Dissecting Distribution Inference

Anonymous Authors

***Abstract*—A distribution inference attack aims to infer statistical properties of data used to train machine learning models. These attacks are sometimes surprisingly potent, but the factors that impact distribution inference risk are not well understood and demonstrated attacks often rely on strong and unrealistic assumptions such as full knowledge of training environments even in supposedly black-box threat scenarios. To improve understanding of distribution inference risks, we develop a new black-box attack that even outperforms the best known white-box attack in most settings. Using this new attack, we evaluate distribution inference risk while relaxing a variety of assumptions about the adversary's knowledge under black-box access, like known model architectures and label-only access. Finally, we evaluate the effectiveness of previously proposed defenses and introduce new defenses. We find that although noise-based defenses appear to be ineffective, a simple re-sampling defense can be highly effective.**

I. Introduction

Machine learning models are susceptible to several disclosure risks, including leaking sensitive information related to training data. Distribution inference considers what a model reveals about its entire underlying training data, in contrast with inference attacks that focus on individual records like membership inference and memorization attacks. Initial work focuses on inferring ratios of data with a particular attribute, referring to the attacks as "property inference". Examples include inferring the accent of speakers in voice recognition models [1], targeting ratios of characteristics like gender labels [2], and estimating sentiment across email datasets [3]. More recently, there have been attempts to extend these attacks to properties beyond ratios, such as predicting graph density [4], node/edge properties of groups within graphs [5], and direct regression over graph mean-degree [6] with successful inference adversaries with just black-box access.

State-of-the-art distribution inference attacks achieve non-trivial distinguishing accuracies [2], [6], [7] and thus pose a privacy risk, but the actual amount of leakage achieved is often minimal. Leakage varies significantly across different datasets, but for most settings the best current attacks leak no more information than what one or two samples from the distribution would reveal [6]. This is enough to have a significant advantage in distinguishing highly dissimilar distributions, but seems unlikely to pose a serious privacy risk in most cases.

Contributions. We advance the understanding of distribution inference risk on several fronts including an improved attack, analysis of risk, and development and evaluation of defenses. We introduce a new black-box attack, the KL Divergence Attack, that uses distributional similarity in predictions (Section II-B). and substantially outperform the current state-of-the-art (Section III), increasing previous estimates of inference leakage for various datasets in the literature. Surprisingly, we find that in most settings our black-box KL Divergence Attack is more effective than the best known white-box attack. We also evaluate the black-box attacks in more realistic settings where the adversary does not have as much information as is typically assumed in inference experiments (Section IV).

We evaluate the impact of different model architectures (Section IV-A), the lack of common feature extractors (Section IV-B), and relaxing the assumption of prediction probabilities to label-only access settings (Section IV-C). Our experiments find large variances in inference risk due to relaxing assumptions about model architectures and feature extractors, but demonstrate that attacks can be effective in label-only settings.

Section V evaluates defenses, including both previously proposed ones and new ideas. Most privacy-related defenses for machine learning involve adding noise at some stage of the training; for instance, at the gradient-level in differentially private training [8], or at the data level with most implementations of adversarial training [9]. Our experiments find that these noise-based defenses provide little mitigation against distribution inference (Section V-A). At some level, this is unsurprising since these defenses are designed to protect individual training records, not distributional properties, and we show a connection between how well a model generalizes to the task distribution and its susceptibility for distribution inference leakage (Section V-B). We then develop a simple and inexpensive mitigation based on data re-sampling (Section V-C) which can protect against distribution inference in most cases where an adversary knows the statistical property to protect.

II. Attacks

Before introducing our attacks, we summarize the formal definition of distribution inference we use (Section II-A). Then, we describe previous and our new attack in the black-box (Section II-B) and white-box (Section II-C) settings.

A. Defining Distribution Inference

The goal of distribution inference is to infer sensitive properties of the distribution used to train a machine learned model given some level of access to that model. Mahloujifar et al. [3] formalized a notion of distribution inference that can capture the proportion of a dataset satisfying some property. In this setting, each data-point can either have an attribute value (linked to the property, such as 'female') equal to zero or one, and sampling data from $\mathcal{D}_+$ and $\mathcal{D}_-$ corresponds to data with this attribute set to zero or one respectively. The two distributions to distinguish, then, correspond to a mixture of

these distributions for some α corresponding to the probability of sampling a record from $\mathcal{D}_+$:

$$\alpha \cdot \mathcal{D}_+ + (1-\alpha) \cdot \mathcal{D}_-$$

Using this definition, the authors focus on the task of distinguishing between distributions with different α values, like ones with 40% ($\alpha = 0.4$) and 50% ($\alpha = 0.5$) females. Suri and Evans [6] generalize this notion by using distribution transformation functions $\mathcal{G}_0$ and $\mathcal{G}_1$ to transform an underlying distribution $\mathcal{D}$ (which essentially corresponds to domain knowledge, such as the distribution of face images), instead of explicitly assuming binary attributes and the two distributions being mixtures with different α values. When considering proportional properties (related to α, as described above), the transformers can correspond to sampling records with a given attribute value for any two valid probabilities, α_0 and α_1. As explained via example by the authors: consider $\mathcal{D}$ as a distribution of emails with labels for spam/ham. "$\mathcal{G}_0$ is applied over $\mathcal{D}$ to yield a modified distribution $\mathcal{G}_0(\mathcal{D})$ with 0.8 probability of sampling an email that has negative sentiment, i.e. a dataset sampled uniformly at random from this distribution would have approximately 80% of the emails in it with negative sentiment. Similarly, $\mathcal{G}_1(\mathcal{D})$ could be another distribution with this probability as 0.5 (equally likely to be positive or negative)." α_0 and α_1 here would thus be 0.8 and 0.5 respectively.

We follow this formalization, since this is more generic and better captures assumptions like access to data from some underlying distribution $\mathcal{D}$. The adversary's task in this setting thus corresponds to distinguishing models trained on data from $\mathcal{G}_0(\mathcal{D})$ and $\mathcal{G}_1(\mathcal{D})$, given access to a dataset sampled from $\mathcal{D}$ and some level of access to the trained model.

B. Black-Box Attacks

Black-box attacks assume the adversary has the ability to submit inputs to the trained model and observe the response but does not have direct access to the model. In addition, the adversary has access to some representative data from some distribution $\mathcal{D}$, and seeks to infer which of the transformations, $\mathcal{G}_0$ or $\mathcal{G}_1$, corresponds to the victim's training distribution. Using knowledge of $\mathcal{D}$ and the transformation functions $\mathcal{G}_0$ and $\mathcal{G}_1$, the adversary is able to train shadow models locally. Knowledge of the candidate distributions is necessary to be able to distinguish between them, and it is reasonable to assume an adversary with enough computational resources to train models locally. Most research assumes that the victim and adversary use the same model architecture (e.g., [7], [10]), and that the adversary has access to model prediction confidence vectors (e.g., [7], [11], [12]). In Section IV, we consider settings where the adversary has less knowledge of the victim model and the model API only outputs labels. Next, we review previous black-box distribution inference attacks, and then introduce our new KL Divergence Attack.

Prior Work. Zhang et al. [7] propose meta-classifier attacks that use probability vectors from models for a specific set of query points. Similar ideas are explored in related tasks [11]. The attack works by collecting model predictions for a fixed set of query points (chosen at random); using local shadow models to train a meta-classifier on these concatenated predictions, and finally generating predictions for unseen models using the meta-classifier. Suri and Evans [6] propose the Loss Test and Threshold Test attacks that compare model accuracies on candidate distributions to predict training distributions. The Threshold Test performs best of these: it uses locally trained models to derive a threshold on observed accuracy on a given data sample, which is then used to predict the training distribution of a model. This attack yields non-trivial inference accuracies in many cases but falls short of the white-box attacks by huge margins for most settings. These attacks have also been extended to settings where active adversaries that can poison the victim's training data [13]. The only previous attacks designed to work with label-only predictions are by Juarez et. al. [14] that performs a statistical test based on attribute-wise model performance, and Mahloujifar et al.'s attack in the setting of active adversaries [3].

KL Divergence Attack (KL). Recent work by Hartley et al. [15] demonstrates how the presence of unique features, even if present in one training record, can impact output probability distributions. Motivated by their use of KL divergence to differentiate between the two scenarios (instance present or not), we propose an attack that compares the KL divergence in output probabilities of the victim model using local models.

