"You Can't Fix What You Can't Measure": Privately Measuring Demographic Performance Disparities in Federated Learning

Marc Juarez
University of Southern California
marc.juarez@usc.edu

Aleksandra Korolova
University of Southern California
korolova@usc.edu

ABSTRACT

Federated learning allows many devices to collaborate in the training of machine learning models. As in traditional machine learning, there is a growing concern that models trained with federated learning may exhibit disparate performance for different demographic groups. Existing solutions to measure and ensure equal model performance across groups require access to information about group membership, but this access is not always available or desirable, especially under the privacy aspirations of federated learning.

We study the feasibility of measuring such performance disparities while protecting the privacy of the user's group membership and the federated model's performance on the user's data. Protecting both is essential for privacy, because they may be correlated, and thus learning one may reveal the other. On the other hand, from the utility perspective, the privacy-preserved data should maintain the correlation to ensure ability to perform accurate measurements of the performance disparity. We achieve both of these goals by developing locally differentially private mechanisms that preserve the correlations between group membership and model performance. To analyze the effectiveness of the mechanisms, we bound their error in estimating the disparity when optimized for a given privacy budget, and validate these bounds on synthetic data. Our results show that the error rapidly decreases for realistic numbers of participating clients, demonstrating that, contrary to what prior work suggested, protecting the privacy of protected attributes is not necessarily in conflict with identifying disparities in the performance of federated models.

KEYWORDS

differential privacy, algorithmic fairness, federated learning

1 INTRODUCTION

Federated learning (FL) has become a popular way to distribute the training of machine learning models across multiple organizations and multiple devices. Even though most of the academic literature in FL has focused on deployments across multiple organizations, also known as cross-silo federated learning (SFL), there are several large-scale deployments of cross-device federated learning (DFL) in the industry, such as Google's next-word prediction model for Android's Gboard [35, 50, 70], and Apple's speaker identification model in Siri [31]. A key motivation for DFL is the aspiration of training powerful machine learning models while ensuring data minimization and privacy of the individuals involved.

In parallel, machine learning models have been shown to exhibit differences in performance across groups, often falling short for people from historically marginalized groups, in the domains of vision, natural language processing, and healthcare [14, 16, 52, 58]. The challenge of disparate performance and, therefore, disparate impact, is relevant also for models trained with federated learning, and has begun to be observed both in SFL [12, 12, 18, 56, 67] and in DFL versions of it [68, 69, 72].

Performance disparity in the DFL setting may be harmful beyond merely the individual's experience of worse quality of service [19]. For example, Amazon has proposed to use DFL to improve Alexa's wake-word detection [64]. At the same time, research has shown that Alexa's model and other non-federated speech recognition models perform worse when interpreting voices of women and people of non-white ethnicities [5, 38, 62]. When the wake-word is incorrectly detected, Alexa sends the samples of unrelated speech to a server for speech recognition. A recent study shows that such mistaken activations result in up to a minute of speech recording being uploaded to the cloud [1]. Thus, apart from worse user experience, members of these groups may be subject to more surveillance.

Another example specific to federated learning is Mozilla's experimentation with DFL for ranking recommendations in the Firefox's URL bar [37]. By default, Firefox preemptively resolves the domain name for the URL bar autocompletes. When used in conjunction with a DFL system, groups that have worse ranking recommendations resolve irrelevant domain names, and thus may unnecessarily be more exposed to network adversaries. Further, in applications of DFL to the security domain, such as authentication [40], a performance disparity may lead to a lower security level for certain groups.

Overall, federated learning has been getting tremendous traction in industry and academia in recent years. If it were to become a de-facto data minimization or privacy standard for much of machine learning, as current trends suggest, an unknown performance disparity dependent on group membership could have tremendous negative consequences in many application domains.

Like in traditional ML, one of the challenges in detecting and mitigating disparate performance of the DFL model is that access to information about the attributes related to group membership is often limited or noisy, or puts the questions of privacy and fairness in direct opposition [3, 9, 20, 63]. Furthermore, regulations, such as the GDPR in the EU, mandate that protected attributes (determining relevant group membership for the measurement of performance disparity), such as gender or race, not be collected at all, with the exception of organizations that request explicit informed consent. As a consequence, companies—especially in *Big Tech* [9]—tend to not collect the attributes, hindering the detection of the disparities.

Prior work in traditional ML has attempted to reconcile the goals of protecting the privacy of group membership and mitigation of performance disparity by engaging a third-party in the collection of the protected attributes [63], use of secure multi-party computation [46], and ensuring differential privacy of the attributes [42]. A major distinction between the models considered in these works and the models obtained through DFL is how they are applied: these prior works mitigate disparity on a model that is used to make decisions

about individuals, whereas existing DFL models are distributed to clients for use on their own data. Credit score prediction models are an example of the former, and wake-word detection models, an example of the latter. Due to this distinction, the fairness definitions used by the previously employed techniques are not relevant for these DFL applications. We focus instead on the measure of *unfairness* determined by the difference in the performance of the global model across groups of clients, an informative metric about the disparate performance of DFL models.

We overcome the seeming contradiction between fairness and privacy in FL via a deliberate use of differential privacy (DP). We design DP mechanisms that allow to detect performance disparities while protecting the privacy of the group membership (or attributes determining such membership), albeit at the cost of introducing an error in the measurements. We propose and explore several DP mechanisms to achieve this goal and compare their trade-offs between privacy and the ability to detect a performance disparity. Our theoretical analysis shows that the mechanisms ensure strong privacy guarantees while the error in the measurements is relatively low for the typical number of clients in a DFL setting. With our tools, a regulatory agency, or even the aggregator of the models—if meeting the consent requirements for an exception in the relevant law framework—could identify cases of performance disparity in applications where such disparity might be undesirable or harmful for some of the groups.

Our contributions towards the private measurement of demographic disparities in the performance of FL models are:

- We propose local differential privacy mechanisms that provide strong privacy guarantees of the group membership while enabling measurements of the difference in a DFL model's mean performance across demographic groups. These mechanisms also protect the federated model's performance on the user's data, which may be correlated with group membership.
- We provide an analysis of the mechanisms over a wide range of privacy levels, number of clients, and group sizes. To draw the comparison between the mechanisms, we characterize their measurement error as a function of the privacy budget, and find the budget allocation that minimizes this error under a privacy constraint.
- We discuss the trade-offs between privacy and utility that the operator of the mechanisms has to make, and illustrate how the mechanisms can enable mitigation strategies to promote equal model performance across groups in the DFL setting. Our findings indicate that existing DFL deployments can afford the privacy budgets required by a high-accuracy measurement with strong privacy guarantees.
- We evaluate these trade-offs on data synthetically generated from a real-world dataset. The results show that the theoretical bounds not only hold, but are significantly more conservative than the empirical errors, indicating that, in practice, the measurements might be more accurate at the same privacy level.

The rest of the paper is structured as follows. Section 2 provides the background on DP and FL. In Section 3, we provide the definition of the performance gap or disparity that we consider, describe the adversary model, and discuss related work. Section 4 describes the proposed mechanisms for protecting privacy while computing the performance gap. Section 5 provides a theoretical analysis of the error the DP mechanisms induce in the performance gap measurements, and thus the privacy vs. utility trade-offs they impose. In addition, we describe how we use a real-world activity detection dataset to generate synthetic data and evaluate the error of the mechanisms for a realistic number of clients. Section 6 compares our approach to alternative approaches, details an implementation of the mechanisms, describes the strategies that the mechanisms enable to mitigate a performance disparity, and outlines potential future work. Finally, Section 7 summarizes the implications of our findings.

2 BACKGROUND

The following is the necessary background on differential privacy and federated learning to follow the rest of the paper.

2.1 Differential Privacy (DP)

DP is a privacy framework that bounds the amount of individual information that an analyst can infer from the data. We have two main DP trust models: the *central* and the *local* model. The central model assumes a trusted curator that holds a dataset and enforces the DP guarantee on queries performed by external entities. By contrast, in the local model the individuals protect their inputs before sharing them. The local model does not need a curator but requires larger perturbations than the central model for the same privacy guarantee, thus incurring greater error on the utility of the statistics.

