factor is dominant, size or data distribution: Our analysis so far unveils both size and data distribution of subgroups affect PLD. It brings up the question that which factor is the dominant one, or how the two factors impact PLD synthetically. Our observation is when difference in the data distributions is less intensive, size plays a more significant role and the minority subgroup tends to leak more privacy. Otherwise, data distribution may dominate and overturn the impact of size. Below we use the exceptional case discussed in §5.1 to demonstrate our observation.

We measure the distribution of CSV for G_1^+ and G_0^+ in that case. For comparison, the same measurement is done for G_1^+ and G_0^+ from other datasets with Gender as the protected attribute and + as the

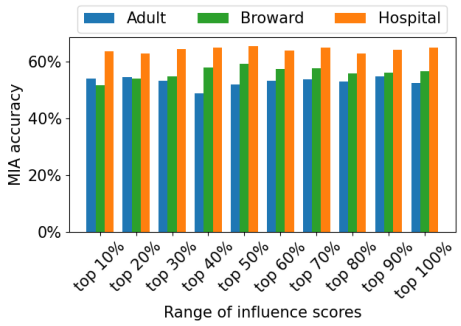


Figure 3: MIA accuracy of subgroups with different influence scores. Top x% at x-axis means [top-(x-10)%, top-x%].

label. All the results are shown in Figure 2. The major observation of our analysis is that the CSV distribution for G_1^+ and G_0^+ in Adult dataset varies more dramatically, compare to the Broward and Hospital datasets. We use the discrete Kolmogorov–Smirnov (KS) tests to quantify the distance between the CSV distribution of G_1^+ and G_0^+. It turns out that the distance with Adult dataset is much higher than the other two datasets, as illustrated in Figure 2. Because of this large distribution difference, G_1^+ has less privacy leakage than G_0^+, even though it has a smaller size.

6 IMPACT OF MIA DEFENSE ON PLD

The research community has designed multiple defenses against MIA. The defenses aim to reduce the privacy leakage of the entire population. An interesting question is *whether (and why) the defenses have disparate impacts across different subgroups and consequently mitigate PLD* (i.e., research question **RQ3** (§1))? This section seeks to answer this question from an empirical perspective.

Evaluation Metrics: We use the metrics described in §4.1 to measure the performance of the target model and the attack model. To measure PLD, we calculate the difference of MIA accuracy between two subgroups with same target label (+ or -), following §4.2.

6.1 Defense Mechanisms

In general, MIA defenses can be classified into two categories. One category modifies the training process of the target model such that it leaks less membership information to MIA. The other category adds perturbations to the CSVs instead of modifying the training process of the target model. We consider the following defenses in the first category, because of their wide use by the community.

Differential privacy (DP): [20] has become the de facto standard in privacy-preserving data analytics. Roughly speaking, a differentially private ML algorithm ensures that the inclusion of an individual training sample does not significantly affect the model output. We consider ϵ-*differential privacy*, where ϵ is called *privacy budget* and specifies the level of guaranteed privacy. A larger ϵ provides weaker privacy protection but smaller accuracy/utility loss of the model. We use the implementation of *TensorFlow-Privacy* [43] to enforce DP with {0.3, 0.5, 1.0, 3.0, 5.0} as the privacy budget ϵ.

L_2-Regularizer: Overfitting has been identified as one of the major reasons why MIA can be effective. Thus, a natural solution to defend against MIA is to reduce model overfitting by using regularization.

Table 6: Performance of target model and MIA under DP

Dataset	Privacy Budget (ϵ)	Train Acc.	Test Acc.	MIA Acc.
Adult	No DP	90.4%	82.5%	54.1%
	5.0	83.8%	82.9%	50.9%
	3.0	83.2%	82.7%	50.8%
	1.0	82.4%	82.3%	50.8%
	0.5	77.5%	77.3%	50.6%
	0.3	75.3%	75.2%	50.6%
Broward	No DP	71.3%	67.3%	51.8%
	5.0	68.4%	67.5%	50.2%
	3.0	67.7%	66.8%	50.1%
	1.0	66.6%	66.5%	49.8%
	0.5	65.0%	65.1%	49.7%
	0.3	56.9%	56.9%	49.8%
Hospital	No DP	91.8%	64.9%	65.1%
	5.0	70.1%	67.6%	51.6%
	3.0	68.7%	67.2%	51.2%
	1.0	66.3%	66.0%	50.7%
	0.5	65.4%	65.5%	50.5%
	0.3	65.4%	65.5%	50.5%

Shokri et al. [52] has shown the effectiveness of the conventional L_2-regularizer as a defense mechanism. L_2-regularizer adds the norm of the parameters to the loss function with a parameter λ, which controls the weight of the regularization. We use {0.00005, 0.0001, 0.001, 0.005, 0.01, 0.1} for the λ parameter, which controls the weight of parameter L_2 norm in the loss function. The higher the value we assign to λ, the stronger the defense is.

Dropout: also reduces overfitting [49]. It randomly deletes a proportion (*dropout ratio*) of edges in a fully connected neural network model in each training iteration. We use four dropout ratios {1%, 5%, 10%, 20% }. Larger dropout ratio indicates stronger defense.

6.2 Effectiveness of MIA Defenses

To validate the effectiveness of the defenses and their impacts on the target model, we measure the accuracy of both target and attack models after deploying the defenses. Table 6 shows the results of DP. The results for L_2-regularizer and Dropout are similar, which are presented in Table 11. First, DP is effective in mitigating MIA. For all the datasets, MIA accuracy decreases to close to 50% (i.e., random guess) when the privacy budget ϵ decreases. Second, DP harms the accuracy of the target model when providing stronger defense against MIA. Consider Adult dataset as an example. The target training accuracy drops from 90.4% to 75.3% and the target testing accuracy drops from 82.5% to 75.2% when the privacy budget ϵ approaches 0.3. These results align with the well-known trade-off issue between privacy and model accuracy.

6.3 Impacts of MIA Defenses on PLD

To understand the impacts of MIA defenses on PLD, we measure MIA accuracy of each subgroup. Table 7 summarises the results with DP as the defense. The results for L_2-regularizer and Dropout defenses are similar and can be found in Appendix (Tables 12 & 13).

Table 7: Impacts of DP on PLD. The columns of G_1^+ and G_0^+ show the MIA accuracy of the corresponding subgroup.