The adversary prepares by training a collection of local models $\{M_0^1, M_0^2, ...M_1^1, M_1^2, ...\}$, where M_0^i and M_1^i (for some i) denote models from training distributions $\mathcal{G}_0(\mathcal{D})$, $\mathcal{G}_1(\mathcal{D})$ respectively. Let X denote some data randomly sampled by the adversary from the distributions $\mathcal{G}_0(\mathcal{D})$ and $\mathcal{G}_1(\mathcal{D})$, with an equal number of samples ($|X|/2$) from both distributions. We first define a way to estimate the KL-Divergence between two models using predictions:

$$\mathbb{E}[D_{KL}(N \parallel M)] = \mathbb{E}_{x \in X}\left[\sum_{c \in \mathcal{C}} N(x)_c \log\left(\frac{N(x)_c}{M(x)_c}\right)\right] \quad (1)$$

where $M(x)_c$ corresponds to the prediction probability corresponding to class c (out of all classes $\mathcal{C}$) for some point x for model M, and the expectation $\mathbb{E}[]$ is taken over the adversary's data X. We use the same data X in computing KL-Divergence values. Next, the adversary defines a "weighted vote" for a pair of models (N, P) with respect to M:

$$\lambda(M, N, P) = \mathbb{E}[D_{KL}(N \parallel M)] - \mathbb{E}[D_{KL}(P \parallel M)]. \quad (2)$$

A positive quantity $\lambda(M, N, P)$ thus indicates that the model M has its predictions distributed closer to P than N, since a lower KL-divergence between distributions indicates higher similarity. Using its collection of local models trained on the two candidate distributions, the adversary then computes and aggregates this "weighted vote" across all pairs of its local models (M_0^i, M_1^j):

$$\hat{b} = \mathbb{I}\left[\sum_i \sum_j \lambda(M, M_0^i, M_1^j) > 0\right] \quad (3)$$

The rule above thus effectively checks all its pairs of local models and compares similarities in prediction distributions with a given victim model. Since the core idea here is to compare distributions of model predictions, other metrics to compare distributions, like Jensen-Shannon Divergence, or TV Distance, can be used instead of KL-Divergence.

C. White-Box Attacks

In the white-box setting, the adversary additionally has direct access to the victim's model including its trained parameters. Although this access model assumes a stronger adversary, it is a realistic adversary for many scenarios, like when models are deployed on client devices. It is also useful in two ways: 1) gauging the extent of inference leakage, helping bound risk and understand it better, and 2) studying patterns and trends across properties and models to help better understand distribution risk and come closer to inventing effective defenses.

Prior Work. The main previous white-box distribution inference attacks are based on permutation-invariant networks [2]. These attacks assume that information related to the training distribution can be somehow extract from trained model parameters. They look at model parameters across all layers of a multi-layer perceptron to generate a feature representation for the entire model that is insensitive to arbitrary reorderings of neurons. The method works by constructing feature representations for each layer using learnable parameters (with prior layers as context) using shadow models, and trains the meta-classifier via back-propagation, using labels indicating which training distribution the shadow models correspond to. These attacks, originally designed for networks with linear layers, have been extended to support convolutional layers [6]. Other distribution inference attacks in the literature follow a similar meta-classifier approach: using parameter extraction for support vector machines [1], using model gradients [10], or intermediate node embeddings in graph neural networks [5].

III. Results

We evaluate our proposed attacks on several datasets, including both established benchmarks used in prior work and new configurations and property-task combinations previously unexplored in the literature. Code for reproducing our experiments is available at https://anonymous.4open.science/r/dissecting_distribution_inference-0B2F/ (anonymized link for reviewing, will be in public github repository).

Our new attacks are significantly more potent than the previous state-of-the-art. Our KL Divergence Attack (KL) outperforms all previous black-box attacks by huge margins (Section III-C). Even more interestingly, the KL Divergence Attack, with only black-box access, outperforms Permutation Invariant Networks (PIN) by a large margin in nearly all settings. We study trends between the correlation of the task and property, and its impact on inference risk (Section III-D).

A. Datasets and Models

We evaluate our attacks on twelve task-property pairs across five datasets, summarized in Table I. These were selected to directly compare results with previous works (RSNA Bone Age, ogbn-arxiv, CelebA), to study the impact of task-property correlation on inference risk (various property-task pairs for CelebA), and to include datasets representing real-world use-cases, like Census19 and Texas-100X.

Census19 [16] is an updated and expanded version of the Adult Census dataset [17] based on data from the US Census Bureau. It contains a mixture of numerical and categorical features, and the same prediction task. We focus on the ratio of whites (race) and females (sex) as properties, and use a two-layer feed-forward neural-network as the architecture.

Texas-100X [18] contains demographic and medical information for patients across hospitals. The original dataset uses 100 possible classes for surgical procedure prediction. We slightly modify the task and focus only on data from the top 20 classes, reducing it to a 20-class classification task. We focus on the ratio of whites (race), females (sex), and Hispanics (ethnicity) as properties, and use a two-layer feed-forward neural-network.

CelebA [19] contains collections of face images of celebrities. Each image is annotated with attributes. We use three different tasks: smile detection, gender prediction, and mouth-open prediction. We conduct experiments with a convolutional neural network trained from scratch for this dataset, with five convolutional layers and pooling layers followed by three linear layers, which is the smallest network we could find with reasonable task accuracy. For our experiments with feature extractors, we also conduct experiments where the adversary uses a pre-trained FaceNet [20] model trained on the CASIA-WebFace [21] dataset, with a two-layer network. It leads to a drop in performance (from $\sim 92\%$ to $\sim 82\%$), but the point of such an experiment is indeed to assess inference risk in more practical settings. For the attack inference properties, we use the proportion of females (smile-detection task), old people (gender-prediction task), people with wavy hair (mouth-open-prediction task), and people with high cheekbones (mouth-open prediction task). These pairs are useful in comparing results with previous works, and also help cover a spectrum of different correlations between the task and property attributes.

RSNA Bone Age [22] contains x-ray images of hands, and the standard task is to predict the patient's age in months. We convert the task to binary classification based on an age threshold (> 132 months), and focus on the ratios of the females (available as metadata) as properties. We also consider a flipped scenario, where the task is to predict females, with the ratios of people below the age threshold as properties. We use a pre-trained DenseNet [23] model for feature extraction, followed by a two-layer network for classification. Similar to CelebA, we consider a setting where the adversary uses pre-trained feature extractor, while the victim trains models from scratch. Additionally, we also consider a setting where both the victim and adversary use the same feature extractor, but use different model architectures on top of the feature extractor

ogbn-arxiv [24] is a directed graph of citations between computer science arXiv papers, with the task as predicted subject area categories (out of 40) using features extracted from paper documents. We infer the mean node-degree property of

Dataset	Task/Property	Distinguishing Accuracy for $\alpha_1 = 0.2$				Mean Distinguishing Accuracy			
		Black-Box			White-Box	Black-Box			White-Box
		TT [6]	ZTO [7]	KL	PIN [2]	TT [6]	ZTO [7]	KL	PIN [2]
Census19	Income/Females	61.3	54.4	**82.5**	81.0	50.0	53.4	**89.8**	78.6
	Income/Whites	59.4	54.9	**83.8**	75.4	53.2	52.6	**92.4**	74.2
Texas-100X	Procedure/Females	51.2	51.6	**82.5**	51.3	50.0	50.0	**89.3**	50.0
	Procedure/Whites	52.4	50.1	**81.6**	50.6	50.9	50.0	**86.8**	50.0
	Procedure/Hispanic	50.0	50.0	**82.4**	50.1	50.0	50.0	**78.4**	50.0
CelebA	Mouth Open/Wavy Hair	50.6	52.3	**62.1**	56.3	52.0	51.8	56.8	**92.0**
	Smile/Females	50.9	56.2	**89.6**	75.0	55.4	60.9	**85.3**	70.1
	Gender/Young	52.9	55.5	**86.3**	81.2	50.3	52.6	**86.4**	81.0
	Mouth Open/High Cheekbones	50.1	56.2	76.7	**86.1**	50.0	50.0	**84.6**	83.0
RSNA Bone Age	Age/Females	64.0	77.9	**95.2**	94.5	90.0	95.4	**100.0**	99.4
	Females/Age	68.5	78.5	**99.8**	75.2	95.7	99.4	**100.0**	66.0
ogbn-arxiv	Node classification/Mean Degree	50.0	55.4	**92.6**	71.9	50.1	50.1	**100.0**	87.4

TABLE I: Effectiveness of inference attacks. We show results for our KL Divergence Attack (KL) and three prior attacks: Threshold Test (TT) [6], ZTO [7], and Permutation Invariant Networks (PIN) [2]. For the classifiers, the first set of results shows the attack's ability to distinguish between models trained on training sets where the proportion of the property is either $\alpha_0 = 0.5$ or $\alpha_1 = 0.2$ as an accuracy percentage. The second set of results shows the mean distinguishing accuracies (%) between $\alpha_0 = 0.5$ and a set of varying α_1 values $(0.0, 0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9, 0.1)$. For the graph datasets used for ogbn-arxiv, for the first set of results we use $\alpha_0 = 13$ are $\alpha_1 = 10$ as the two distributions; for the second set, we vary the mean node degree as the property, setting $\alpha_0 = 13$ and varying α_1 in $[9, 10, 11, 12, 14, 15, 16, 17]$, and report the mean distinguishing accuracy. For all of the results, for each α_1 value, we compute the median over five trials. The mean distinguishing accuracy is then computed over the mean of these values for all α_1 values. For each setting, results for the most effective attack are bolded.

the graph, and use Graph Convolutional Networks [25].