The local model is better suited for cross-device federated learning (DFL), as it is unclear who would play the role of a curator in the existing deployments of DFL. In addition, the high number of clients in the typical DFL setting can attenuate the error in the statistics while eliminating the need for a trusted third-party.

Definition 2.1 (ϵ-Local Differential Privacy (ϵ-LDP))**.** A randomized mechanism $\mathcal{M}: D \to R$ satisfies ϵ-LDP where $\epsilon > 0$ if, and only if, for any pair of inputs $v, v' \in D$ and for all $y \in R$

$$\frac{\Pr[\mathcal{M}(v) = y]}{\Pr[\mathcal{M}(v') = y]} \leq e^{\epsilon},$$

where the probabilities are taken over the randomness of $\mathcal{M}$.

One of the simplest LDP mechanisms is *Randomized Response* (RR). RR was designed to provide plausible deniability to survey respondents [66]. For a binary protected attribute, the respondents reply with the true value with probability a and give the opposite value otherwise. Generalized RR (GRR) extends RR to a non-binary protected attribute by evenly distributing the probability of *lying* over the other values of the protected attribute [65].

Definition 2.2 (The GRR mechanism)**.** For $x \in \{1,...,d\}$, $d > 1$, and $a \in [\frac{1}{2}, 1]$, the GRR mechanism, $\mathcal{M}_{\text{GRR}}$, is defined by

$$\Pr[\mathcal{M}_{\text{GRR}}(x; d, a) = y] := \begin{cases} a & \text{if } y = x \\ \frac{1-a}{d-1} & \text{if } y \neq x \end{cases}$$

If $a = \frac{e^{\epsilon}}{e^{\epsilon}+d-1}$, $\mathcal{M}_{\text{GRR}}$ is ϵ-LDP [65], and when $d = 2$, $\mathcal{M}_{\text{GRR}}$ is the binary RR mechanism.

For continuous values, the classic LDP mechanism is the Laplace mechanism [24], denoted by $\mathcal{M}_{\text{Lap}}$. The Laplace mechanism draws noise from a Laplace distribution and adds it to the inputs.

Definition 2.3 (The Laplace mechanism)**.** For $x \in [-1, 1]$, the Laplace mechanism, $\mathcal{M}_{\text{Lap}}$, is defined by

$$\Pr[\mathcal{M}_{\text{Lap}}(x;\epsilon)=y] := f_{\text{Lap}(0,\frac{2}{\epsilon})}(y-x),$$

where $f_{\text{Lap}(0,\frac{2}{\epsilon})}$ is the probability density function of a Laplace r.v. with zero mean and scale $\frac{2}{\epsilon}$. Note that since we consider input values in $[-1,1]$, $\mathcal{M}_{\text{Lap}}$ achieves ϵ-LDP when the scale of the Laplace distribution is greater or equal than $\frac{2}{\epsilon}$ [24].

2.2 Federated Learning (FL)

Federated learning (FL) allows a number of clients, each with its own set of data, to collaborate in the training of a model on *all* the data, without the clients having to share the data among them or with third parties. In its most basic form, FL distributes the training over the clients by following a multi-round protocol with a central server known as the *aggregator*. At the start of the protocol, the aggregator holds a pre-trained model that they wish to refine using the data available at the clients. To refine the model, in each round, the clients download from the aggregator the parameters of the model, known as the *global model*, and perform a local update using the data on their devices. This update could be, for example, to take a step of mini-batch gradient descent. With the update, the clients have derived new models, and can share their parameters with the server, who computes a weighted average to obtain a new global model.

There are two types of FL depending on who the clients are: cross-device FL (DFL) and cross-silo FL (SFL). In DFL, the clients run on different devices (e.g., mobile phones) and the training data usually belongs to the same user. By contrast, in SFL the clients are different organizations or institutions, such as hospitals or banks, who hold data from many individuals. In addition, although the server in the FL protocol and the company that owns the rights to the model could be two separate entities, in DFL they are the same entity. In the rest of the paper, we refer to this entity as the *aggregator*.

In this paper, we focus on DFL as it is becoming increasingly popular in the Big Tech industry [31, 70]. As described in [11], the way DFL is deployed is by selecting a few hundreds of clients for training, and a disjoint set of clients of a similar size for testing. After hyper-parameter tuning and evaluation of the federated model, the model is deployed to the whole population of clients, which is orders of magnitude larger than the training and testing sets. In these applications of DFL, the clients apply the federated model to inform decisions that concern the client's user, as opposed to making decisions about others, as it is often the case in SFL. This is an important difference between the DFL setting that we consider and other FL settings, as it will motivate the (un)fairness notion that we propose.

3 PROBLEM STATEMENT

We consider K clients and each client has evaluated the global model on a fraction of their data to obtain a performance value $v_k \in [-1,1]$[1]. We consider a disjoint protected attribute, $A = \{1,...,d\}$, with $d > 1$, which partitions the set of clients: $\mathcal{P} = \{G_1,...,G_d\}$. The protected attribute could also be the intersection of several disjoint protected attributes. The k-th client's group value is $g_k \in A$. We drop the subindices of v and g when the client we are referring to is clear.

The (un)fairness notion. Motivated by the harms of disparate performance of the global model in DFL, the notion of *unfairness* that we consider in the DFL setting is disparate performance across groups of clients defined by a demographic attribute, such as sex or race. The mean performance of a group is defined as follows.

Definition 3.1 (**Group mean performance).** The mean performance of a group $G \in \mathcal{P}$ is $m_G := \frac{1}{n}\sum_{i=1}^{n} v_i$, where $n = |G|$.

To quantify the difference in performance between two groups, we measure the absolute difference between the group mean performances. We call this quantity the *performance gap*.

Definition 3.2 (**Performance gap).** The performance gap between any two $A, B \in \mathcal{P}$ is defined by $\Delta m := |m_A - m_B|$.

This notion of unfairness is in contrast with traditional fairness definitions that measure the performance of the classifier on individual predictions [7, 8, 41, 45, 55]. Those definitions are suitable for scenarios where data points represent people and each single prediction concerns an individual. However, these definitions are not suited for the typical DFL scenario, as the global model is not applied to make predictions on instances that represent individuals, but rather—to solve a learning task with the data that the client has collected on their device. The performance gap in solving the task captures the disparity that the groups are subjected to in the DFL scenario.

Yue et al. propose a fairness definition that does consider the DFL setting [73]. According to this definition, a model is more fair if its performance has less variance across groups. However, similarly to previous definitions, the group performance is not measured on each individual device and is thus not suited for typical DFL applications.

As with fairness definitions for decision-making about individuals, the choice of an appropriate performance metric depends on various factors, such as the learning task and the potential harms. For example, in the application of FL to wake-word detection of the introduction, the aggregator may aim at minimizing the false positive rate gap, as that would reduce the difference between the groups in the average percentage of recordings that are sent to the cloud by mistake. We abstract the performance metric and our methods are not specific to a particular application.

Adversary model. Because a performance gap in solving the task across groups may be detrimental for some of the groups, we consider an entity interested in measuring this gap. In the rest of the paper, we assume that such entity is the aggregator (i.e., the entity that averages the local models of FL), but our mechanisms could also be used by an external entity, such as a regulatory agency or a public interest auditor. As in popular DFL deployments [49], we assume that the aggregator uses Secure Aggregation [10] to obtain the client updates. Therefore, the aggregator cannot infer any information about the protected attributes from the updates.

Both the group and the model performance value are privacy-sensitive information (the group—because it corresponds to an attribute such as race; the performance—because of the potential correlation with the group). Thus, the clients apply a perturbation mechanism ensuring local DP, $\mathcal{M}$, on their group-value tuples before sending them to the aggregator. We denote the perturbed values $(g'_k, v'_k) :=$

[1] Although performance metrics are often in $[0,1]$, we take $[-1,1]$ because it simplifies parts of the analysis of the LDP mechanisms. Note that mapping $[0,1]$ to $[-1,1]$ does not have any implication for the privacy of the mechanisms.