Dataset	Privacy budget ϵ	Gender						Race					
		G_1^+	G_0^+	PLD	G_1^-	G_0^-	PLD	G_1^+	G_0^+	PLD	G_1^-	G_0^-	PLD
Adult	No DP	51.9%	53.2%	1.3%	65.4%	57.2%	8.2%	52.9%	52.6%	0.3%	69.8%	57.3%	12.5%
	5.0	50.6%	50.8%	0.2%	51.3%	51.4%	0.1%	50.5%	50.7%	0.2%	52.1%	51.3%	0.8%
	3.0	50.4%	50.8%	0.4%	51.8%	51.3%	0.5%	50.6%	50.7%	0.1%	51.9%	51.3%	0.6%
	1.0	50.5%	50.6%	0.1%	51.3%	51.4%	0.1%	50.3%	50.6%	0.3%	51.3%	50.5%	0.8%
	0.5	50.3%	50.7%	0.4%	51.2%	50.9%	0.3%	50.5%	50.5%	0.0%	51.0%	50.7%	0.3%
	0.3	50.5%	50.5%	0.0%	51.0%	50.8%	0.2%	50.4%	50.5%	0.1%	50.2%	50.6%	0.4%
Broward	No DP	52.6%	51.5%	1.1%	54.0%	51.5%	2.5%	51.0%	52.4%	1.4%	52.7%	51.5%	1.2%
	5.0	49.6%	50.0%	0.4%	49.3%	50.7%	1.4%	49.9%	49.9%	0.0%	51.0%	49.7%	1.3%
	3.0	50.8%	50.5%	0.3%	48.9%	49.8%	0.9%	50.5%	49.8%	0.7%	50.5%	48.5%	2.0%
	1.0	49.7%	49.9%	0.2%	49.7%	49.8%	0.1%	49.8%	49.9%	0.1%	49.4%	50.4%	1.0%
	0.5	50.1%	49.5%	0.6%	49.3%	49.8%	0.5%	49.6%	49.7%	0.1%	50.0%	49.4%	0.6%
	0.3	50.2%	49.6%	0.6%	50.1%	49.7%	0.4%	50.0%	49.8%	0.2%	49.9%	49.7%	0.2%
Hospital	No DP	71.2%	73.5%	2.3%	62.1%	60.6%	1.5%	72.7%	71.4%	1.3%	62.6%	60.7%	1.9%
	5.0	52.3%	52.0%	0.3%	51.2%	51.3%	0.1%	52.5%	52.0%	0.5%	51.4%	51.2%	0.2%
	3.0	51.7%	51.5%	0.2%	51.0%	51.0%	0.0%	51.9%	51.4%	0.5%	51.0%	51.0%	0.0%
	1.0	50.8%	50.9%	0.1%	50.7%	50.5%	0.2%	50.9%	50.8%	0.1%	50.7%	50.6%	0.1%
	0.5	50.6%	50.7%	0.1%	50.5%	50.2%	0.3%	50.9%	50.5%	0.4%	50.3%	50.4%	0.1%
	0.3	50.7%	50.6%	0.1%	50.5%	50.5%	0.0%	50.8%	50.5%	0.3%	50.4%	50.5%	0.1%

MIA defenses mitigate PLD: When the privacy budget of DP decreases, MIA accuracy of the subgroups also drops, all gradually getting closer to 50%. In turn, this leads to a reduction of PLD. Even with a larger privacy budget (e.g., $\epsilon = 5$), PLD is significantly decreased. The results indicate that while mitigating MIA, DP also helps reduce PLD. Similar trends are observed on L_2-regularizer and Dropout, as shown in Tables 12 and 13 in the Appendix.

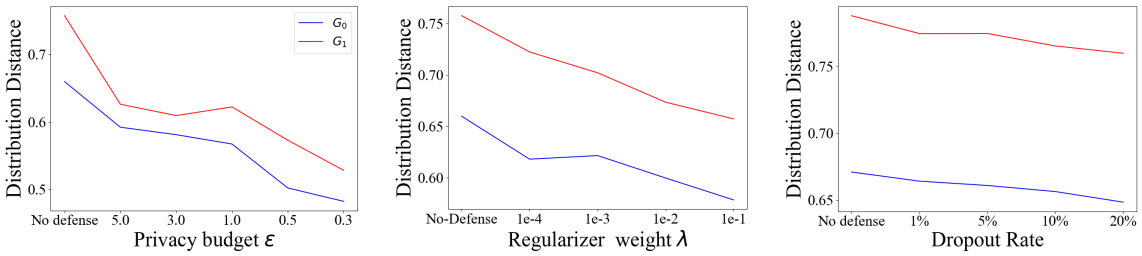
A follow-up question is how exactly DP and other MIA defenses reduce PLD. In principle, MIA defenses work by making samples less distinguishable from each other such that the attacker cannot recognize the members. Indirectly, the defenses make different subgroups similar and thus, shrink the distance between the distributions of their target model output (or CSV). Further considering our previous observation that PLD decreases with the CSV distribution distance (see Figure 1), it is expected that MIA defenses reduce PLD. To verify this reasoning, we measure the distance between the CSV distribution of the minority and majority groups, before and after the MIA defenses are applied. Figure 4 (a)-(c) show the results on Adult dataset with Race as the protected attribute. As the defense strength increases, the CSV distribution distance between the two subgroups decreases, well supporting our reasoning.

Defense unfairness exists: MIA defenses reducing PLD leads to an interesting phenomenon: *the defense mechanisms provide different amounts of protection across different subgroups.* Connecting the phenomenon to that the minority subgroup tends to have higher privacy leakage, we derive another observation: *the defense mechanisms provide stronger protection on the minority subgroups.* Consider Adult dataset with Race as the protected attribute as an example (see Table 6). MIA accuracy of the minority subgroup G_1^- drops from 69.8% (without DP) to 50.2% ($\epsilon = 0.3$), presenting a 28% of reduction. In contrast, MIA accuracy of the majority group G_0^- decreases from 57.3% to 50.6%, only showing a 12% of reduction. Similar disparity also arises in other settings and the other two defenses, as illustrated in Tables 12 and 13 in the Appendix. This raises an interesting question: *why MIA defenses provide more protection on the minority subgroup than the majority group?*