B. Experimental Setup

We build upon the experimental setup described in earlier works on distribution inference, using implementations and trained models provided by previous works [2], [6]. For each dataset, we create non-overlapping splits of data for the victim and adversary. For each dataset, we simulate $\mathcal{D}$ using the dataset itself. $\mathcal{G}_0(\mathcal{D})$ is simulated by sampling from the dataset, such that the resulting distribution has $\alpha = 0.5$ (or 13, for ogbn-arxiv), while $\mathcal{G}_1(\mathcal{D})$ is simulated for some α (which we vary across experiments). We include the datasets and victim/adversary splits used in previous experiments, and include results on two new datasets. For all of the experiments, we follow the processing pipeline described in Suri and Evans [6] to obtain non-overlapping splits. Essentially, both parties sub-sample from their data splits (to achieve specific α values) with different random seeds, and train models on the sampled data.

We perform each experiment five times and report mean values with standard deviation in all of our experiments. Full experimental details are provided in the Appendix A.

KL Divergence Attack. Since using all pairs of adversary's models can be expensive, the attack uses a set fraction (0.8) of randomly chosen pairs to compute the expectation in Equation 3. We experiment with multiple values of this fraction, and observe comparable performance. For each pair of local models, the attack collects the difference in KL values. These differences are then normalized across all differences observed for local models, after which the adversary uses voting-based aggregation to generate the final prediction. We also experimented with variants that do not include voting, as well as ones that flip the inputs to KL-Divergence computation (so $D_{KL}(B||A)$ instead of $D_{KL}(A||B)$), but find the current version to perform best.

C. Results

Table I summarizes the results of our distribution inference experiments. For each experiment, we report the distinguishing accuracy between two distributions as well as the mean distinguishing accuracy across a set of different distributions, as detailed in the table caption. For the classifiers, we vary α_1 in $[0.0, 1.0]$ at intervals of 0.1, and set $\alpha_0 = 0.5$ for the case of ratio-based properties, where a certain α value for a distribution means datasets sampled uniformly at random would have α fraction of the data with the property attribute 1, for e.g. ratio of females. The distinguishing accuracies thus correspond to predicting whether a model has the training distribution corresponding to α_0 or α_1 where random guessing would be 50% accuracy, and perfect predictions would be 100%.

The majority of experimental evaluations in the literature follow a binary classification scenario, where the adversary is supposed to distinguish between two potential training distributions with property values α_0 and α_1. Although regression-based adversaries have been demonstrated [6] as being strictly more powerful, they are much more computationally expensive and we leave evaluations in that scenario to future work.

Trends across datasets. Inference leakage varies significantly across different datasets, with very little leakage for most cases in Texas-100X, substantial leakage for Census19, and

Dataset/Task	Number of Shadow Models				
	5	10	25	100	400
Census19 (Sex)	73.0	76.5	81.3	86.4	89.7
Census19 (Race)	77.2	79.3	81.3	84.2	84.7
RSNA Bone Age (Age)	97.3	98.3	99.3	99.7	99.8
CelebA (Sex)	74.0	78.6	80.9	86.9	89.2

TABLE II: Impact of varying the number of shadow models used by the adversary per distribution to launch its attacks. Values are mean distinguishing accuracies (%) for KL Divergence Attack(computed as described in Table I). Even with only 5 models, the adversary is able to achieve considerable inference accuracy.

exceptionally high leakage for the graph-based ogbn-arxiv dataset. The lack of virtually any inference risk in Texas-100X is surprising, as the features contain the property label, and data splits are processed per hospital during generation, making the victim and adversary distributions highly similar. This difference in inference risk between Census19 and Texas-100X, despite both being tabular datasets, reveals how just the nature of data (tabular, images) does not by itself determine inference risk and risk can vary unpredictably (at least based on current understanding) with aspects of the data. As previously observed by Suri et al. [12], leakage is quite high for RSNA Bone Age. Our new improved attacks identify vulnerable datasets, such as Census19, that would have been considered low leakage risks using previous state-of-the-art attacks.

Comparing black-box attacks. The KL Divergence Attack outperforms Threshold Test (TT) and the black-box attack by Zhang et. al. [7] (which we refer to as ZTO) in all cases with large margins. Across all of the settings, TT and ZTO rarely achieve distinguishing accuracies above 75%, whereas the KL Divergence Attack produces meaningful leakage for all of the datasets. The superiority of the KL Divergence Attack can be attributed to the use of pairs of local models and their trends (which grow in the order $\binom{n}{2}$ for n models), as opposed to using information from models in isolation in the other attacks.

Number of shadow models. The black-box attacks use 50 shadow models per training distribution. We vary this number to 1) get an empirical lower bound on the number of shadow models required to achieve non-trivial leakage, and 2) study increase in information leakage with an increase in shadow models. Leakage is significant with only five shadow models per distribution and in most cases, and improves with more local shadow models (Table II).

White-box attacks. The black-box KL Divergence Attack performs surprisingly well despite the weaker threat model, outperforming the best white-box attack in nearly all experimental settings. Since an adversary in the white-box setting has access to more information than just the data and model predictions, it should be at least as powerful as a black-box adversary. We attribute the relative ineffectiveness of the white-box attacks to two main reasons. First, in Permutation Invariant Networks, model parameters are directly used as features for the meta-classifier, unlike comparisons in model prediction distributions in KL Divergence Attack. Secondly, white-box attacks have a larger feature space and learning meta-classifiers additionally requires learning to recognize relevant patterns in model parameters, a huge and complex data distribution. The black-box attacks, on the other hand, are agnostic to parameters in the victim model and thus much easier to scale, resulting in better performance.

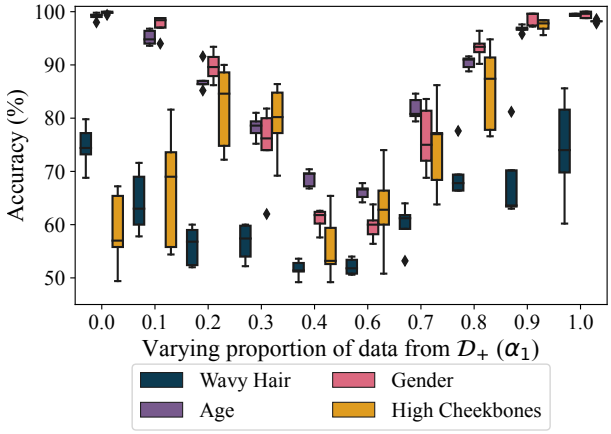

Fig. 1: Distinguishing accuracy for different task-property pairs for CelebA with varying correlation, for KL Divergence Attack.

D. Correlation

The impact of correlation between the underlying task of a model and the property of its training distribution being inferred has been touched upon briefly in the literature [7], but not studied extensively. Intuition suggests there should be some positive relationship between inference risk with increasing task-property correlation, but prior studies do not evaluate inference risk across a range of property-task correlations. We carefully pick pairs of properties and tasks for the CelebA dataset, such that there is a good range of correlations.

We conduct experiments with property correlations of ≈ 0 (Mouth Slightly Open–Wavy Hair), ≈ 0.14 (Smiling–Female), ≈ 0.28 (Female–Young), and ≈ 0.42 (Mouth Slightly Open–High Cheekbones). Across this range of correlations, mean distinguishing accuracies are 62.1%, 85.3%, 86.3%, 76.7% as correlation values increase. The lack of any clear trend between correlation and inference risk is consistent with observations in the literature around task-property correlation and inference risk [7]. As observed, inference risk is non-trivial as long as the task-property correlation is non-zero. While the case of non-zero inference risk is obvious (with zero correlation, loss optimization would not use the property as an indicative feature), changes in inference risk with varying correlation values may be tied to observed correlation versus actual causality, and methods for causal learning may help alleviate this inference risk [26].

We perform similar analyses for the RSNA Bone Age dataset, where we flip the property and task. In this case, the correlation between the task and property remains the same,

thus helping identify potential changes in inference leakage arising purely from the choice of the property itself. While switching from Age–Females to Females–Age, we observe a huge bump in mean distinguishing accuracies: from 95.2% to 99.8%. Although the choice of property and task are expected to impact inference risk, our results suggest that this choice may be more relevant to evaluating inference risk than property-task correlation itself.

IV. Impact of Adversary's Knowledge

Research on inference privacy typically considers threat models with one of two simplistic adversarial assumptions: white-box settings, where the adversary has full access to the model; and black-box settings, where the adversary has only API access to the model but receives full confidence vectors for each prediction and has complete knowledge of aspects of the training process and model architecture. The specific information available to an adversary in the black-box setting, however, may have a significant impact on inference risk. For instance, access to labeled data (for attacks) with prediction probabilities is often implicit, as is the use of the same model architectures and feature extractors between the victim and adversary. We study the impact of these common assumptions, and how relaxing them impacts inference risk. We measure impact on risk when the victim and adversary use different model architectures (Section IV-A), do not share feature extractors (Section IV-B), and when the available model API only provides label predictions (Section IV-C). Inference risk is somewhat robust to differences in model architectures, as long as the victim and adversary's models have similar capacity. The absence of shared feature extractors reduces inference risk significantly, but we find attacks can still succeed when only label predictions are available.

A. Model Architecture

In the white-box setting an adversary can directly observe the target model's architecture, but in black-box settings it is unrealistic to assume the adversary knows the target model architecture. Likely model architectures may be limited in certain domains like image data, where the victim is likely to use a popular model architecture such as DenseNet [23]. But a variety of models like random forests, support vector machines, and clustering-based classifiers can be used for tabular data and may even be picked by model trainers via automated tools [27].