$\mathcal{M}(g_k,v_k)$. From a privacy perspective, we assume that the aggregator is an honest-but-curious adversary, i.e., they follow the protocol as intended but may try to learn g_k or v_k from the perturbed tuples.

In this study, we investigate the question: *how can the aggregator measure the performance gap while protecting the privacy of the clients' (g_k,v_k) tuples with an LDP mechanism?* To address this question, we design novel LDP mechanisms and study the trade-offs they impose in terms of their privacy guarantees and the error they induce in the measurements. *We hypothesize that the number of clients of current DFL deployments is sufficient to allow LDP mechanisms that achieve low-error and high-privacy.*

3.1 Related Work

The fields of algorithmic fairness and FL have both rapidly developed in recent years. We find an array of works that design methods to measure and mitigate unfairness for traditional machine learning models [30, 36, 43, 74, 75]. Recently, researchers have adapted these techniques to the FL setting [29, 54, 73]. However, these techniques, including the ones in the non-FL setting, assume access to the protected attributes. This assumption is a significant limitation of these prior works because, in practice, access to the protected attributes is often restricted or not available for legal and privacy reasons [9, 63].

In the traditional machine learning literature, Veale and Binns first noted the legal, institutional, and commercial deterrents against collecting demographic data [63]. This lack of demographic data renders the methods that assume their availability impractical. To address this issue, they envision third-parties who collect the demographic data, and use privacy-preserving protocols to collaborate with the companies in detecting and mitigating discrimination [63].

Researchers materialized these protocols with DP and Secure Multi-Party Computation (SMPC) [42, 46]. Jagielski et al. proposed DP versions of existing post- and in-processing techniques to train classifiers that satisfy the Equalized Odds constraint [42]. In contrast, our work defines the notion of performance disparity as it is more suitable for the DFL setting. In addition, our focus is on the measurement of the disparity rather than its mitigation, although we argue that our mechanisms can enable post-processing techniques to reduce the performance gap (see Section 6).

Kilbertus et al. propose an SMPC protocol that enables a company to train a model on the encrypted data – including demographic data– providing confidentiality of the protected attributes. A recent technical report by Meta describes a system to measure the performance gap with privacy of the protected attributes that is also strongly based on SMPC [2]. The challenges of an SMPC-based solution for DFL are its high computational and bandwidth costs, and its brittleness to client dropouts. Moreover, SMPC and DP provide different privacy guarantees; in particular, SMPC does not set any limit on the information about individual group membership that the aggregator can infer from the measurements.

Another difference between our work and the prior works is that the prior works assume the model holder's collaboration towards the goal of identifying disparities. By contrast, depending on how DFL is implemented, our approach may allow the clients and the mechanism operator to measure the performance gap without the aggregator's collaboration. These measurements could be performed by a regulatory agency, public interest researchers, or a collective of clients who cooperate to identify a performance gap in the global model.

Fairness with noisy attributes. Several works have proposed methods to train a classifier that enforces a fairness constraint in the presence of noisy protected attributes [4, 17, 48]. Our LDP mechanisms also add noise to such attributes. The fundamental difference between those works and ours is that they consider a different kind of noise: noise that arises from randomly changing the protected attributes. For example, Celis et al. consider a binary protected attribute and individuals who flip their attribute with different probabilities [17].

The noise that our mechanisms add, although unknown to the aggregator due to the randomness of the mechanism, is chosen according to a publicly known algorithm; thus, its effect on the aggregate utility, such as computing the average performance for a group, is different, more limited, and suitable for analysis. Furthermore, our goal is to ensure the rigorous privacy guarantees of DP precisely to discourage the kind of noise that [17] considers. With the privacy guarantees of our mechanisms, individuals do not have to fear retaliation from providing the true group membership value, thus the mechanisms dis-incentivize them to lie about it.

Fairness without protected attributes. Ensuring non-discrimination when access to demographic data is not available has been a central open question in both technical and policy circles [3, 9, 20, 63].

To overcome the lack of demographic data, some prior works have designed methods to train fair classifiers without the protected attributes. One approach is to define a proxy, a feature that correlates with the protected attribute, and apply traditional techniques on it [21, 34]. This approach is related to the works on fair classification with noisy attributes discussed above, as the proxy can be seen as a noisy version of the original attribute.

The work of [39] trains a classifier that enforces a fairness constraint without requiring any access to the protected attributes by minimizing the classifier's loss on the worst-case distribution of latent groups. However this approach is not robust to outliers, which can substantially degrade the performance of the resulting model [47].

This line of work is orthogonal to ours, as it assumes that the attributes are simply unavailable. By contrast, we argue that if their unavailability is due to privacy concerns, then these concerns can be overcome using mechanisms that ensure the strong privacy guarantees of local differential privacy, and thus be made available subject to DP constraints.

LDP mechanisms for group mean value estimation. Prior work on LDP mechanisms to protect sensitive attributes is too extensive to be covered in detail. Most of the LDP mechanisms in the related literature are designed to estimate statistics, e.g., mean, frequency, and heavy hitters, over the whole set of data [22, 23, 53]. We are, however, interested in computing the mean value for a subset defined by a specific attribute.

Recent work has proposed LDP mechanisms to estimate mean and frequency values for subsets of the data [33, 71]. However, the subsets they consider are not necessarily defined by demographic attributes, and thus do not take into account the unique characteristics of demographic data. For example, they do not study the error of

the mechanisms for different sizes of the sets, which is crucial to assess the feasibility of a mechanism to measure the performance gap. Moreover, these works are both based on $\mathcal{M}_{\text{RR}}$ and the value discretization proposed by Nguyên et al. [53]. In this paper, we explore using $\mathcal{M}_{\text{Lap}}$ to perturb the values and show that for some privacy budgets, it can ensure lower error under the same privacy constraint. Finally, this literature does not consider the perturbation of performance values for performance gap measurements, and therefore, is focused on slightly different privacy - utility tradeoffs than we are.

4 MEASURING THE PERFORMANCE GAP

We develop mechanisms that are able to measure the performance gap accurately, even when they only have privacy-protective access to the group membership. Since the performance gap is defined as the difference between the mean performance values of the groups, we reduce the problem to estimating the mean performance value of the model for a group. We show that LDP guarantees are not affected by this reduction and, in Section 5, provide an expression for the error of estimating the performance gap from the two group mean estimates. Thus, our mechanisms are designed to solve the problem of mean group value estimation while protecting the privacy of the group membership.

Our mechanisms use GRR to perturb the group values. The intuition for perturbing the group with GRR is that it provides plausible deniability for group membership. As a result, clients have less incentives to lie: since there is a significant probability that their group was swapped, they can always claim that the mechanism assigned them to a different group.

If the model exhibits performance disparity, then the performance values could be correlated with the group. Thus, the value perturbation part of the mechanisms is designed to preserve the *overall* aggregate correlation between the group and the values, while preventing the reliable inference of the group that an individual client belongs to from the performance values. Our mechanisms use GRR to perturb the group and differ in the noise that is added to perturb the values. We experiment with different perturbations, based on $\mathcal{M}_{\text{GRR}}$ and $\mathcal{M}_{\text{Lap}}$, because we aim at exploring the trade-offs between privacy and utility that different types of DP mechanisms can provide. As in much of the DP literature, we do not find a universally optimal mechanism; different mechanisms perform better in different privacy regimes.

Next, we describe the mechanisms and provide their LDP bounds as a function of the privacy budgets for the group and value perturbations. The following sections categorize the mechanisms into $\mathcal{M}_{\text{GRR}}$-based and $\mathcal{M}_{\text{Lap}}$-based mechanisms.

4.1 $\mathcal{M}_{\text{GRR}}$-Based Mechanisms

We draw inspiration from the LDP mechanisms to protect key-value pairs [33, 71] to define a value perturbation based on $\mathcal{M}_{\text{GRR}}$. Like the key-value pair mechanisms, we use discretization as a primitive, and introduce a dependency between the group and value perturbations that preserves the *overall* correlations in the data for the measurement. By contrast, because we consider a partition of the set of clients, our mechanisms are considerably simpler than the ones in the key-value literature.

Algorithm 1: Privacy Mechanism: $\mathcal{M}_{\text{R}}$.