6.4 Analysis of Defense Unfairness

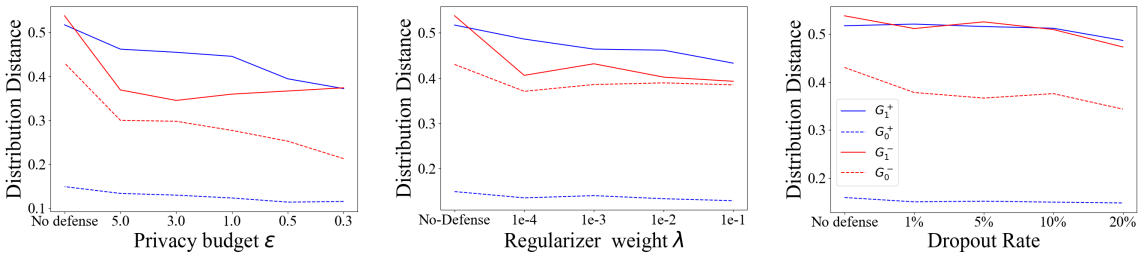
Prior work [5] has shown that DP has disparate impacts on model accuracy. In particular, the accuracy of target model deteriorates more for the minority subgroups. However, such disparate impacts on target model accuracy cannot explain the disparate impacts on MIA accuracy. Therefore, to understand the above question, we once again borrow insights from the principle of MIA defenses. As pointed out above, MIA defenses work by making samples less distinguishable from each other. As such, MIA defenses shall concurrently make subgroups less different from the entire population. We accordingly measure the distance between the CSV distributions of each subgroup and the entire population, before and after MIA defenses. Figure 4 presents the results on Adult dataset with Race as the protected attribute. The results on Broward and Hospital datasets are similar and shown in Figures 7 & 8. Evidently, MIA defenses reduce the CSV distribution distance between each subgroup and the entire population. Another trend is MIA defenses reduce the distance for the minority subgroup more dramatically. But how does this connect to defense unfairness? We give an analysis below.

Recalling §5.1, the target model leaks more privacy of the minority subgroup because it memorizes more about that subgroup. A related, more fundamental matter is why the target model memorizes more about the minority subgroup. We believe a reason is



(a) DP (b) L_2-regularizer (c) Dropout

Change of CSV distribution distance between subgroups with increase of defense strength



(d) DP (e) L_2-regularizer (f) Dropout

Change of CSV distribution distance each subgroup and the entire population with increase of defense strength

Figure 4: Impact of defense mechanisms on subgroups (Adult dataset, Race). In sub-figures (a)-(c), each curve shows the distance between two subgroups with the same label (we use G_0 and G_1 to denote the two subgroups). In sub-figures (d)-(f), the two subgroups of the same target label have the same color; The minority/majority subgroups are labeled with solid/dotted lines.

the minority subgroup deviates more from the entire population. Consider Figure 4 (d)/(e)/(f) as an example. Without MIA defenses, the minority subgroup has a larger CSV distribution distance to the overall population, compared to the majority subgroup. As a consequence (presumably), the target model memories more about the minority subgroup and leaks more privacy from that subgroup (see Table 4). Connecting everything together, MIA defenses reduce the distances between the minority subgroup and the overall population quicker and thus, de-memorize the minority subgroup quicker. This results in stronger protection for the minority subgroup.

Correlation between defense unfairness and distribution distance: Our analysis above correlates defense unfairness to distribution distance between two subgroups. We took a further step to understand how the two metrics are correlated exactly. We first introduce a metric to quantify defense unfairness. Specifically, we consider the delta between the *reduced MIA accuracy* of two subgroups when a defense is applied as defense unfairness:

$$\text{Defense Unfairness} = |\Delta(acc(G_0^+)) - \Delta(acc(G_1^+))|$$

where $\Delta(acc(G_0^+))/\Delta(acc(G_1^+))$ denotes the change of MIA accuracy of G_0^+/G_1^+ before and after the defense.

Leveraging the above metric, we measure the level of defense unfairness and the change of distribution distance when gradually increasing the defense strengthen. Figure 5 shows the results with DP as the defense mechanism on Hospital dataset. The results on Adult and Broward datasets are similar and can be found in Appendix (Figure 6). The distribution distance is measured by Kolmogorov–Smirnov test. A key observation, reflected by the results, is that defense unfairness is more evident when the distribution distance is reduced to a larger extent. This is not surprising. According to what we discussed in §6.2, the reduction of distribution distance shall lead to reduction of PLD, which is essentially defense unfairness. Another interesting observation is that the defense tends to reduce the distribution distance more rapidly when the distance value is larger and, in turn, present higher defense unfairness.

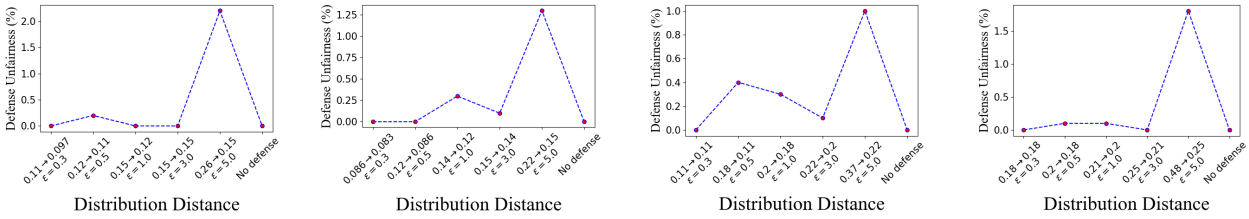
6.5 Privacy Unfairness *v.s.* Defense Unfairness

PLD arises because MIA presents disparate accuracy on different subgroups. To mitigate PLD, a defense must decrease the MIA accuracy of one group harder, which essentially causes defense unfairness. In this regard, privacy unfairness principally conflicts with defense unfairness, and the latter is a cue to the former. While it is also interesting to study defense unfairness and the related mitigation, the matter goes beyond the scope of this paper.

7 RELATED WORK

MIA and Defenses: MIA [30, 50, 51] predicts whether a given record was used in training a target model, typically under a black-box setting where the target model exposes a prediction API to the attacker. Some recent works [42, 46, 49, 56] provide more detailed study of MIA. MIAs have also been developed to attack federated/collaborative learning [44, 46], generative adversarial networks (GANs) [14, 28], adversarially robust deep learning models [54], embedding models [53], and GNNs [29]. On the horizon, new variants of MIAs, e.g., label-only MIAs [16, 41], are arising.