Differences in victim and adversary architectures have not been previously explored, except by Mahloujifar et al. [3] for poisoning-based adversaries. In their setting, the victim and adversary can have different model architectures—the adversary uses logistic regression while the victim can use a variety of different feed-forward neural networks. They note a drop in inference risk with an increase in victim model complexity. It is thus unclear if these trends are specific to the model architecture themselves. Additionally, the adversary's model architecture is kept the same, so they did not explore the potential for higher inference risk with better local models.

Victim Model	**Adversary Model**			
	RF	LR	MLP_2	MLP_3
Random forest (RF)	95.1	78.9	86.7	85.6
Linear regression (LR)	93.2	100	76.4	80.8
Two-layer perceptron (MLP_2)	69.7	56.6	82.5	82.7
Three-layer perceptron (MLP_3)	69.3	56.3	82.2	81.1

TABLE III: Variation by model type. Each values is the observed mean distinguishing accuracy (%) (measured as described in Table I) of the KL attack for Census19 (Sex), for different combinations of model types for victim and adversary.

To identify trends in inference risk with differences between architectures, we train multiple models with different architectures for both the victim and adversary. For Census (Gender), we try all possible combinations out of linear regression (LR), multi-layer perceptrons with two and three layers (MLP_2, MLP_3), and a random forest classifier (RF). We also consider using a two-layer perceptron (MLP_2) and a support vector machine (SVM) for the case of RSNA Bone Age (Gender). For this experiment and the rest of this section, we report results with KL Divergence Attack.

We observe several interesting trends for Census19 while varying model types for the victim and adversary (Table III). For instance, inference risk is significantly higher when the adversary uses models with learning capacity similar to the victim, like both using one of (MLP_2, MLP_3) or (RF, MLP). Concretely, mean distinguishing accuracy is 84.5% when learning capacities match, as opposed to 72.7% when learning capacities do not match. Interestingly though, we also observe a sharp increase in inference risk when the victim uses models with low capacity, like linear regression and random forest instead of multi-layer perceptrons. For example, mean distinguishing accuracy is 72.5% when victim models have high learning capacity (MLP_2, MLP_3), but increases to 87.1% when the victim models have low learning capacity (RF, LR). These trends hint at possible connections between distribution inference risk and model learning capacity.

B. Feature Extractors

When dealing with high-dimensionality datasets and a scarcity of data, it is common to use techniques such as transfer learning [28], [29] to boost model performance with reduced data and computational requirements. Using a pre-trained model for feature extraction should intuitively limit distribution-related privacy leakage, since there are fewer trainable parameters that can potentially contain revealing information. At the same time, fewer parameters also reduce the space of models, making it easier for an adversary to launch attacks. Even in a black-box setting, the adversary may be able to use the same feature extractor as the victim, either as a result of the adversary snooping and gaining information, or just by assuming the use of popular pre-trained models (like BERT [30]). This setting has been explored previously [1], [2], [6], but the effect of the adversary using the same or different extraction models from the victim model has not been previously explored.

Victim Model	Adversary Model	
	FE+MLP$_2$	FE+SVM
Feature extractor, perceptron (FE+MLP$_2$)	94.5	93.0
Feature extractor, SVM (FE+SVM)	99.5	99.6
DenseNet (CNN)	88.0	94.4

TABLE IV: Mean distinguishing accuracies (%) for RSNA Bone Age (Sex), for different combinations of model types for victim and adversary (as computed in Table I).

For RSNA Bone Age (Sex), we consider two configurations: one where the victim and adversary use the same feature extractor, and another where the victim trains DenseNet [23] models from scratch (CNN). For the first setting, we explore an SVM (FE+SVM) and a two-layer perceptron (FE+MLP$_2$). There is a considerable drop in distinguishing accuracies (from 96.7% to 91.2%) when the victim and adversary no longer share feature extractors (Table IV). For the settings where they do, we observe leakage to be highest for similar model architectures, and note a sharp increase when the victim uses an SVM. Nonetheless, inference risk stays sufficiently high. Interestingly, for the case where feature extractors are not shared, using a lower learning-complexity model (FE+SVM) seems to lead to higher leakage, than FE+MLP$_2$. While leakage is high in both cases, the increase can be explained by the chances of adversary's local models overfitting being less than that with an MLP.

For CelebA (Sex), we explore a setting where the adversary utilizes a feature extractor to train its models, while the victim trains CNNs from scratch. We observe a similar diminishing of inference risk when a shared feature extractor is not available to the adversary, consistent with the RSNA Bone Age results. Compared to the scenario where the adversary uses the same model architecture as the victim without any pre-trained feature extractors, mean distinguishing accuracy drops from 85.3% to 71.0%.

C. Label-Only Access

Most black-box attacks in the literature related to distribution inference assume access to prediction confidence vectors. This is not an unreasonable assumption—many APIs return prediction scores, especially for top-k classes (for example Google Vision API and ClarifAI Prediction API return scaled confidence scores for the top 10 or 20 classes, respectively). It is unclear, however, what kind of performance drops to expect for distribution inference attacks in settings where the model's API only returns a label. The only previous works to explore distribution inference in the label-only setting are in the context of group distribution shift auditing [14], and active adversaries with poisoning capabilities [3].

With some straightforward modifications, KL Divergence Attack can be launched with access to just label predictions, with negligible drops in inference leakage in most cases. The attack requires prediction confidence scores to compute the KL-divergence values. However, these scores are absent in the label-only setting, and the labels effectively correspond to confidence values of 0 and 1. This makes the KL computations (Equation 1) invalid, since the log of 0 or $1/0$ is undefined. To tackle this, we replace $0/1$ labels with confidence scores ϵ and $1-\epsilon$ for some small value ϵ (set to 0.01 in our experiments).

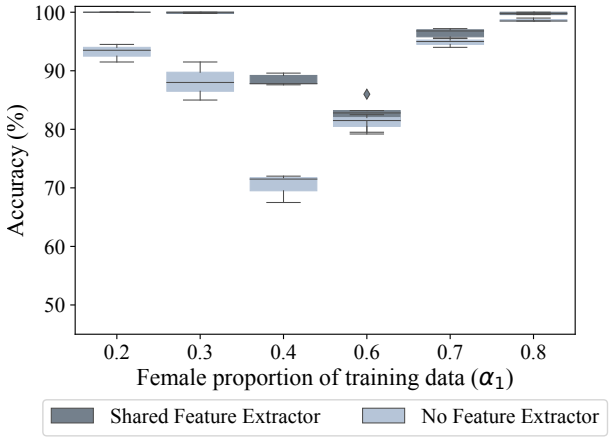

Fig. 2: Distinguishing accuracy for the KL Divergence Attack for RSNA Bone Age (Sex), when the adversary uses the same feature extractor as the victim, and when the victim does not use or share any pre-trained feature extractor. While there is an obvious drop in performance, inference risk still stays high.

Dataset/Task	Confidence Scores	Prediction Label	
		Direct	Sampling
Census19 (Sex)	82.5	77.3	80.5
CelebA (Sex)	85.3	78.0	79.3
RSNA Bone Age (Age)	99.8	96.3	97.1

TABLE V: Effectiveness of label-only attacks. Each value is the observed mean distinguishing accuracy (%) for the attack (as computed in Table I). The label-only setting leaks less information, but the attacks still are effective even when confidence scores are unavailable. 'Direct' uses a single query, while 'Sampling' uses 10 samples around each test point.

We observe mixed trends across datasets and attacks. For instance, switching to the label-only setting has little impact in the case of Census19, while mean distinguishing accuracies drop by more than 8% for CelebA (Table V). However, the drop in performance for CelebA is not uniform across all ratios. Inference risk is still quite high for many values of α (Figure 3). Similar trends hold for RSNA Bone Age, where distinguishing accuracy is $> 75\%$ for all ratios. We also experiment with using probabilistic sampling to extract more information. For each datapoint, we sample k random points in its neighborhood by adding random noise from $\mathcal{N}(0, \sigma^2)$ to each feature, and average the generated label predictions to estimate confidence scores. We observe slight improvements in attack performance from the sampling, at the cost of additional queries.