Input : The client's group and value: (g,v).
Privacy budgets $\epsilon_1,\epsilon_2 \in [0,+\infty)$.
Output : The perturbed tuple: (g',v').

1 $g' \leftarrow \mathcal{M}_{\text{GRR}}(g;d,\frac{e^{\epsilon_1}}{e^{\epsilon_1}+d-1})$
2 $v' \leftarrow v$ if $g=g'$ and $v' \leftarrow 0$ otherwise
3 $v' \leftarrow discretize(v')$
4 $v' \leftarrow \mathcal{M}_{\text{GRR}}(v';2,\frac{e^{\epsilon_2}}{1+e^{\epsilon_2}})$
5 **return** (g',v')

Algorithm 1 describes the mechanism. The mechanism first applies $\mathcal{M}_{\text{GRR}}$ to perturb the group (line 1). Only if GRR perturbed the group, the mechanism sets the value to 0 (line 2), introducing a correlation between the group and the value. In line 3, the mechanism discretizes the value to $\{-1,1\}$ following the discretization primitive of Harmony [53], defined by

$$discretize(v) = \begin{cases} 1 & \text{w.p. } \frac{1+v}{2} \\ -1 & \text{w.p. } \frac{1-v}{2} \end{cases}$$

This discretization does not bias the estimates of the values: $\mathbb{E}[\tilde{v}] = \frac{1+v}{2} - \frac{1-v}{2} = v$, where v and $\tilde{v}$ denote the values before and after the discretization, and the expectation is taken over the coin tosses of the discretization primitive. In addition, because the clients with perturbed group discretize the value uniformly, $\mathbb{E}[\tilde{v}] = 0$ for those clients. Therefore, they do not contribute to the other group's mean value. These two observations will be relevant when we analyze the error of the statistics on the perturbed data.

After the discretization, the values are still unprotected, thus the mechanism applies $\mathcal{M}_{\text{GRR}}$ on them (line 4).

Now we show that this mechanism achieves ϵ-LDP with ϵ the maximum of the privacy budget of the two individual RRs.

Theorem 4.1. $\mathcal{M}_{\text{R}}$ is ϵ-LDP with $\epsilon = \max\{\epsilon_1, \epsilon_2\}$.

PROOF. We denote $a = \frac{e^{\epsilon_1}}{e^{\epsilon_1}+d-1}$ and $b = \frac{e^{\epsilon_2}}{1+e^{\epsilon_2}}$. Let $x_0 = (g_0, v_0)$ and $x_1 = (g_1, v_1)$ be two different inputs and $y = (g', v')$ be an output of the mechanism. From the mechanism's definition, we have that for an arbitrary input $x = (g, v)$,

$$\Pr[y \mid x] = \begin{cases} \frac{a(1+(2b-1)v'v)}{2} & \text{if } g' = g \\ \frac{1-a}{2(d-1)} & \text{if } g' \neq g \end{cases}$$

Since $v \in [-1,1]$ and $v' \in \{-1,1\}$, an upper bound of $\Pr[y \mid x]$ when $g' = g$ is

$$\Pr[y \mid x] \leq \frac{ab}{2} \tag{1}$$

and a lower bound is

$$\Pr[y \mid x] \geq a(1-b) \tag{2}$$

Now, we bound $\Pr[y \mid x_0]/\Pr[y \mid x_1]$, where x_0 and x_1 differ in either group or value. If they have the same group but differ in value, we consider two cases: $g' = g$ and $g' \neq g$.

Case 1: $g' = g$. Using the upper- and lower-bounds, we obtain:

$$\frac{\Pr[y \mid x_0]}{\Pr[y \mid x_1]} \leq \frac{ab}{2a(1-b)} = \frac{e^{\epsilon_2}}{2} \leq e^{\epsilon_2} \tag{3}$$

Case 2: $g' \neq g$. Using the probability of $\Pr[y|x_1]$ when $g' \neq g$:

$$\frac{\Pr[y|x_0]}{\Pr[y|x_1]} = 1 \leq e^{\epsilon_2}, \text{ as } \epsilon_2 \in [0, +\infty) \tag{4}$$

This shows that if the inputs have the same group, the differential privacy guarantee boils down to the guarantee of the value-perturbing GRR mechanism.

If x_0 and x_1 differ in group and may or may not differ in value, we again break down the analysis into two cases: $g' = g$ and $g' \neq g$.

Case 1: $g' = g$. Using the upper-bound, we obtain:

$$\frac{\Pr[y|x_0]}{\Pr[y|x_1]} \leq \frac{ab}{1-a} = \frac{e^{\epsilon_2+\epsilon_1}}{1+e^{\epsilon_2}} \leq e^{\epsilon_1} \tag{5}$$

Case 2: $g' \neq g$. Using the lower-bound, we have:

$$\frac{\Pr[y|x_0]}{\Pr[y|x_1]} \leq \frac{1-a}{2a(1-b)} = \frac{1+e^{\epsilon_2}}{2e^{\epsilon_1}} \leq \frac{2e^{\epsilon_2}}{2e^{\epsilon_1}} = e^{\epsilon_2-\epsilon_1} \tag{6}$$

Combining the equations above, we conclude that $\mathcal{M}_{\mathrm{R}}$ is ϵ-DP with $\epsilon = \max\{\epsilon_1, \epsilon_2, \epsilon_2 - \epsilon_1\} = \max\{\epsilon_1, \epsilon_2\}$. □

This result bounds the ratio of the probabilities in Definition 2.1; it is not only tighter than the bound that the basic theorem on sequential composition of mechanisms [51] gives, but it is a tight bound. The tightness of the LDP bound is important to provide an uppper bound for the privacy of the mechanism, when comparing it with other mechanisms as well as quantifying the privacy vs. utility trade-off.

4.2 $\mathcal{M}_{\mathrm{Lap}}$-Based Mechanisms

In this section, we describe $\mathcal{M}_{\mathrm{L}}$, a mechanism that uses $\mathcal{M}_{\mathrm{Lap}}$ to perturb the values. In Algorithm 2, we detail the steps of this mechanism. As in $\mathcal{M}_{\mathrm{R}}$, the group is perturbed with $\mathcal{M}_{\mathrm{GRR}}$ (line 1) but, to perturb the value, it instead samples additive noise from a Laplace distribution, following $\mathcal{M}_{\mathrm{Lap}}$ (lines 2-6). The scale of the noise may vary depending on whether the client's group has been perturbed. Clients whose group flipped do not require as much noise to *hide* in the value distribution of the other group, as those values are also perturbed with Laplace noise. We parametrize the scale of the Laplace distribution for the additive noise that the mechanisms add to the clients who swapped by $0 < k \leq 2$. In addition, the value of the clients that swap to the other group is set to zero, such that, like in $\mathcal{M}_{\mathrm{R}}$, they do not contribute to that group's mean value – their expected value is zero.

Algorithm 2: Privacy Mechanism: $\mathcal{M}_{\mathrm{L}}$.

Input : The client's group and value: (g, v).
Privacy budgets $\epsilon_1, \epsilon_2 \in [0, +\infty)$, and $k \in (0, 2]$.
Output : The perturbed tuple: (g', v').

1 $g' \leftarrow \mathcal{M}_{\mathrm{GRR}}(g; d, \frac{e^{\epsilon_1}}{e^{\epsilon_1}+d-1})$
2 **if** $g = g'$ **then**
3 $\quad v' \leftarrow \mathcal{M}_{\mathrm{Lap}}(v; \epsilon_2)$
4 **else**
5 $\quad v' \leftarrow \mathcal{M}_{\mathrm{Lap}}(0; 2\frac{\epsilon_2}{k})$
6 **return** (g', v')

Even though the scale of the noise that we add to the values of users who flip is smaller than the scale required to achieve ϵ_2-LDP, we show that $\mathcal{M}_{\mathrm{L}}$ achieves ϵ-LDP for an ϵ that is a function of k, ϵ_1 and ϵ_2.