Several defenses have been designed to defend against MIA. As described in §6, the defenses belong to two major categories. The first category of defenses modify the training process of the target model such that it leaks less membership information. Exemplary defenses include differential privacy [20], dropout and model stacking [49], and adversarial regularization [45]. In contrast, the



(a) Gender, $y = +$ (b) Gender, $y = -$ (c) Race, $y = +$ (d) Race, $y = -$

Figure 5: Defense unfairness *v.s.* distance between subgroup distributions (DP as defense, Hospital dataset). The x-axis is the change of distance between the subgroup distributions under different privacy budgets; The y-axis is the defense unfairness.

second category of defenses add perturbations to the CSV instead of modifying the training process of the target model [34].

Algorithmic Fairness in Machine Learning: Fairness has caught increasing attention from the ML community [18]. Several competing notions of algorithmic fairness in machine learning have been recently proposed. These definitions can be grouped into two broad classes, namely *group fairness* and *individual fairness*. Group fairness [7, 9, 10, 23, 36] is concerned with a small number of *protected groups* (*e.g.*, females) that are defined by the *protected attributes* (*e.g.*, gender). It requires that the protected groups should have some form of statistical parity (*e.g.*, between positive outcomes or errors) compared with either the advantaged groups (*e.g.*, males) or the populations as a whole. On the other hand, individual fairness [21] requires people who are "similar" receive similar outcomes. Our PLD fairness definition belongs to the category of group fairness. In particular, we adapt accuracy parity [6, 8, 15] to our setting to evaluate disparity of privacy vulnerability.

Algorithmic Fairness and Privacy The interaction between privacy and fairness has attracted increasing attention recently. Several works [26, 31, 48, 55] have explored how to achieve fairness and privacy jointly. These works consider fairness and privacy as two independent objectives, while we consider the intersection of fairness and privacy, or *fair privacy*. Dwork et al. [21] initialize the exploration of the relationship between algorithmic fairness and privacy. They show that differential privacy techniques can be adapted to satisfy fairness in ML. In a later position paper [22], the authors propose a set of high-level research questions of understanding the interaction between fairness and privacy. Our work answers one of the questions that whether the privacy attacks are more effective against particular members of protected groups.

The disparate effects of ML models across different groups have been observed in several recent works. Bagdasaryan et al. [5] show that DP has disparate effects on model accuracy - the differentially private models have larger accuracy reduction on the underrepresented groups. It only used the image data for the empirical evaluation of disparate impact of DP. We extend the study to tabular data. We also consider the other two defenses (L_2-regularizer and Dropout) besides DP. Jagielski et al. [32] investigated the MIA with the presence of DP. While their motivation is very different from ours, they show some similar observation as ours that membership inference does not affect each training sample uniformly. Chang et al. [12] recently explored whether enforcing fairness in ML can incur privacy risks, which is complementary to our work. The most related to ours is probably the recent work by Yaghini et al. [57] (developed independently and in parallel with ours). Similar to us, they identified the existence of disparate vulnerability across different demographic groups against MIA. They also identified subgroup data distribution and subgroup size as two factors influencing disparate vulnerability, but without deeper analysis of why and how. Furthermore, they only considered DP [19] as the defense mechanism and showed similar results as ours. However, they did not provide detailed analysis why DP has such effect.

8 CONCLUSION AND FUTURE WORK

This paper studies the topic of fair privacy in the context of MIAs. We focus on two major issues, *privacy unfairness* (i.e., disparate vulnerability of MIAs across different subgroups) and *defense unfairness* (i.e., disparate protection by the defenses against MIA across subgroups). First, we formally define the notion of privacy-leakage disparity (PLD) to measure the disparity of privacy vulnerability to MIA across different subgroups. Then we show that PLD exists, through extensive empirical studies on real-world datasets. We investigate why and how subgroup size and subgroup data distribution impact PLD. Finally, we show that defense unfairness also exists for three widely-used MIA defenses (DP, L_2-regularizer, and Dropout), which actually mitigates privacy unfairness.

There are multiple interesting research directions to explore. First, our PLD metric measures the difference in MIA accuracy of different demographic subgroups. It remains to be investigated if other privacy-leakage disparity metrics based on the difference in MIA precision, recall (as shown in Appendix (§ C)), or F-measure can be adapted, and whether PLD will change if the metrics are different. Second, we focus on disparity of different subgroups. Disparity can also be defined at an individual level (known as *individual fairness*). Briefly speaking, individual privacy fairness requires that similar objects receive similar treatment. It is interesting to examine if individual PLD exists, and how to design methods to mitigate PLD at the individual level.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their feedback. This project was supported by the National Science Foundation (#CNS-2029038; #CNS-1937786) and an IBM Faculty Award. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

REFERENCES

[1] Adult dataset. https://archive.ics.uci.edu/ml/datasets/Adult.

[2] Broward dataset. https://farid.berkeley.edu/downloads/publications/scienceadvances17/.

[3] Keras python deep learning toolkit. https://keras.io/.

[4] Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics 26*, 10 (2010), 1340–1347.

[5] Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems* (2019), pp. 15479–15488.

[6] Barocas, S., and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev. 104* (2016), 671.

[7] Biddle, D. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

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

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

[10] Calders, T., and Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery 21*, 2 (2010), 277–292.

[11] Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)* (2019), pp. 267–284.

[12] Chang, H., and Shokri, R. On the privacy risks of algorithmic fairness. *arXiv preprint arXiv:2011.03731* (2020).

[13] Chang, H., and Shokri, R. On the privacy risks of algorithmic fairness. In *Proceedings of IEEE European Symposium on Security and Privacy (EuroSP)* (2021).

[14] Chen, D., Yu, N., Zhang, Y., and Fritz, M. Gan-leaks: A taxonomy of membership inference attacks against gans. *arXiv preprint arXiv:1909.03935* (2019).

[15] Chi, J., Zhao, H., Gordon, G., and Tian, Y. Understanding and mitigating accuracy disparity in regression, 2021.

[16] Choquette-Choo, C. A., Tramer, F., Carlini, N., and Papernot, N. Label-only membership inference attacks. In *International Conference on Machine Learning* (2021), PMLR, pp. 1964–1974.

[17] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*, 2 (2017), 153–163.

[18] Chouldechova, A., and Roth, A. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[19] Dwork, C. Differential privacy. In *the International Colloquium on Automata, Languages and Programming (ICALP)* (2006), Springer, pp. 1–12.

[20] Dwork, C. Differential privacy. *Encyclopedia of Cryptography and Security* (2011), 338–340.

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

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

[23] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 259–268.

[24] Feldman, V. 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.

[25] Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (2015), ACM, pp. 1322–1333.

[26] Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., and Giannotti, F. Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery 29*, 6 (2015), 1733–1782.

[27] Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (2016), pp. 3315–3323.