V. Defenses

Several defenses against distribution inference have been proposed, but most of them (except differential privacy [2], which has shortcomings as we discuss later) have not been

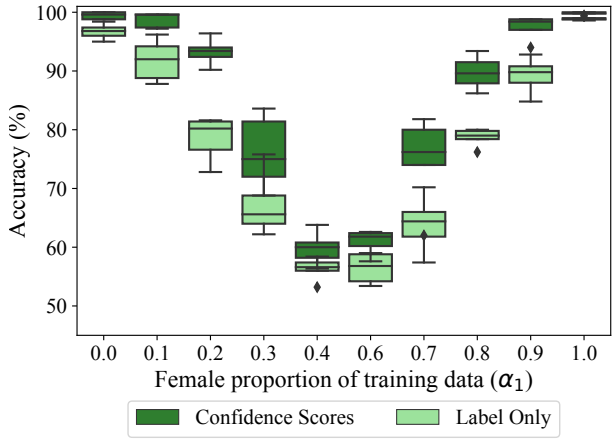

Fig. 3: Comparing the distinguishing accuracy for the KL Divergence Attack for CelebA (Sex), when the target model returns prediction confidence scores and when it returns only prediction labels. Performance drops most for certain ratios like 0.2 and 0.8, but remains high and roughly the same for more extreme ratios like 0.0, 0.1, and 1.0.

actually evaluated. Like most defenses designed to limit privacy leakage, these defenses involve adding noise in some parts of the training process. This can include the data itself [1] or model parameters [2], [10]. One notable exception is work by Hartmann et. al. [31], where the authors study causes of leakage in distribution inference attacks, and evaluate mitigation strategies based on causal learning (IRM [32]), correcting inductive biases, and increasing the amount of training data, for synthetic datasets. We evaluate some of these noise-based defenses in Section V-A, and find that they seem unlikely to successfully mitigate distribution inference risks. Our exploration of inference risk with model generalization reveals interesting trends and a potential trade-off between learning and inference risk (Section V-B). Section V-C introduces and evaluates a simple defense based on data re-sampling, which can prevent distribution inference in settings where the model trainer knows which distributional property to hide.

Prior Work. Unlike membership inference where differentially private training can provide a guaranteed bound on inference risk, there are no defenses against distribution inference from trained models with theoretical guarantees. Chen and Ohrimenko [33] recently proposed a defense mechanism that builds upon formal notions of distributional privacy [34] to protect against property inference attacks on statistical queries. This is the first known theoretically-grounded defense against distribution inference, but it does not apply to protecting machine-learning models.

The only previous defense that has demonstrated meaningful protection against distribution inference attacks on machine learning models is NoSnoop [35], proposed for a collaborative learning setting. In their threat model, the adversary seeks to infer sensitive information about exact training batches and has access to intermediate model losses from clients. The defense works utilize a discriminator-generator setup, where gradient updates are used to minimize property leakage while preserving task-based utility. Although this defense is highly effective, it defends against a very narrow type of configuration, including properties limited to the presence/absence of sensitive data.

Recent work by Stock et al. [36] proposes a defense against distribution inference based on gradient updates from meta-classifiers. The victim, being aware of two distributions that an adversary may test for, trains multiple copies of its models with these training distributions. Then, it trains a meta-classifier and computes gradient updates for its own local models such that inference risk is minimized. This method requires the victim to train hundreds and thousands of models locally for the meta-classifier, leads to large drops in task performance, and does not generalize to settings where the victim and adversary use different kinds of meta-classifiers.

Other proposed defenses include removing sensitive attributes from features [7], using node-multiplicative transforms, or encoding arbitrary information into the models [2]. Since black-box attacks only utilize relationships between inputs and model outputs, they are unaffected by such changes as long as model functionality remains unaffected. Additionally, the Permutation-Invariant Network architecture can be modified to have some form of scale-invariance as well, which in turn can also bypass such multiplicative defenses. Further, these defenses seem unlikely to diminish black-box attacks, which our experiments have shown to be more effective than known white-box attacks, hence we do not evaluate them here.

A. Noise-Based Defenses

Several proposed defenses against distribution inference involve adding noise in various ways—differentially privacy training incorporates crafted noise in the training process and label poisoning adds noise to the training data. We also consider using adversarial training, which augments training with adversarial perturbations.

Differentially Private Training. Differential privacy (DP) is a formal privacy notion that provides theoretical guarantees that bound an adversary's ability to distinguish between neighboring input datasets from the output of a computation. Differential privacy can provide theoretical bounds limiting membership inference. Evaluations by Ateniese et al. [1] suggest differentially private training is not an effective defense against distribution inference attacks. However, their experiments used a setup with some overlap between the victim and adversary's data, so it is possible the observed lack of protection is related to the overlapping data available to the adversary. Although differential privacy in itself does not guarantee protection against distribution inference, evaluating risk for models trained with these guarantees can help better understand how such noise-based mechanisms can affect inference risk, and assess the vulnerability of models meant to provide membership inference privacy. Empirical evidence can thus beneficial and more concrete than relying on pure intuition (or argumentative reasoning about why a defense may or may not work).

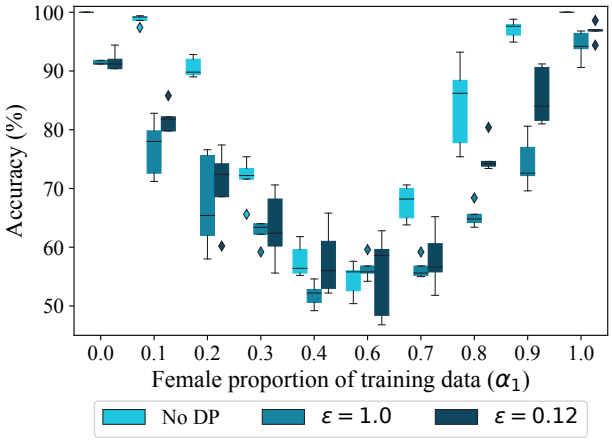

Fig. 4: Distinguishing accuracy for different for Census19 (Sex), using KL Divergence Attack. Attack accuracy drops with stronger DP guarantees (decreasing privacy budget ϵ).

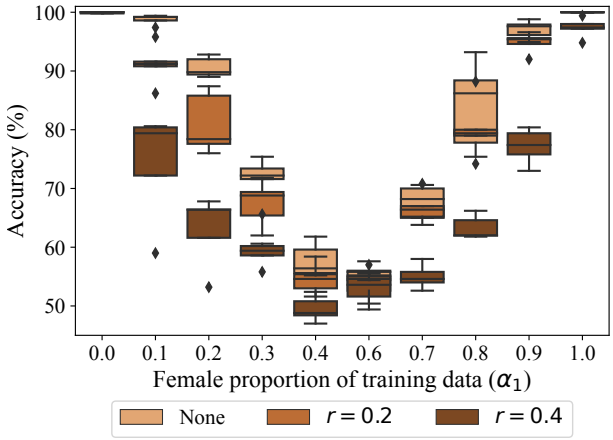

Fig. 5: Distinguishing accuracy for different for Census19 (Sex), for varying levels of label poisoning. Inference risk drops considerably with increasing levels of label poisoning, but is also followed with non-trivial drops in task accuracies.

We use DP-SGD [8] to train victim models with Reńyi Differential Privacy [37], with privacy loss budgets of $\epsilon = 1.0$ and $\epsilon = 0.12$, with $\delta = 4.9 \times 10^{-6}$. We evaluate this defense on Census19, since it is the only tabular dataset with non-trivial inference leakage. We observe a drop in distinguishing accuracies, but inference risk stays high for ratios further away from $\alpha_0 = 0.5$ (Figure 4).

We hypothesize that this drop in effectiveness may not be because of the differential privacy noise itself, though, but could be because either the model does not learn the distribution well enough (and hence does not reveal it), or it produces arbitrary differences that cause a mismatch between the victim's models trained with using DP-SGD and the adversary's shadow models trained without privacy noise. Inspection of task accuracy for the differential-privacy models suggests lower learning effectiveness as one potential factor (Table VIII). To test whether the decrease in prediction accuracy is mainly due to arbitrary differences in the models, we evaluate results for the setting where the adversary also trains its models using DP-SGD with the same privacy loss budget. Compared to an adversary that does not use DP, there is a clear increase in inference risk—mean distinguishing accuracy increases to 86.4% for $\epsilon = 1.0$, and 91.5% for $\epsilon = 0.12$ (compared to 82.5% without any DP).

Assuming adversary's knowledge of the use of differentially-private training and the specific privacy loss budget is not a far-fetched assumption. Organizations that release differentially private models often document their exact levels of privacy budget [38], [39]. An adversary in such scenarios can thus train its models with the same privacy parameters.

Label Poisoning. Ganju et al. [2] proposed to mitigate distribution inference by adding noise to the training data via label poisoning. The underlying idea is to perturb the training data in a way that will alter the model parameters and make the adversary's task harder. Although changing data labels is prone to detrimentally harming the model's task performance, a model trainer may be able to find an acceptable trade-off between accuracy and inference risk. For a given noise ratio r, the defense comprises randomly flipping task labels for r fraction of the training data. We evaluate this defense for CelebA (Male) and RSNA Bone Age (Age) with a label noise ratio of 0.2, and Census19 (Gender) for label noise ratios 0.2 and 0.4. As expected, this defense harms task performance (Table VIII), reducing task accuracy: by $\sim 1-2\%$ for $r = 0.2$ for all three datasets, and $\sim 3\%$ for $r = 0.4$ off Census19. Average inference risk drops for Census19, but remains is still quite high for ratios like $\alpha_1 < 0.2$ and $\alpha_1 > 0.8$, as shown in Figure 5. It may be possible to find a desirable tradeoff for a simple task like Census19, but this approach is not effective for more complex tasks. For instance, using a label noise ratio of 0.4 in CelebA completely destroys task performance, reducing the classifier to only slightly better than random guessing.