Theorem 4.2. The mechanism $\mathcal{M}_{\mathrm{L}}$ is ϵ-LDP with

$$\epsilon = \max\left\{\epsilon_2, \ln(\frac{2}{k}) + \frac{\epsilon_2}{2} - \epsilon_1, \ln(\frac{k}{2}) + \frac{\epsilon_2}{k} + \epsilon_1\right\} \tag{7}$$

Proof. Let $x_0 = (g_0, v_0)$ and $x_1 = (g_1, v_1)$ be two different inputs and $y = (g', v')$ be an output of the mechanism. From the mechanism's definition, we have that for an arbitrary input $x = (g, v)$,

$$\Pr[y|x] = \begin{cases} \frac{e^{\epsilon_1}}{e^{\epsilon_1}+d-1} f_{\mathrm{Lap}(0, \frac{2}{\epsilon_2})}(v' - v) & \text{if } g' = g \\ \frac{1}{e^{\epsilon_1}+d-1} f_{\mathrm{Lap}(0, \frac{k}{\epsilon_2})}(v') & \text{if } g' \neq g \end{cases}$$

This is because when the mechanism preserves the group, $v' = v + Y$ where $Y \sim \mathrm{Lap}(0, \frac{2}{\epsilon_2})$, hence the probability of the new value is the probability of sampling $v' - v$ from the Laplace distribution with zero mean and scale of $\frac{2}{\epsilon_2}$. When the group is flipped, the mechanism sets v to zero, thus the probability of the new value is just the probability of sampling v' from $\mathrm{Lap}(0, \frac{k}{\epsilon_2})$.

As in the proof of Theorem 4.1, we follow a case-based reasoning. If x_0 and x_1 have the same group but differ in value, we consider two cases: $g' = g$ and $g' \neq g$.

Case 1: $g' = g$.

$$\frac{\Pr[y|x_0]}{\Pr[y|x_1]} \leq e^{\epsilon_2} \tag{8}$$

Case 2: $g' \neq g$.

$$\frac{\Pr[y|x_0]}{\Pr[y|x_1]} = 1 \tag{9}$$

If x_0 and x_1 differ in group, we again consider two cases: $g' = g$ and $g' \neq g$.

Case 3: $g' = g$.

$$\frac{\Pr[y|x_0]}{\Pr[y|x_1]} = \frac{\frac{e^{\epsilon_1}}{e^{\epsilon_1}+d-1} f_{\mathrm{Lap}(0, \frac{2}{\epsilon_2})}(v' - v_0)}{\frac{1}{e^{\epsilon_1}+d-1} f_{\mathrm{Lap}(0, \frac{k}{\epsilon_2})}(v')} = \frac{k}{2} e^{\epsilon_1 + \epsilon_2 (\frac{|v'|}{k} - \frac{|v'-v_0|}{2})} \leq \frac{k}{2} e^{\epsilon_1 + \frac{\epsilon_2}{k}} \tag{10}$$

Observe that, for any v', $|v' - v_0|$ is 0 when v_0 is v'. Therefore, $\frac{|v'|}{k} - \frac{|v'-v_0|}{2} \leq \frac{|v'|}{k} \leq \frac{1}{k}$ and the value $\frac{1}{k}$ is attained when $v' \in \{-1, 1\}$ and $v_0 = v'$.

Case 4: $g' \neq g$.

$$\frac{\Pr[y|x_0]}{\Pr[y|x_1]} = \frac{2}{k} e^{\epsilon_2 (\frac{|v'-v_1|}{2} - \frac{|v'|}{k}) - \epsilon_1} \leq \frac{2}{k} e^{\frac{\epsilon_2}{2} - \epsilon_1} \tag{11}$$

The inequality follows from the triangle inequality: $\frac{|v'-v_1|}{2} - \frac{|v'|}{k} \leq \frac{|v'-v_1|}{2} - \frac{|v'|}{2} \leq \frac{|v_1|}{2} \leq \frac{1}{2}$. The maxima are in the extremes of $[-1, 1]$.

Finally, combining all the inequalities above, we obtain that the probability ratio must be bounded by

$$e^{\max\left\{\epsilon_2, \ln(\frac{2}{k}) + \frac{\epsilon_2}{2} - \epsilon_1, \ln(\frac{k}{2}) + \frac{\epsilon_2}{k} + \epsilon_1\right\}}$$

□

As shown by Theorem 4.1 and Theorem 4.2, the privacy budgets of $\mathcal{M}_{\mathrm{R}}$ and $\mathcal{M}_{\mathrm{L}}$ have a different impact on the LDP bound of the mechanism. In Section 5.2, we show how to allocate these privacy

budgets and how to choose k, such that they minimize the error on utility for a fixed overall privacy budget ϵ.

5 PERFORMANCE EVALUATION

Unsurprisingly, ensuring local differential privacy of the group membership and performance value, will introduce an error into the accuracy of the performance gap measurements that are possible. What we aim to understand next is how this error varies depending on the mechanism choice and the desired level of the differential privacy guarantee.

To measure the privacy-induced error, we follow the LDP literature by treating the measurement as an estimator of m_G under the randomness of the mechanisms. A key metric of the quality of an estimator is its Mean Squared Error (MSE), as it captures the error due to both the estimator's variance and its bias. By showing that our estimators are unbiased, we can compare their MSEs by simply comparing their variance (Section 5.2). Further, knowing the estimators' variance allows us to probabilistically bound the distance of a performance gap measurement to its true value as a function of the number of clients, which is informative to assess the feasibility of our mechanisms in a DFL setting (Section 5.4).

Our results show that, although the mechanisms achieve similar performance, neither outperforms the other for all possible privacy budgets. They both allow for accurate measurements while ensuring that the error is below a certain threshold for numbers of clients orders of magnitude lower than the ones of current DFL deployments.

5.1 Unbiased Estimators

We now describe the estimators that the operator should use to estimate m_G from the perturbed group-performance value tuples.

Definition 5.1 (Estimators of m_G)**.** The estimators for the mechanisms are defined by

$$\hat{m}_G^L := \frac{1}{an}\sum_{j=1}^{n'} v'_j, \qquad \text{and} \qquad \hat{m}_G^R := \frac{1}{a(2b-1)n}\sum_{j=1}^{n'} v'_j, \tag{12}$$

where $a = \frac{e^{\epsilon_1}}{e^{\epsilon_1}+d-1}$ and $b = \frac{e^{\epsilon_2}}{1+e^{\epsilon_2}}$, and n' is the number of clients in group G after the mechanism's perturbation.

With these estimators, we are assuming that n can be approximated by the operator of the mechanism. This is not an unrealistic assumption for many demographics, for which the Census or other public data can provide an accurate guess. In Section 5.3, we provide a more detailed examination about the impact of an error in the number of clients on our MSE analysis. Our findings show that our analysis stands for up to 50% relative error in the estimations of K/n (recall that K is the total number of clients). Moreover, as discussed in Section 6.4, hypothesis testing techniques for RR can be applied on our mechanisms to test how n deviates from its estimated value [59].

We first show that these estimators are unbiased and hence MSE is just the variance of the estimators.

Proposition 5.1. The estimators of the mechanisms are unbiased, i.e., $\mathbb{E}[\hat{m}_G^L] = \mathbb{E}[\hat{m}_G^R] = m_G$.

Proof. We prove that $\hat{m}_G^L$ is unbiased. The proof for the unbiasedness of $\hat{m}_G^R$ is analogous.