[28] Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies 2019*, 1 (2019), 133–152.

[29] He, X., Jia, J., Backes, M., Gong, N. Z., and Zhang, Y. Stealing links from graph neural networks. In *30th {USENIX} Security Symposium ({USENIX} Security 21)* (2021).

[30] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet 4*, 8 (2008), e1000167.

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

[32] Jagielski, M., Ullman, J., and Oprea, A. Auditing differentially private machine learning: How private is private sgd? *arXiv preprint arXiv:2006.07709* (2020).

[33] Jayaraman, B., Wang, L., Knipmeyer, K., Gu, Q., and Evans, D. Revisiting membership inference under realistic assumptions. *arXiv preprint arXiv:2005.10881* (2020).

[34] Jia, J., Salem, A., Backes, M., Zhang, Y., and Gong, N. Z. Memguard: Defending against black-box membership inference attacks via adversarial examples. *arXiv preprint arXiv:1909.10594* (2019).

[35] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2012), Springer, pp. 35–50.

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

[37] Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS* (2017), pp. 43:1–43:23.

[38] Koh, P. W., and Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning* (2017), PMLR, pp. 1885–1894.

[39] Leino, K., and Fredrikson, M. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. *arXiv preprint arXiv:1906.11798* (2019).

[40] Leino, K., and Fredrikson, M. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th {USENIX} Security Symposium ({USENIX} Security 20)* (2020), pp. 1605–1622.

[41] Li, Z., and Zhang, Y. Membership leakage in label-only exposures. *arXiv preprint arXiv:2007.15528* (2020).

[42] Long, Y., Bindschaedler, V., Wang, L., Bu, D., Wang, X., Tang, H., Gunter, C. A., and Chen, K. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889* (2018).

[43] McMahan, H. B., and Andrew, G. A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210v2* (2018).

[44] Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)* (2019), IEEE, pp. 691–706.

[45] Nasr, M., Shokri, R., and Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* (2018), ACM, pp. 634–646.

[46] Nasr, M., Shokri, R., and Houmansadr, A. 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)* (2019), IEEE, pp. 739–753.

[47] of State Health Services, T. D. Hospital discharge data use agreement. https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm.

[48] Ruggieri, S., Hajian, S., Kamiran, F., and Zhang, X. Anti-discrimination analysis using privacy attack strategies. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2014), Springer, pp. 694–710.

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

[50] Sankararaman, S., Obozinski, G., Jordan, M. I., and Halperin, E. Genomic privacy and limits of individual detection in a pool. *Nature genetics 41*, 9 (2009), 965–967.

[51] Shokri, R., and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security* (2015), ACM, pp. 1310–1321.

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

[53] Song, C., and Raghunathan, A. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security* (2020), pp. 377–390.

[54] Song, L., Shokri, R., and Mittal, P. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)* (2019), IEEE, pp. 50–56.

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

[56] Truex, S., Liu, L., Gursoy, M. E., Yu, L., and Wei, W. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing* (2019).

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

[58] Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer*

Table 8: Setup of groups based on the protected attribute

Dataset	Protected Attribute			
	Gender		Race	
	Group G_1	Group G_0	Group G_1	Group G_0
Adult	Female	Male	Non-white	White
Broward			Non-black	Black
Hospital			Non-white	White

Table 9: Importance and ranking of protected attributes

Dataset	Gender		Race	
	Relative Importance	Ranking	Relative Importance	Ranking
Adult	1.81%	$14^{th}/14$	2.20%	$12^{th}/14$
Broward	2.68%	$6^{th}/8$	3.02%	$5^{rd}/8$
Hospital	1.99%	$15^{th}/16$	2.84%	$7^{th}/16$

Security Foundations Symposium (CSF) (2018), IEEE, pp. 268–282.

[59] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (2017), pp. 1171–1180.

APPENDIX

A GROUPING

Table 8 shows the setup of groups by gender and race. We note that the race attribute has more than two values in every dataset. Accordingly, we assign all race values into two racial groups by following the conventions in the fairness community [9, 27, 35].

B FEATURE IMPORTANCE OF THE PROTECTED ATTRIBUTES

Table 9 shows the feature importance of the protected attribute in the three datasets. The feature importance is measured by permutation feature importance method [4]. The results show that the protected attribute (Gender or Race) are not of high importance for training of the target model. The rankings of features based on feature importance in Table 9 also demonstrate the unimportance of the two protected attributes.

C MIA PRECISION AND RECALL OF SUBGROUPS

Table 10 includes the results of MIA precision and recall for each subgroup. Similar to the observation of MIA accuracy (Table 4), different subgroups have disparate MIA precision and recall. The disparity can be significant in some settings. For example, the difference in MIA recall on Adult dataset (with Gender as the protected attribute) can be as large as around 13% between G_1^+ and G_0^+. On the other hand, while the disparity of MIA precision also exists across all the settings, it is not as as large as that of MIA recall. Furthermore, some minority subgroups may receive lower (or higher) MIA precision, but higher (or lower) MIA recall. For instance, on Adult dataset with Gender as the protected attribute, the minority subgroup G_1^+ has higher MIA precision but lower MIA recall compared with the majority subgroup G_0^+. This raises an interesting research question: which fairness definition is appropriate for evaluation of privacy-leakage disparity? We leave this to the future work.

D DEFENSE EFFECTIVENESS OF L_2-REGULARIZER AND DROPOUT

We show the performance of both Dropout and L_2 defense mechanisms in Table 11. Similar to the observations of DP (Table 6), both Dropout and L2-regularizer mechanisms are effective to defend against MIA. In particular, MIA accuracy eventually achieve 50% (random guess) when both defenses get stronger. The only exceptional case is when Dropout is applied on the target model on Hospital dataset. The best MIA accuracy that Dropout can reduce to is still around 60%. The possible reason is that the model is still overfitting even when the dropout ratio is as high as 20%. We also observe that both training and testing accuracy decreases when the defense gets stronger. This is expected due to the trade-off between privacy and accuracy.

E IMPACTS OF L_2-REGULARIZER AND DROPOUT ON PLD

Table 12 shows the MIA accuracy per subgroup as well as PLD when Dropout is used as the defense mechanism with various dropout ratios. The main observation is that Dropout mitigates PLD on all the three datasets - PLD eventually approaches to zero when the dropout ratio increases (i.e., stronger defense). Furthermore, MIA accuracy decreases at different speeds for different subgroups. In most of the settings, MIA accuracy decreases the fastest on the minority subgroups. This is consistent with our findings when DP is the defense mechanism (§ 6). We believe the underlying reasons of such defense disparity are the same as what we discovered for DP (§ 6). We have similar observations when L_2-regularizer is used as the defense (Table 13).