Adversarial Training. Adversarial training [9] involves using a training loss function that encourages the model to learn features that are robust to perturbations in the input, and produces models that are less prone to overfitting dataset-specific patterns [40]. This can be especially useful when the data includes properties that are not correlated with the task, and a model should not capture signals related to the irrelevant property. A model trained with adversarial robustness objective, using this reasoning, should be less susceptible to distribution inference. To test this hypothesis and explore the impact of training for robustness. We train adversarially-robust models for varying L_∞ norms for the Gender and Age properties on CelebA, since the other datasets are either tabular or do not contain sufficient samples for adversarial training with acceptable performance. Figure 6 shows distinguishing accuracies for varying settings for the perturbation strength used in adversarial training. Training for adversarial robustness, as documented in the literature, leads to drops in task accuracy.

We observe very interesting trends with respect to inference

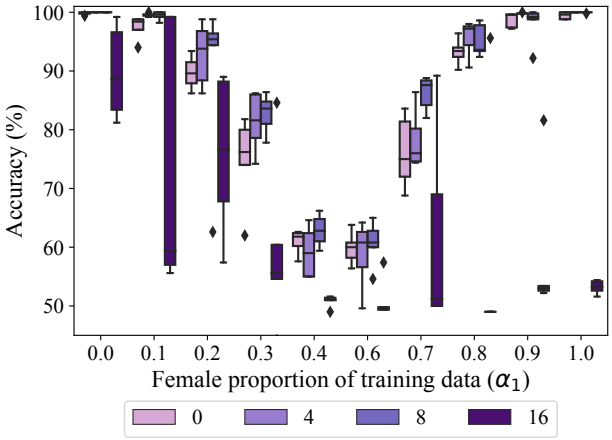

Fig. 6: Distinguishing accuracy for different using KL, for varying levels of adversarial robustness ϵ (/255) in L_∞ norm, for CelebA (Sex). Inference risk lowers with increasing robustness.

Dataset/Task	Adversarial Training (ϵ)			
	0/255	4/255	8/255	16/255
CelebA (Sex)	85.3	86.8	88.3	58.9
CelebA (Age)	86.3	90.0	88.9	83.6

TABLE VI: Impact of adversarial training. Values are mean distinguishing accuracies (%) (as computed in Table I) for KL on models trained with adversarial robustness, with varying norms ϵ (/255) of perturbation budget (L_∞ norm).

risk. Risk increases with increasing perturbation strength (ϵ) until $8/255$, and then drops to near-zero (Figure 6).

Since adversarial training helps models remove focus from spurious correlations, it is naturally aligned with causal inference [41]. This then leads to these models using signals relevant to the property being inferred (like Age or Sex) even more, since it is related to the task. However, as this perturbation norm increases, task accuracy drops accordingly, thus leading to lower inference risk since the model itself performs poorly at learning causal connections, like the one between the property being inferred and the given task. One notable exception here is $\epsilon = 8/255$ for CelebA (Age), where inference risk seems to slightly increase. One possible explanation is the stronger age (property) and sex (task) relationship in this case, leading to the causal relationship-accuracy tradeoff leaning in favor of the former, in terms of inference risk.

B. Generalization

The defenses discussed in Section V-A have one thing in common: they nearly always lead to non-trivial drops in task performance. This is not only unacceptable for most deployments, but raises the question of whether the defenses are doing anything useful or just reducing distribution inference by producing models that learn the underlying distribution less well. Concretely, we observe a positive correlation between model task accuracy and mean distinguishing accuracies

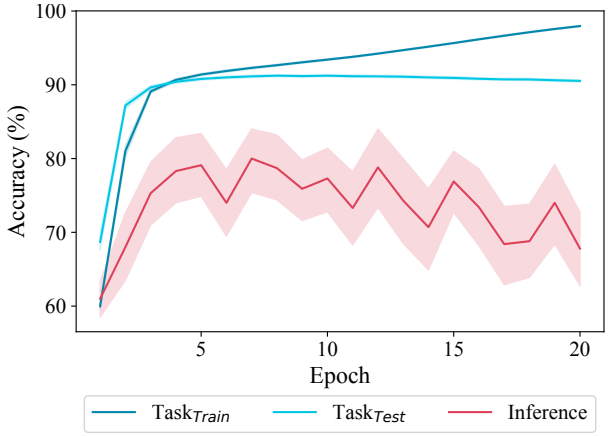

Fig. 7: Mean distinguishing accuracy (as computed in Table I) of the KL Divergence Attack on CelebA (Sex), for varying number of training epochs for victim models. Shaded regions correspond to error bars. Distribution inference risk increases as the model trains, and then starts to decrease as the model starts to overfit.

(Pearson's correlation coefficient > 0.7). A model with poor task performance possibly fails to learn useful signals from the training distribution, and is thus cannot leak properties it has not learned; while good performance means a model has learned the distribution well and is prone to more leakage.

To investigate these correlations, we inspect trends between inference risk and generalization across training epochs. For this experiment, we train the models longer than the other experiments (which were optimally selected for best generalization using validation data), allowing us to better study trends between overfitting and inference risk. We observe interesting trends in inference risk with model training. In most cases, inference risk is high after even one epoch of model training (Figure 7). This is especially surprising because the model takes a few epochs to get good performance on the task itself, but shows that the model is learning and exposing aspects of the distribution even early in its training.

These trends clearly suggest that model under-training is not a feasible defense. Training beyond minimum generalization gap does lead to significantly reduced distribution inference risk. However, this region of the training corresponds to overfitting, which is known to be positively correlated with increased risk to membership inference [42]. Thus, a model trainer that is willing to overfit its models to avoid distribution inference adversaries would risk making the model more vulnerable to membership inference.

C. Re-Sampling Data

If the victim is aware of the property that an adversary might target, or only has a few known properties of the distribution that it wants to protect, the easiest mitigation is to just modify the training distribution (or the sampling mechanism) such that the property is no longer present for the training dataset. Knowing the particular property to hide

is a plausible assumption that is often assumed in work on distribution inference defenses. For instance, Chen and Ohrimenko [33] propose a theoretically-grounded defense that builds upon the distributional privacy framework [43] and modifies feature values to provide privacy guarantees against distribution inference adversaries.

Zhou et al. [44] propose over-sampling to reduce inference risk, but do so by adding new samples to their training data. Although this defense eliminates distribution risk (at the cost of model performance), the availability of new data is not always possible. Model trainers typically use all available data. We explore two variations of re-sampling defenses: *over-sampling* and *under-sampling*. In both cases, the model trainer re-samples data from its available datasets such that the resulting dataset is indistinguishable from a dataset sampled from a different distribution. For over-sampling we experiment with two flavors: using simple replacement and over-sampling based on inserting augmented data. These defenses rely on the key assumption that the model trainer knows the property they want to hide, and that there are a only few such properties so re-sampling to hide the desired properties will not unduly hard the model's task performance. When this assumption holds, resampling defenses can virtually eliminate inference risk. We evaluate this defense on configurations with low (CelebA–Sex), medium (Census19–Sex), and high (RSNA Bone Age–Age) inference risk to measure the impact of this defense.

Under-Sampling. The model trainer can simply under-sample its data such that the resulting dataset has a ratio corresponding to some other distribution. For example, a model trainer, with a dataset containing 70% females who wants to hide the ratio of females in the dataset from an inference adversary can simply under-sample examples with the 'female' attribute such that its data is balanced. This defense should prevent any disclosure about the pre-sampled distribution since there should be no difference between the cases where the training data was balanced to begin with and when it was adjusted with this defense, so long as the distribution is not distorted by the under-sampling. In our experiments, we find that under-sampling lowers inference risk significantly, but does not completely eliminate it—mean distinguishing accuracy drops below 54% for CelebA with significantly lower leakage for Census19 ($< 57\%$) and RSNA Bone Age ($\sim 60\%$) (Table VIII).

Over-Sampling. A model trainer not willing to sacrifice available training data by under-sampling may prefer to over-sample. The most basic variant over-samples the data before training begins, duplicating training records, and then trains its models like usual. Although this defense leads to complete utilization of data, the presence of repeated data can be revealing and may reveal the property the adversary wants to hide to an adversary adversary aware of the defense. It could, for instance, lead to a change in group-wise accuracies, which an adversary can learn to identify and still succeed at distribution inference.

Augmentation Based Over-Sampling. The ideal scenario for the defense would comprise of injecting fresh labeled data to adjust the desired property, as was assumed by Zhou et al. [44]. However, labeled data is scarce and may be expensive to acquire, and using techniques like pseudo-labeling can still leak information. In such scenarios, the model trainer can use augmentation techniques to synthetically generate additional samples. This avoids repeating samples, and may have the added benefit of potentially increasing the model's robustness to augmentations. But, the use of augmented data in an imbalanced way may still reveal information to a distribution inference adversary. For this defense, we focus only on the CelebA dataset, since designing augmentation for tabular datasets is much harder, and augmentations for RSNA Bone Age are limited. Task accuracy remains comparable and inference risk drops significantly (slightly higher than other forms of sampling), but is not completely eliminated and still higher than standard under and over-sampling.