We model the values in G after applying $\mathcal{M}_{\mathrm{L}}$ with the following mutually independent random variables

$$V_i = B_i(v_i + Y_i), \quad i = 1,..., n, \tag{13}$$

$$\bar{V}_j = \bar{B}_j(0 + \bar{Y}_j) = \bar{B}_j\bar{Y}_j, \quad j = 1,..., K-n \tag{14}$$

where V_i and $\bar{V}_j$ are the final, perturbed values in group G that originate from group G and $\bar{G}$, respectively. In our notation, the bar denotes that the random variable relates to group $\bar{G}$, the complement of G. The random variables $B_i \sim \text{Bernoulli}(a)$ and $\bar{B}_j \sim \text{Bernoulli}(1-a)$ model $\mathcal{M}_{\mathrm{RR}}$, and $Y_i \sim \text{Lap}(0, 2/\epsilon_2)$ and $\bar{Y}_j \sim \text{Lap}(0, k/\epsilon_2)$ model $\mathcal{M}_{\mathrm{Lap}}$. Thus, the expected value of the estimator is

$$\mathbb{E}[\hat{m}_G^L] = \frac{1}{an}\left(\sum_{i=1}^{n}\mathbb{E}[V_i] + \sum_{j=1}^{K-n}\mathbb{E}[\bar{V}_j]\right) \quad \text{Linearity of } \mathbb{E} \tag{15}$$

$$= \frac{1}{an}\sum_{i=1}^{n}\mathbb{E}[B_i(v_i + Y_i)] \quad \mathbb{E}[\bar{V}_j] = 0 \tag{16}$$

$$= \frac{1}{an}\sum_{i=1}^{n}\mathbb{E}[B_i](v_i + \mathbb{E}[Y_i]) \quad \text{Mutual independence} \tag{17}$$

$$= \frac{1}{an}\sum_{i=1}^{n}\mathbb{E}[B_i]v_i \quad \mathbb{E}[Y_i] = 0 \tag{18}$$

$$= \frac{a}{an}\sum_{i=1}^{n} v_i \quad \mathbb{E}[B_i] = a \tag{19}$$

$$= m_G \tag{20}$$

We used that $\mathbb{E}[\bar{V}_j] = 0$ because $\mathbb{E}[\bar{Y}_j] = 0$ and that the random variables are mutually independent. □

Next, we give closed-form expressions for the variance of these estimators.

Proposition 5.2.

$$\text{Var}[\hat{m}_G^R] = \frac{1}{a(2b-1)^2 n}\left(1 - a(2b-1)^2\nu^2 + \frac{K-n}{n}\frac{1-a}{a}\right) \tag{21}$$

$$\text{Var}[\hat{m}_G^L] = \frac{1}{n}\left(\nu^2 e^{-\epsilon_1} + \left(1+e^{-\epsilon_1}\right)\left(\sigma_L^2 + \frac{K-n}{n}\bar{\sigma}_L^2 e^{-\epsilon_1}\right)\right) \tag{22}$$

where $\nu^2 = \frac{1}{n}\sum_{i=1}^{n} v_i^2$, and $\sigma_L^2, \bar{\sigma}_L^2$ are the variances of the Laplace noise distributions, for clients who do not swap and those who do, respectively. We obtain lower and upper bounds of these variances using that $0 \le \nu^2 \le 1$.

Proof outline. Using the probabilistic model defined in the proof of Proposition 5.1, we can write the variance of the estimator as

$$\text{Var}[\hat{m}_G^L] = \frac{1}{a^2 n^2}\text{Var}\left[\sum_{i=1}^{n}(v_i + Y_i)B_i + \sum_{j=1}^{K-n}\bar{Y}_j\bar{B}_j\right].$$

Note that the noise terms have positive variance and therefore do not cancel out. We can use the fact that the variables are mutually independent to write the variance of the sum as the sum of variances.

We will then obtain variances of products and will use the well-known formula for the variance of the product of two independent random variables. Rearranging the terms gives the expression in the Proposition 5.2.

The proof for $\hat{m}^R_G$ is analogous.

Definition 5.2 (Estimator of Δm)**.** The estimator of the performance gap is $\Delta\hat{m}^* := \left|\hat{m}^*_G - \hat{m}^*_{\bar{G}}\right|$, where the asterisk is a wildcard for the mechanism.

Theorem 5.3 shows that the variance of $\Delta\hat{m}^*$ is the sum of the variances of the group mean value estimates.

Theorem 5.3. For both mechanisms, $\Delta\hat{m}^*$ is an unbiased estimator of Δm and its variance is

$$\mathrm{Var}\left[\Delta\hat{m}^*\right] = \mathrm{Var}[\hat{m}^*_G] + \mathrm{Var}[\hat{m}^*_{\bar{G}}].$$

PROOF OUTLINE. First, we prove the unbiasedness of $\Delta\hat{m}^*$. Due to the Proposition 5.1 and the linearity of expectation, the expected value of $\Delta\hat{m}^*$ is Δm. Assuming that G is the advantaged group and thus $\hat{m}^*_G \geq \hat{m}^*_{\bar{G}}$, we have that $\mathbb{E}[|\hat{m}^*_G - \hat{m}^*_{\bar{G}}|] = |m_G - m_{\bar{G}}|$.

To show that the variance of $\Delta\hat{m}^*$ is the sum of the variance of the mean group value estimators, it suffices to show that $Cov(\hat{m}^*_G, \hat{m}^*_{\bar{G}}) = 0$, which is true if, and only if, $\mathbb{E}[\hat{m}^*_G \hat{m}^*_{\bar{G}}] = m_G\, m_{\bar{G}}$. Calculating the value of that expectation explicitly, we observe that many of its terms have an independent Laplace r.v. as a factor and, consequently, these terms are zero. Finally, we can apply Bienaymé's identity to obtain the result of the theorem.

The proof for mechanism $\mathcal{M}_\mathrm{R}$ is similar, as the expected value of clients with the group perturbed is zero.

The intuition behind Theorem 5.3 is that even though $\hat{m}^*_G$ and $\hat{m}^*_{\bar{G}}$ are not independent, the errors are uncorrelated, and thus they add up.

We have validated all the theoretical results numerically. In Appendix A, we show an example of the validation of Theorem 5.3.

5.2 Allocation of the Privacy Budget

Combining Theorem 5.3 with Proposition 5.2 we can obtain a closed-form expression for the variance and hence the MSE of $\Delta\hat{m}^*$. Such expression allows to compare the estimators for specific values of the parameters of the mechanisms, such as the privacy budget allocated to the $\mathcal{M}_\mathrm{GRR}$ and $\mathcal{M}_\mathrm{Lap}$ components. It is unclear a priori how to divide a fixed privacy budget into the mechanism's components to maximize utility. Our approach is to find an allocation that minimizes the upper bound of the MSE of the estimators, for the total privacy budget of the mechanism (ϵ).

Mechanism $\mathcal{M}_\mathrm{R}$. For a fixed ϵ, the optimal allocation is $\epsilon_1 = \epsilon_2 = \epsilon$, as it minimizes the MSE under the LDP constraint.

Mechanism $\mathcal{M}_\mathrm{L}$. In Eq. (22), we see that the variance of the unbiased estimator for $\mathcal{M}_\mathrm{L}$ is dominated by ϵ_2. Therefore, since ϵ_1,ϵ_2, and k must satisfy Eq. (7), we minimize the MSE by first setting $\epsilon_2 = \epsilon$ and, then, finding the k that maximizes ϵ_1 under the LDP constraint in Eq. (7).

If we take $\epsilon_2 = \epsilon$ in Eq. (7), we obtain bounds for ϵ_1

$$\ln(\frac{2}{k}) - \frac{\epsilon}{2} \leq \epsilon_1 \leq \ln(\frac{2}{k}) + \frac{\epsilon}{2}\lambda(k), \tag{23}$$

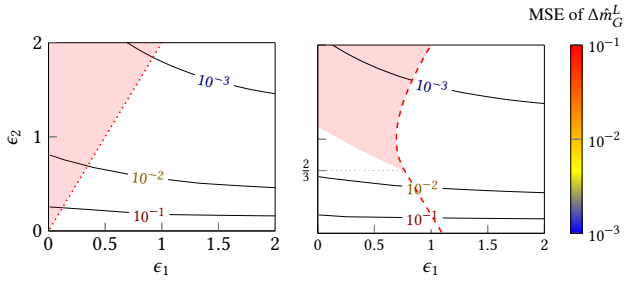

Figure 1: Contour plot of the MSE of $\Delta\hat{m}^L$ for $k = 2$ (left) and $k = 2/3$ (right), as a function of ϵ_1 and ϵ_2. The colored area is the region where the parameters satisfy ϵ-LDP for $\epsilon = \epsilon_2$. The curves represent the optimal allocations of $\mathcal{M}_\mathrm{L}$ (dotted) and $\mathcal{M}_\mathrm{L}$ opt. (dashed).