F ANALYSIS OF DEFENSE UNFAIRNESS ON BROWARD AND HOSPITAL DATASETS

Figures 7 and 8 present how the distribution distance changes with increase of defense strength on Broward and Hospital datasets respectively. The observations are similar to Adult dataset (Figure 4 in §6.4) and thus are omitted due to limited space.

G DEFENSE UNFAIRNESS *V.S.* DISTRIBUTION DISTANCE

Figure 6 presents the level of defense unfairness and the change of distribution distance when gradually increasing the defense strengthen on Adult and Broward datasets. The observation is similar as on Hospital dataset (Figure 5). We omit the detailed discussion due to limited space.

Table 10: MIA precision and recall of subgroups. Between every two subgroups of the same label, the minority subgroup is marked with green, the higher MIA precision is marked with orange and the higher recall is marked with pink.

Dataset	Label y	Subgroup	Gender			Race		
			size	MIA precision	MIA Recall	size	MIA precision	MIA Recall
Adult	+	G_1^+	6512	51.1%	97.7%	2654	51.4%	98.3%
		G_0^+	10493	52.0%	84.8%	14351	51.6%	88.2%
	−	G_1^-	833	60.3%	90.5%	495	63.0%	96.3%
		G_0^-	4773	54.3%	89.4%	5111	54.4%	88.9%
Broward	+	G_1^+	453	52.3%	75.1%	1084	50.6%	73.7%
		G_0^+	1530	51.0%	66.6%	899	52.1%	48.9%
	−	G_1^-	249	55.1%	51.4%	678	52.4%	62.2%
		G_0^-	1375	50.9%	60.5%	946	51.1%	66.4%
Hospital	+	G_1^+	4849	66.6%	85.4%	3022	68.4%	84.4%
		G_0^+	4264	69.0%	85.0%	6092	66.3%	86.9%
	−	G_1^-	8253	60.1%	72.1%	5051	61.0%	70.2%
		G_0^-	9022	58.3%	74.8%	12225	58.4%	74.8%

Table 11: Performance of Dropout and L_2-regularizer on both target model and MIA

Dataset	Dropout Ratio	Train acc.	Test acc.	MIA Acc
Adult	Original	90.4%	82.5%	54.1%
	1%	89.9%	83.1%	53.7%
	5%	89.0%	83.5%	53.3%
	10%	88.2%	83.8%	52.8%
	20%	86.9%	84.4%	51.9%
Broward	Original	71.3%	67.3%	51.8%
	1%	70.4%	62.3%	51.2%
	5%	66.7%	61.3%	51.6%
	10%	64.4%	60.5%	50.4%
	20%	63.1%	60.8%	50.6%
Hospital	Original	91.8%	64.9%	65.1%
	1%	91.4%	62.1%	64.0%
	5%	91.0%	62.3%	63.8%
	10%	90.8%	62.7%	63.3%
	20%	89.8%	62.0%	62.6%

(a) Dropout

Dataset	Regularizer parameter λ	Train acc.	Test acc.	MIA Acc
Adult	Original	90.4%	82.5%	54.1%
	0.0001	85.9%	82.9%	52.0%
	0.001	84.9%	84.7%	50.9%
	0.01	83.1%	83.1%	50.8%
	0.1	75.1%	75.3%	50.4%
Broward	Original	71.3%	67.3%	51.8%
	0.0001	71.2%	65.3%	51.4%
	0.001	70.5%	67.3%	50.8%
	0.01	69.0%	67.6%	50.7%
	0.1	67.8%	68.4%	50.6%
Hospital	Original	91.8%	64.9%	65.1%
	0.0001	92.3%	64.8%	64.9%
	0.001	87.5%	66.2%	61.6%
	0.01	70.5%	69.1%	51.0%
	0.1	65.4%	65.5%	50.2%

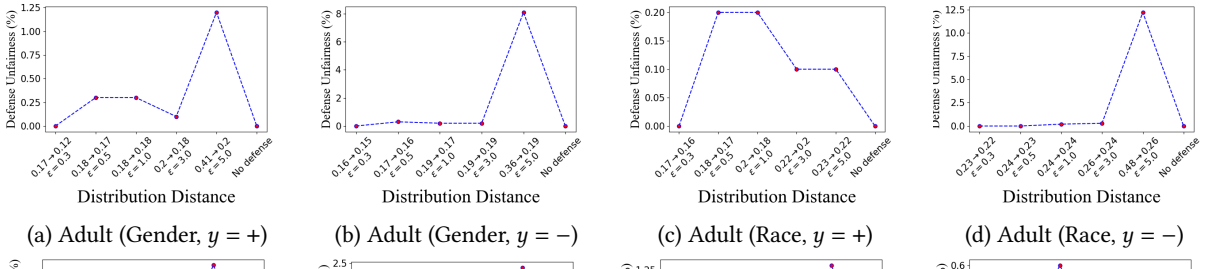
(b) L_2-regularizer

Table 12: Impacts of Dropout as the defense mechanism on privacy-leakage disparity (PLD)