Re-Sampling	α	**Precision**		**Recall**	
		not-white	white	not-white	white
Under-Sampling	< 0.5	$\downarrow 2\%$	$\sim$	$\uparrow 1\%$	$\downarrow 1\%$
	> 0.5	$\downarrow 1\%$	$\sim$	$\downarrow 1\%$	$\sim$
Over-Sampling	< 0.5	$\downarrow 2\%$	$\downarrow 1\%$	$\uparrow 1\%$	$\downarrow 1\%$
	> 0.5	$\downarrow 1\%$	$\downarrow 1\%$	$\downarrow 1\%$	$\downarrow 1\%$

TABLE VII: Relative change (%) in precision and recall metrics for white and not-white (race attribute), for Census19 (gender) for under-sampling and over-sampling. We consider cases where data for males ($\alpha < 0.5$) or females ($\alpha > 0.5$) is under-sampled for equalization.

Impact on Fairness. This form of re-sampling is common in research related to improving fairness in machine learning [45], commonly known as "unbiasing". However, re-sampling data can impact different sub-groups and populations of the distributions unequally, creating issues related to fairness in model predictions [46]. To investigate such potential impacts, we measure the impact of under-sampling and over-sampling-based mitigation strategies on fairness. We compare the precision and recall for another group and its possible values, for both undersampling and oversampling. Re-sampling based defenses have negligible impact on fairness in the case of CelebA, but result in disparate impacts of both under/over-sampling on the two groups. for Census19 (Table VII). For instance, over-sampling from a ratio $\alpha < 0.5$ lowers both precision and recall for whites, but increases recall and decreases precision more greatly for not-whites. These changes are even more severe for RSNA Bone Age, where changes in precision can be as high as 20% in opposite directions for different groups.

Adaptive Attacks. An adversary with knowledge of the under-sampling approach may be able to derive the original distribution by estimating the size of the training data to learn the sampling ratio. The strongest adversary would be one that starts with knowledge of specific records in the original training dataset, and can use membership inference attacks to estimate

Defense	Task Accuracy (%)	Dist. Acc. (%) $\alpha_1 = 0.2$	Dist. Acc. (%) Mean
Census19 (Sex)			
No Defense	77.9 ± 0.9	89.8	82.5±17.9
DP ($\epsilon = 1.0$)	77.0 ± 1.0	65.4	69.3±14.6
DP ($\epsilon = 0.12$)	75.6 ± 1.0	72.4	73.4±14.9
Label Poisoning ($r = 0.2$)	77.3 ± 1.0	78.4	78.9±17.4
Label Poisoning ($r = 0.4$)	74.9 ± 1.2	66.4	70.0±17.9
Under-sampling	77.5 ± 0.5	50.0	56.7±6.9
Over-sampling	77.3 ± 0.6	50.0	51.9±2.6
RSNA Bone Age (Age)			
No Defense	65.8 ± 2.0	100	99.8±0.4
Label Poisoning ($r = 0.2$)	64.3 ± 2.3	100	95.7±6.2
Under-sampling	65.4 ± 3.2	73.4	59.1±13.3
Over-sampling	64.6 ± 2.8	70.4	60.2±11.2
CelebA (Sex)			
No Defense	91.6 ± 0.8	89.6	85.3±15.8
Label Poisoning ($r = 0.2$)	90.0 ± 5.0	82.0	78.3±15.6
Adv. Training ($\epsilon = 4/255$)	90.4 ± 0.8	93.8	86.8±16.4
Adv. Training ($\epsilon = 8/255$)	88.5 ± 1.2	95.4	88.3±15.0
Adv. Training ($\epsilon = 16/255$)	75.7 ± 11.9	76.6	58.9±13.1
Under-sampling	90.8 ± 1.1	50.0	53.7±6.1
Over-sampling	90.6 ± 0.8	50.0	53.8±4.1
Augmentation Based Over-Sampling	91.7 ± 1.6	74.8	61.0±14.5

TABLE VIII: Effectiveness of considered defenses. Each distinguishing accuracy reported is the observed prediction accuracy of KL. The first results are for predicting between $\alpha_0 = 0.5$ and $\alpha_1 = 0.2$; the last column reports mean distinguishing accuracy (as described in Table I). Mean distinguishing accuracy numbers are reported with $\pm$ standard deviation, over different α_1 values. Most noise-based defenses harm model task accuracies, and the only defenses that diminishes leakage without harming task accuracy are based on data re-sampling.

how many of those records are included in the under-sampled dataset. We evaluate such attacks in Appendix B, and find they are unlikely to be effective without dramatic improvements to membership inference attacks.

VI. Limitations and Conclusions

Distribution inference attacks are known to reveal sensitive properties about underlying training distributions, but their effectiveness has been established only in very controlled settings so far. Our work advances understanding of this risk in more realistic settings, but is still far from understanding all of the issues that might impact a real deployment and attack.

Our proposed black-box attacks are highly efficient and maintain their effectiveness even when access to exact prediction probabilities is unavailable. Our experiments find that differences in model architectures harm inference accuracies, but the impact is not very severe as long as models with similar complexity are used. Even the lack of common feature extractors, a common setting in many evaluations in the literature, does not completely eliminate inference risk.

Like nearly all inference privacy work, we assume an adversary with access to a dataset that matches the underlying distribution (in this case, before the transformation to the actual training distribution as modeled by $\mathcal{G}_0$ and $\mathcal{G}_1$). This is a strong assumption, which may be realistic in some cases but is often unlikely. All our attacks (and nearly all previous ones) require representative data for training models locally. Exploring adversaries with limited data access to these distributions and how it impacts inference risk is left as part of future work.

Configurations beyond single-party learning have seen growing interesting lately. Active adversaries with data poisoning capabilities have been demonstrated to be highly effective in both single-party [3] and multi-party [47] settings. Our proposed attacks are highly potent for many configurations, but the exact impact of different training setups like federated learning remains largely unexplored.

The general approach to achieve security and privacy for machine-learning models is to add noise, but our evaluations suggest this approach is not a principled or effective defense against distribution inference. The main reductions in inference accuracy that result from these defenses seem to be due to the way they disrupt the model from learning the distribution well, so observed reductions in inference risk are related to drops in task performance. Our experiments with different model architectures and differentially private training support this—inference risk increases significantly when the victim and adversary use the same learning algorithms or model architectures. The only reliably effective defense from our experiments is to re-sample data, which depends on the assumption that the model training is aware of the adversary's inference goals (or at least of the properties that should be protected). These re-sampling defenses too are not perfect, however, as they seem to negatively impact fairness of groups related to the property attribute.

There is a need for more theoretical connections between distribution inference risk and general useful notions of machine learning, like model complexity and fairness. Our work suggests such connections do exist, and we hope they will be better understood as both empirical and theoretical understanding of inference privacy advances.

References

[1] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers," *International Journal of Security and Networks*, vol. 10, no. 3, pp. 137–150, 2015.

[2] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations," in *ACM Conference on Computer and Communications Security*, 2018.

[3] S. Mahloujifar, E. Ghosh, and M. Chase, "Property Inference from Poisoning," in *IEEE Symposium on Security and Privacy*, 2022.

[4] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang, "Inference Attacks Against Graph Neural Networks," in *USENIX Security Symposium*, 2022.

[5] X. Wang and W. H. Wang, "Group Property Inference Attacks Against Graph Neural Networks," *arXiv preprint arXiv:2209.01100*, 2022.

[6] A. Suri and D. Evans, "Formalizing and Estimating Distribution Inference Risks," in *Privacy Enhancing Technologies Symposium*, 2022.

[7] W. Zhang, S. Tople, and O. Ohrimenko, "Leakage of Dataset Properties in Multi-Party Machine Learning," in *USENIX Security Symposium*, 2021.

[8] 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.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in *International Conference on Learning Representations*, 2018.

[10] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting Unintended Feature Leakage in Collaborative Learning," in *IEEE Symposium on Security and Privacy*, 2019.

[11] X. Xu, Q. Wang, H. Li, N. Borisov, C. A. Gunter, and B. Li, "Detecting AI Trojans Using Meta Neural Analysis," in *IEEE Symposium on Security and Privacy*, 2021.

[12] A. Suri, P. Kanani, V. J. Marathe, and D. W. Peterson, "Subject Membership Inference Attacks in Federated Learning," *arXiv preprint arXiv:2206.03317*, 2022.

[13] H. Chaudhari, J. Abascal, A. Oprea, M. Jagielski, F. Tramèr, and J. Ullman, "SNAP: Efficient Extraction of Private Properties with Poisoning," *arXiv preprint arXiv:2208.12348*, 2022.

[14] M. Juárez, S. Yeom, and M. Fredrikson, "Black-Box Audits for Group Distribution Shifts," *arXiv preprint arXiv:2209.03620*, 2022.

[15] J. Hartley and S. A. Tsaftaris, "Measuring Unintended Memorisation of Unique Private Features in Neural Networks," *arXiv preprint arXiv:2202.08099*, 2022.

[16] J. Su, "Census19," https://github.com/JerrySu11/CensusData, 2022.

[17] S. D. Bay, D. Kibler, M. J. Pazzani, and P. Smyth, "The UCI KDD Archive of Large Data Sets for Data Mining Research and Experimentation," *ACM SIGKDD Explorations Newsletter*, 2000.