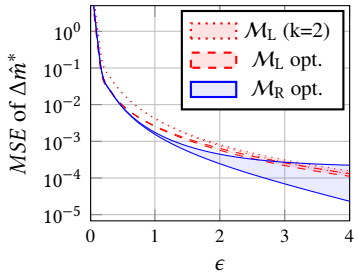

Figure 2: Upper and lower bounds of the estimators' MSE for different overall privacy budgets ϵ. We have set $n_G = n_{\bar{G}} = 10^4$.

where $\lambda(k) = 2\left(1 - \frac{1}{k}\right)$. Thus, this inequality holds iff $\frac{2}{3} \leq k$.

To find the k that maximizes ϵ_1, we consider two cases: $0 < \epsilon < 2/3$, and $2/3 \leq \epsilon$. If $2/3 \leq \epsilon$, we write ϵ_1 as the upper bound of ϵ in Eq. (23), a function of k, and find that $k = \epsilon$ is a maximum for a constant ϵ. However, for $0 < k < 2/3$, Eq. (23) does not hold and hence $k = \epsilon$ would not satisfy ϵ-LDP. When $0 < \epsilon < 2/3$, we take $k = 2/3$, the minimum k that satisfies ϵ-LDP, as that minimizes the scale of the Laplace noise. In that case, ϵ_1 is equal to the upper and lower bounds in Eq. (23).

Thus, the maximum ϵ_1 as a function of ϵ is

$$\epsilon_1 = \begin{cases} \ln(\frac{2}{\epsilon}) + \epsilon - 1 & \text{if } \frac{2}{3} \leq \epsilon \\ \ln(3) - \frac{\epsilon}{2} & \text{if } 0 < \epsilon < \frac{2}{3} \end{cases}$$

Fig. 1 shows the allocations of the privacy budgets that satisfy the LDP constraint (colored area). The dashed and dotted borders of the area show the allocations that minimize the MSE for a total privacy budget of $\epsilon = \epsilon_2 \in (0,2]$ for $k = 2$ and optimal k, respectively. A closer look at the MSE contour lines reveals that the mechanism with optimal k achieves lower MSE values than for $k = 2$. This difference is apparent in Fig. 2, where we plot the MSE of the mechanisms as a function of ϵ for two groups of the same size.

In Fig. 2, we plot the MSE for the performance gap estimator for two groups G and $\bar{G}$ of the same size. We observe that $\mathcal{M}_\mathrm{L}$ opt. achieves lower MSE than $\mathcal{M}_\mathrm{L}$ with $k = 2$ for the range that we consider. $\mathcal{M}_\mathrm{R}$ has lower MSE than $\mathcal{M}_\mathrm{L}$ opt. only when $0.31 \lessapprox \epsilon \lessapprox 2.6$

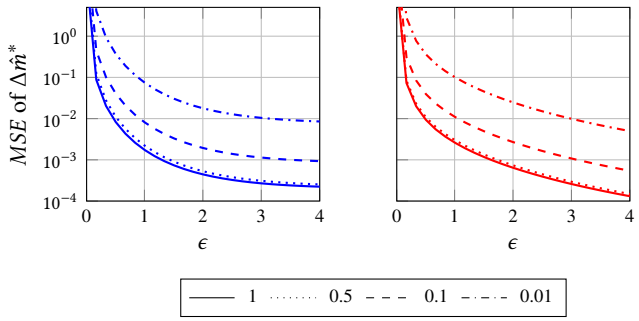

Figure 3: Upper bound of the MSE of $\mathcal{M}_{\mathrm{R}}$ (left) and $\mathcal{M}_{\mathrm{L}}$ (right) for a fixed $K = 2 \cdot 10^4$ and different group ratios $n_G / n_{\bar{G}}$.

but, for $\epsilon \gtrapprox 2.6$, the upper bound of $\mathcal{M}_{\mathrm{R}}$'s MSE is larger than the upper bound of $\mathcal{M}_{\mathrm{L}}$'s and the gap between the two rapidly widens for larger privacy budgets. As a consequence, the operator should select the mechanism that best suits their privacy budget.

5.3 Unbalanced Groups

We also study the impact that an unbalance of the group sizes has on the MSE of the estimators. Fig. 3 depicts the upper bounds of the MSEs, $\mathcal{M}_{\mathrm{R}}$ opt. (left) and $\mathcal{M}_{\mathrm{L}}$ opt. (right), for two groups with different size ratios. We observe that for typical privacy budgets (around 1), the mechanisms can cope with relatively small groups, but the MSE rapidly grows for minorities that account for less than 1% of the client population. Consequently, the mechanisms would incur in high MSE with protected attributes that include small minorities, but would maintain low MSEs for common protected attributes, such as gender and race, in scenarios where they follow a distribution similar to the the US Census distribution.

Throughout the paper we assume that the entity performing the measurements knows n. The figures in this section also show the difference of the mechanisms' MSE when the group fractions are incorrectly estimated. For example, when the groups are balanced and the estimate of the group fractions has a relative error of up to 50%, the difference in the MSE values (i.e., the difference between the bold and dotted lines) is insignificant. This means that the assumption of knowing n can be relaxed in practice.

5.4 Confidence Intervals

In this section, we provide confidence intervals for the estimates of Δm as a function of the number of clients. Our results show that the mechanisms allow for high-precision measurements with realistic numbers of clients in the DFL setting.

Using Chebyshev's inequality and Theorem 5.3, we can give confidence intervals for the distance between the estimator and its expected value, Δm, as a function of the number of clients.

By Chebyshev's inequality, for any real number $M > 0$,

$$\Pr\left[|\Delta\hat{m}^* - \Delta m| < M\sqrt{\mathrm{Var}[\Delta\hat{m}^*]}\right] \geq 1 - \frac{1}{M^2}. \tag{24}$$

If we assume $n_G / n_{\bar{G}}$ constant for all G, i.e., adding more clients does not change the distribution of group sizes, we can factor out $1/K$ from the upper bounds of $\mathrm{Var}[\Delta\hat{m}^*_G]$. Then, we can use Eq. (24) to numerically find the value of ϵ (the overall privacy budget) such that $|\Delta\hat{m}^*_G - \Delta m| < \alpha$, for a small $\alpha > 0$ with high probability. The α gives a confidence interval for the performance gap measurement.

As in the evaluations above, we assume a protected attribute that partitions the clients into two groups. Table 1 shows the minimum privacy budget required to ensure that the error is at most α for various values of α and numbers of clients, with probability 0.99. If the operator can afford a higher privacy budget, the bound would still hold but if their privacy budget is lower, the mechanism does not guarantee the bounds.

Table 1: Minimum privacy budget (ϵ) required to bound the error by α, given K clients, with 0.99 probability. Highlighted are the ϵ's that are considered reasonable in LDP applications.

	$\mathcal{M}_{\mathrm{R}}$			$\mathcal{M}_{\mathrm{L}}$		
K	$\alpha = 10^{-1}$	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$	$\alpha = 10^{-1}$	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$
10^5	1.86	29.78	30.45	2.56	17.89	178.89
10^6	0.63	34.02	29.18	0.71	6.32	56.57
10^7	0.23	1.86	28.60	0.21	2.56	17.89
10^8	0.08	0.63	35.93	0.07	0.71	6.32
10^9	0.02	0.23	1.86	0.02	0.21	2.56

We only consider performance metrics that can be interpreted as percentages, and represent them as a number in the unit interval. Therefore, the most relevant error bounds are the ones of at most one percentage point (columns $\alpha = 10^{-2}$ and $\alpha = 10^{-3}$). These results show that $\mathcal{M}_{\mathrm{R}}$ achieves a better trade-off than $\mathcal{M}_{\mathrm{L}}$ between privacy and error in the measurements except when $\epsilon < 0.31$. The budgets that the mechanism operators have to afford to achieve a one percentage point and a decimal place precision are reasonable for $K \geq 10^7$ and $K \geq 10^9$ clients, respectively. Even though these may look like large numbers of clients, scenarios where DFL is deployed have this many, and even more, clients. For example, in 2018 Apple reported a total of half a billion active Siri clients[2] and, in the same year, Gboard surpassed 1 billion installs, according to Google Play Store[3] statistics.