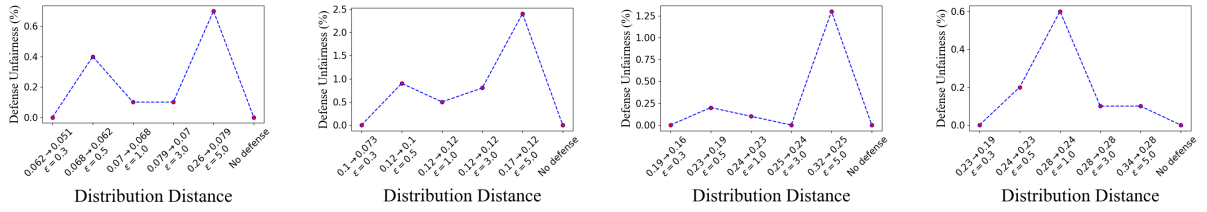
Dataset	Dropout Ratio	Gender						Race					
		G_1^+	G_0^+	PLD	G_1^-	G_0^-	PLD	G_1^+	G_0^+	PLD	G_1^-	G_0^-	PLD
Adult	Original	51.9%	53.2%	1.3%	65.4%	57.2%	8.2%	52.9%	52.6%	0.3%	69.8%	57.3%	12.5%
	1%	51.5%	52.9%	1.4%	64.4%	56.6%	7.8%	52.4%	52.4%	0.0%	66.7%	56.9%	9.8%
	5%	51.3%	52.6%	1.3%	63.1%	55.8%	7.3%	53.1%	51.9%	1.2%	64.9%	56.1%	8.8%
	10%	51.4%	52.2%	0.8%	60.8%	54.6%	6.2%	52.6%	51.8%	0.8%	65.0%	54.6%	10.4%
	20%	51.2%	51.6%	0.4%	58.0%	52.7%	5.3%	51.6%	51.4%	0.2%	59.7%	52.8%	6.9%
Broward	Original	52.6%	51.5%	1.1%	54.0%	51.5%	2.5%	51.0%	52.4%	1.4%	52.7%	51.5%	1.2%
	1%	52.2%	52.0%	0.2%	54.3%	52.2%	2.1%	51.2%	52.5%	1.3%	52.6%	51.7%	0.9%
	5%	52.3%	51.5%	0.8%	53.3%	51.4%	1.8%	50.6%	52.1%	1.5%	51.7%	50.9%	0.8%
	10%	50.6%	50.6%	0.0%	52.7%	51.1%	1.6%	50.8%	51.0%	0.2%	51.5%	51.2%	0.3%
	20%	50.7%	50.5%	0.2%	51.8%	50.9%	0.9%	50.2%	50.9%	0.7%	51.3%	51.0%	0.3%
Hospital	Original	71.2%	73.5%	2.3%	62.1%	60.6%	1.5%	72.7%	71.4%	1.3%	62.6%	60.7%	1.9%
	1%	70.9%	72.8%	1.9%	60.6%	60.7%	0.1%	71.7%	70.6%	1.1%	60.9%	59.8%	1.1%
	5%	70.8%	70.5%	0.3%	60.0%	59.6%	0.4%	70.4%	69.8%	0.6%	61.0%	60.0%	1.0%
	10%	69.3%	70.4%	1.1%	59.7%	58.4%	1.3%	69.1%	69.1%	0.0%	60.8%	59.3%	1.5%
	20%	68.7%	70.1%	1.4%	59.1%	58.1%	1.0%	66.6%	67.0%	0.4%	60.5%	59.1%	1.4%

Table 13: Impacts of L_2-regularizer as the defense mechanism on privacy-leakage disparity (PLD)

Dataset	Regularizer weight λ	Gender						Race					
		G_1^+	G_0^+	PLD	G_1^-	G_0^-	PLD	G_1^+	G_0^+	PLD	G_1^-	G_0^-	PLD
Adult	Original	51.9%	53.2%	1.7%	65.4%	57.2%	8.2%	52.9%	52.6%	0.3%	69.8%	57.3%	12.5%
	1.00E-04	51.0%	51.7%	0.7%	57.3%	53.1%	4.2%	51.3%	51.4%	0.1%	58.3%	53.3%	5.0%
	1.00E-03	50.8%	50.6%	0.2%	51.2%	51.5%	0.3%	50.8%	50.6%	0.2%	51.8%	51.4%	0.4%
	1.00E-02	50.3%	50.8%	0.5%	51.1%	51.2%	0.1%	50.6%	50.6%	0.1%	52.5%	51.1%	1.4%
	1.00E-01	50.2%	50.4%	0.2%	50.7%	50.7%	0.0%	50.5%	50.3%	0.2%	51.2%	50.6%	0.6%
Broward	Original	52.6%	51.5%	1.1%	54.0%	51.5%	2.5%	51.0%	52.4%	1.4%	52.7%	51.5%	1.2%
	1.00E-04	52.4%	51.5%	0.9%	53.5%	51.7%	1.8%	51.7%	52.5%	0.8%	52.5%	51.2%	1.3%
	1.00E-03	51.2%	50.9%	0.3%	51.9%	50.9%	1.0%	50.9%	51.1%	0.2%	51.9%	50.8%	1.1%
	1.00E-02	50.0%	50.2%	0.2%	50.9%	50.5%	0.4%	50.7%	50.4%	0.3%	51.2%	50.1%	1.1%
	1.00E-01	50.5%	50.4%	0.1%	50.6%	50.7%	0.1%	50.2%	50.2%	0.0%	51.3%	50.5%	0.6%
Hospital	Original	71.2%	73.5%	2.3%	62.1%	60.6%	1.5%	72.7%	71.4%	1.3%	62.6%	60.7%	1.9%
	1.00E-04	68.7%	70.8%	1.9%	61.0%	62.2%	1.2%	68.4%	67.1%	1.3%	60.3%	61.7%	1.4%
	1.00E-03	65.5%	67.7%	2.2%	58.6%	59.5%	0.9%	61.8%	58.9%	2.9%	56.8%	57.9%	1.1%
	1.00E-02	51.2%	51.9%	0.7%	50.9%	51.1%	0.2%	51.1%	51.0%	0.1%	50.8%	51.0%	0.2%
	1.00E-01	50.8%	51.4%	0.6%	50.2%	50.1%	0.1%	50.5%	50.3%	0.2%	50.1%	50.2%	0.1%