[18] B. Jayaraman, "Texas-100X," https://github.com/bargavj/Texas-100X, 2022.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale CelebFaces Attributes (CelebA) Dataset," 2018.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[21] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning Face Representation from Scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[22] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala *et al.*, "The RSNA Pediatric Bone Age Machine Learning Challenge," *Radiology*, vol. 290, no. 2, pp. 498–503, 2019.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[24] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft Academic Graph: When experts are not enough," *Quantitative Science Studies*, 2020.

[25] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations*, 2017.

[26] S. Tople, A. Sharma, and A. Nori, "Alleviating Privacy Attacks via Causal Learning," in *International Conference on Machine Learning*, 2020.

[27] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and Robust Automated Machine Learning," in *Advances in Neural Information Processing Systems*, 2015.

[28] Y. Xie and D. Richmond, "Pre-training on Grayscale ImageNet Improves Medical Image Classification," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[29] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned Language Models are Zero-Shot Learners," in *International Conference on Learning Representations*, 2021.

[30] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[31] V. Hartmann, L. Meynent, M. Peyrard, D. Dimitriadis, S. Tople, and R. West, "Distribution inference risks: Identifying and mitigating sources of leakage," *arXiv preprint arXiv:2209.08541*, 2022.

[32] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant Risk Minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[33] M. Chen and O. Ohrimenko, "Protecting Global Properties of Datasets with Distribution Privacy Mechanisms," *arXiv preprint arXiv:2207.08367*, 2022.

[34] W. Zhang, O. Ohrimenko, and R. Cummings, "Attribute Privacy: Framework and Mechanisms," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 757–766.

[35] X. Ma, B. Li, Q. Jiang, Y. Chen, S. Gao, and J. Ma, "NOSnoop: an Effective Collaborative Meta-Learning Scheme against Property Inference Attack," *IEEE Internet of Things Journal*, 2021.

[36] J. Stock, J. Wettlaufer, D. Demmler, and H. Federrath, "Lessons Learned: How (Not) to Defend Against Property Inference Attacks," *arXiv preprint arXiv:2205.08821*, 2022.

[37] I. Mironov, "Rényi Differential Privacy," in *IEEE Computer Security Foundations Symposium (CSF)*. IEEE, 2017.

[38] A. G. Thakurta, A. H. Vyrros, U. S. Vaishampayan, G. Kapoor, J. Freudiger, V. R. Sridhar, and D. Davidson, "Learning new words," *Granted US Patents*, vol. 9594741, 2017.

[39] J. M. Abowd, R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala *et al.*, "The 2020 Census Disclosure Avoidance System TopDown Algorithm," *arXiv preprint arXiv:2204.08986*, 2022.

[40] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial Examples are not Bugs, they are Features," *Advances in neural information processing systems*, vol. 32, 2019.

[41] Y. Zhang, M. Gong, T. Liu, G. Niu, X. Tian, B. Han, B. Schölkopf, and K. Zhang, "Adversarial robustness through the lens of causality," in *International Conference on Learning Representations*, 2022.

[42] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," in *IEEE Computer Security Foundations Symposium*, 2018.

[43] Y. Kawamoto and T. Murakami, "Local Obfuscation Mechanisms for Hiding Probability Distributions," in *European Symposium on Research in Computer Security*. Springer, 2019, pp. 128–148.

[44] J. Zhou, Y. Chen, C. Shen, and Y. Zhang, "Property Inference Attacks Against GANs," *NDSS*, 2022.

[45] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[46] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification," in *Proceedings of the 2018 world wide web conference*, 2018.

[47] Z. Wang, Y. Huang, M. Song, L. Wu, F. Xue, and K. Ren, "Poisoning-Assisted Property Inference Attack Against Federated Learning," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[48] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced Membership Inference Attacks against Machine Learning Models," in *ACM Conference on Computer and Communications Security (CCS)*, 2022.

APPENDIX

A. Experimental Details

For each dataset, we create non-overlapping splits of data for the victim and adversary, where the victim has at least double the amount of adversary's data for ogbn-arxiv and RSNA Bone Age, triple for CelebAand Texas-100X, and $4\times$ for Census19. For each dataset, we simulate $\mathcal{D}$ using the dataset itself. Simulation of distributions with particular α values is is achieved by sampling data with attributes 0 and 1 such that their ratios result in some desired α. For properties not based on binary attributes like ogbn-arxiv, this is achieved by pruning nodes iteratively from the graph (while re-computing neighbor counts along the way) to achieve a desired mean node-degree. We include the datasets and victim/adversary splits used in previous experiments, and include results on two new datasets. For all of the experiments, we follow the processing pipeline described in Suri and Evans [6] to obtain non-overlapping splits. Essentially, both parties sub-sample from their data splits (to achieve specific α values) with different random seeds, and train models on the sampled data.

We perform each experiment five times and report mean values with standard deviation in all of our experiments. For each dataset, we train 250 victim models per distribution. For all black-box attacks, the adversary trains and uses 50 models per distribution for each trial. For white-box attacks, the adversary trains 800 models per distribution, of which 750 are used for training and 50 as the validation set. For cases with very large models (like DenseNet trained from scratch for RSNA Bone Age), we use 100 victim models per distribution.

B. Adaptive Attacks Against Under-Sampling

Assume a more powerful adversary that has access to m training records each corresponding to attribute 0 ($\mathcal{D}_-$) and 1 ($\mathcal{D}_+$), and the original dataset has size n. In this scenario, the adversary is unaware of the original distribution of these attributes (α). Consider the scenario where the victim utilizes under-sampling on its original distribution as a defense to protect α, and re-samples such that both attributes are equally likely.

For our analysis, we assume a near-perfect membership inference adversary, with a false negative rate β. In such a setup, the adversary can check which of its data (the one with attributes zero, and attributes one) still all tests as members. If all zeros still remain members, then data from ($\mathcal{D}_+$) must have been under-sampled, and thus the original α must be > 0.5. Assuming that under-sampling is performed by pruning points uniformly at random, the density of members in the resulting data must remain the same. Thus,

$$\frac{m}{\alpha \cdot n} = \frac{m^{'}}{(1-\alpha)\cdot n} \tag{4}$$

$$m_- = \beta \cdot m \tag{5}$$

$$m_+ = m^{'} \cdot \beta \tag{6}$$

where m_+ is the number of datapoints (out of the known m) with attribute 1 that are inferred as members by the adversary, and m_- for attribute 0. α can thus be estimated as $m_-/(m_- + m_+)$. By symmetry, the case where original $\alpha < 0.5$ yields a similar formula. We test the risk of this adversary while varying the number of data points m, for different values of α. The adversary in this case thus directly predicts α, and mean square error (MSE) values are computed accordingly for the regression case. We use the R attack from Ye et.al. [48], and use the authors' official implementation, with the FPR set to 0.05.

MSE values for varying number of members (m) are reported in Table IX. For the most realistic case, with knowledge of upto 100 members (m) per attribute, the inference risk remains very low, with MSE values as high as ~ 5. This risk is expected to increase with increase in membership knowledge, as in the extreme case, the adversary would have perfect knowledge of the entire training dataset. One notable exception is RSNA Bone Age, where the MSE drops to ~ 0.4 for $m = 500$ This is not surprising, as $m = 500$ for the case of RSNA Bone Age corresponds to $\sim 15\%$ of the victim's training dataset, which is unrealistically high.

For the task of binary distinguishing between $\alpha_0 = 0.5$ and some other α_1, it suffices to see whether the predicted ratio is sufficiently different from 0.5. We do so by checking the predicted α, and predict $\mathcal{G}_1(\mathcal{D})$ if it differs from 0.5 by more than 0.03, and $\mathcal{G}_0(\mathcal{D})$ otherwise. As a baseline, we also consider a simpler adaptive adversary that uses the same re-sampling setup

Dataset/Task	**Number of known members (m)**		
	10	100	500
Census19 (Sex)	9.043	4.588	4.078
RSNA Bone Age (Age)	1.969	0.486	0.372
CelebA (Sex)	4.785	1.466	1.202

TABLE IX: MSE values for direct regression over α for an adversary that utilizes membership inference to infer α for models trained with under-sampling based defense.

Dataset/Task	KL	MI-Based	Same Setup + KL
Census19 (Sex)	56.7	57.1	80.4
RSNA Bone Age (Age)	59.1	65.9	50.0
CelebA (Male)	53.7	50.0	64.7

TABLE X: Mean distinguishing accuracies for the task of binary distinguishing between $\alpha_0 = 0.5$ and α_1, while varying α_1, for the standard adversary (KL), an adversary that utilizes membership inference to infer α for models trained with under-sampling based defense (MI-Based), and a simpler adversary that just copies the victim's re-sampling setup (Same Setup + KL).

as the victim, re-sampling data for its shadow models. Such an adversary can potentially work, since the KL Divergence Attack compares distributions of predictions, which might be sufficiently different between re-sampled and non-sampled ($\alpha = 0.5$) models.

Mean distinguishing accuracies are reported in Table X. Cases where the adversary uses the same setup as the victim (for re-sampling) leads to significant inference leakage in most cases, with mean distinguishing accuracies as high as 80% for Census19. Similarly, the MI-based distribution inference leakage is particularly high for RSNA Bone Age. This is in line with previous observations with the MSE values (Table IX), since the number of members corresponds to a significant portion of the victim's training data.