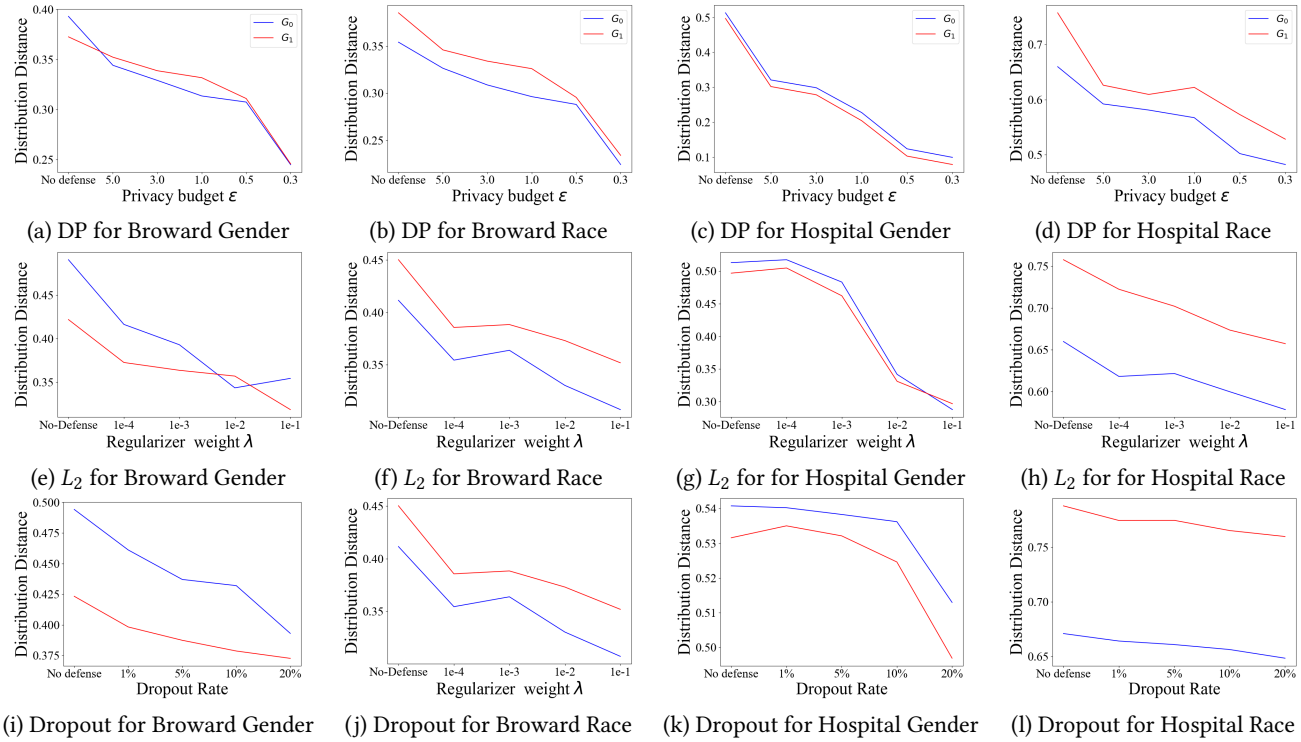


(a) Adult (Gender, $y = +$) (b) Adult (Gender, $y = -$) (c) Adult (Race, $y = +$) (d) Adult (Race, $y = -$)



(e) Broward (Gender, $y = +$) (f) Broward (Gender, $y = -$) (g) Broward (Race, $y = +$) (h) Broward (Race, $y = -$)

Figure 6: Change of defense unfairness with distribution distance under DP. The x-axis represents the distance changes between the distribution of two subgroups under different privacy budgets, and the y-axis represents defense unfairness.

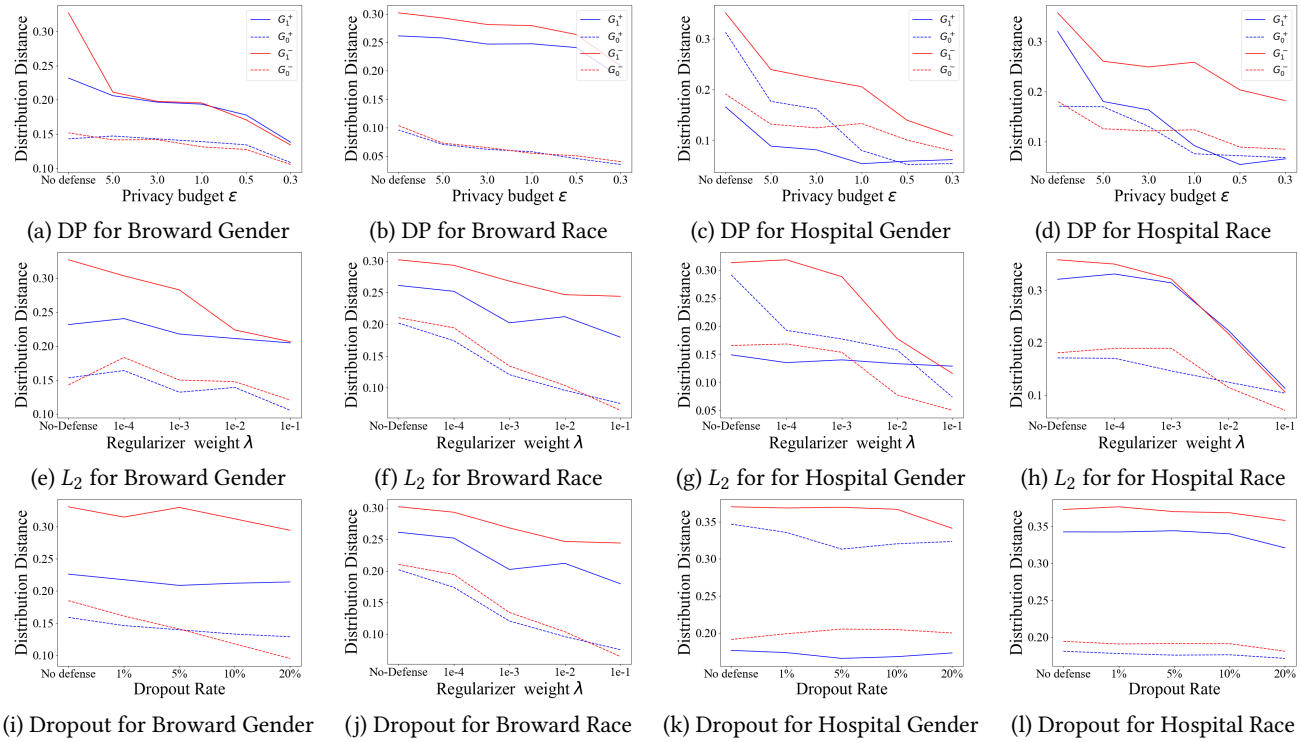


(a) DP for Broward Gender (b) DP for Broward Race (c) DP for Hospital Gender (d) DP for Hospital Race

(e) L_2 for Broward Gender (f) L_2 for Broward Race (g) L_2 for for Hospital Gender (h) L_2 for for Hospital Race

(i) Dropout for Broward Gender (j) Dropout for Broward Race (k) Dropout for Hospital Gender (l) Dropout for Hospital Race

Figure 7: Change of CSV distribution distance between subgroups with increase of defense strength



(a) DP for Broward Gender (b) DP for Broward Race (c) DP for Hospital Gender (d) DP for Hospital Race

(e) L_2 for Broward Gender (f) L_2 for Broward Race (g) L_2 for for Hospital Gender (h) L_2 for for Hospital Race

(i) Dropout for Broward Gender (j) Dropout for Broward Race (k) Dropout for Hospital Gender (l) Dropout for Hospital Race

Figure 8: Change of CSV distribution distance each subgroup and the entire population with increase of defense strength.