5.5 Empirical Evaluation

In this section, we describe the experiments to empirically evaluate the error of the mechanisms. Since we are not aware of public datasets with sufficient data to model a real-world deployment of DFL, we synthesize a dataset by fitting the marginal probability distributions of the protected attribute on a real-world dataset. Our results show that the error of the mechanisms in the synthetic data is orders of magnitude lower than the Chebyshev bounds obtained in the previous section, indicating that an operator who uses the Chebyshev bounds might be overly conservative in their user privacy risk assessment.

Data Generation. Our data generation model is based on the dataset collected by Torres et al. [60] to assess the effectiveness of activity detection, in this case of patients in a hospital room, using a wearable sensor. The dataset comprises the sensor readings for 14 subjects aged between 66 and 86 years old who were instructed to perform a number of scripted daily activities in two different rooms.

[2] https://www.apple.com/newsroom/2018/01/homepod-arrives-february-9-available-to-order-this-friday

[3] https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin

Each subject participated in approximately five trials, resulting in 87 recorded sessions. The features include the sensor's readings of time, accelerometer position, and radio signal's strength, frequency, and phase. The labels describe the activity in which the participant engaged during the recording of the sample: sitting, lying down, or ambulating. We binarized the detection task by relabeling the data to whether or not the subject was lying down.

We define "sex" as the protected attribute in the data. Although the sex of the subject was annotated per each trial – 25 male and 62 female – there is no mapping between trials and subjects. Thus, we assume that each recorded session represents the data of a different FL client, with each client having an average of 864 samples. We split the data into training and testing by stratification: ensuring that all clients are present in both datasets, and that they all have the same distribution of samples between training and test sets: 70% of the samples for training and 30% for testing. This split also ensures that the data preserves the original distribution of the sex attribute in the dataset.

We simulated the federated learning of a model by training a logistic regressor on the training set. We assume that this is the global model trained with the data of all clients. Since the performance of the model was nearly perfect, resulting in almost all the clients having a zero false positive rate, we have dropped some of the accelerometer features to increase the difficulty of the learning task. The global model's accuracy on the test set is 95.66%, with a false positive rate (FPR) of 8.13%, and a true positive rate (TPR) of 97.42%. Then, independently, we test the global model on each client's test set, resulting in two performance values for each client: we take the true positive (TPR) and false positive rates (FPR) as performance metrics. The mean TPRs are 99.93% and 77.87% and the mean FPRs are 10.64% and 25.11% for males and females, respectively. We observe a significant performance gap on both metrics: $\Delta\text{TPR} = 22.06\%$ and $\Delta\text{FPR} = 14.64\%$.

Table 2: Comparison of the Chebyshev bounds with the empirical mean error of 10 different runs of the mechanisms on the synthetic dataset for $K = 10^7$ clients and various privacy budgets. The first column is the privacy budget, followed by the mean error (and standard deviation) of the estimates on the data and the 0.99-probability Chebyshev's bounds (α) for each mechanism.

	$\mathcal{M}_\text{R}$		$\mathcal{M}_\text{L}$	
ϵ	$\lvert\Delta\hat{m}^R - \Delta m\rvert$	α	$\lvert\Delta\hat{m}^L - \Delta m\rvert$	α
0.01	0.1241(±0.1410)	1.2586	0.0504(±0.0337)	1.0525
0.10	0.0082(±0.0059)	0.1206	0.0046(±0.0040)	0.1060
1.00	0.0008(±0.0006)	0.0094	0.0008(±0.0005)	0.0118
10.00	0.0001(±0.0000)	0.0032	0.0001(±0.0000)	0.0009

Error of the DP mechanism. To generate synthetic data for the global model's performance on new clients, we model the marginal distribution of sex to have the same mean and v^2 as the observations. For the purpose of evaluating the error of the mechanisms, the exact distribution that we fit is not important, thus we draw samples with replacement from the set of observations. This sampling methodology ensures that the relevant statistics are preserved and we can scale up the dataset to the number of clients of a realistic DFL deployment.

We draw a balanced sample of the model's TPR for 10^7 clients and apply the mechanisms on the sex and performance value tuples of the synthetic data. Table 2 compares the empirical error with the 0.99-probability bounds (α) obtained with the procedure explained in the previous section, for different privacy budgets (ϵ). The bounds are one order of magnitude larger than the actual error. This means that the budget that the operator would need to allocate to satisfy a certain α for 10^7 clients is substantially lower than the ones shown in Table 1. As a consequence, following the Chebyshev bounds from the previous section would result in an overly conservative measurement with respect to the privacy of the users, and operators with small privacy budgets could afford more accurate measurements without an impact on user privacy.

Finally, we have also obtained the 0.99-probability Chebyshev bounds using the *lower bounds* of the mechanisms' variance (corresponding to the bottom lines in the regions depicted in Fig. 2). Although our theoretical analysis does not guarantee that these bounds would hold, as for some sets of users the mechanisms may exhibit more variance, our results show that they are close to the upper bounds shown in Table 2. This small difference between the Chebyshev bounds obtained with the upper and lower bounds of the variance suggests that the large gap between the empirical error and its theoretical bound is due to the looseness of the Chebyshev's bounds.

Due to its generality, the Chebyshev's inequality gives loose bounds for most probability distributions. It is remarkable that despite the looseness of these theoretical bounds, our results (in Table 1) show that the privacy budgets required to perform an accurate and private measurement of the performance gap are affordable by the scale of current DFL deployments. Since the results in Table 2 show that privacy budget required to achieve the same privacy level would be even lower, we can safely conclude that the large scale of existing DFL deployments allows for private and accurate measurements of the performance gap via our mechanisms.

6 DISCUSSION

This section is structured as follows: in Section 6.1, we argue that due to its low costs and its robustness to client dropouts, our approach is better suited to the DFL setting than prior works aiming to detect disparities. Section 6.2 details a potential implementation of our mechanisms, and illustrates their high practical feasibility through an example application in the speech recognition domain. In Section 6.3, we describe the mitigation strategies that a measurement with our mechanisms would enable and, in Section 6.4, we discuss potential extensions of our work.

6.1 Comparison to Alternative Approaches

Our approach is superior to other approaches considered in the literature in terms of efficiency of implementation, data collection costs, and robustness to client dropouts. We now discuss these strengths in more detail.

Low computational and communication costs. First, the computational cost of adding LDP noise is by construction distributed among the clients and, in contrast to SMPC, our mechanisms do

not have the communication and computation costs associated with the key setup phase. These low costs make our mechanisms ideal candidates for a measurement in DFL as, in such setting, the clients often run on resource-constrained devices (e.g., mobile phones and smart speakers) where efficiency is critical.

Inexpensive and representative data. Previous studies have pointed out that obtaining labeled datasets with sensitive attributes is a major challenge in auditing for algorithmic bias [15, 57]. Recruiting volunteers for these measurements is costly and the costs are unequal across locations, leading to representation issues in the data [15]. The advantage of our approach is that it can piggyback on the existing FL infrastructure, reducing the costs of data collection and eliminating the need inherent in prior auditing methods to obtain a representative sample.

Robustness to client dropouts. In contrast to other solutions, such as those based on SMPC, our LDP mechanisms are robust to client dropouts. If clients end up not participating in the measurement, they can be replaced by dummy clients that send values uniformly at random. Although this approach may increase the error in the estimates of the performance gap, it does not affect the privacy guarantees.

6.2 Implementation Details

In this section, we detail how our proposed mechanisms can be easily implemented by the infrastructure provider or a regulatory agency and illustrate it with an example application. We also clarify that the usage of the mechanisms is within the paradigm of existing LDP deployments and that our ultimate objective is to facilitate the measurement of the performance gap, rather than to constantly monitor it.

Deployment use case. Our approach is practical and can be easily deployed by the infrastructure provider or a regulator with limited access to the infrastructure. To illustrate an implementation of our mechanisms, we go back to the motivating example of applying FL to improve wake-word detection.

Amazon researchers argue that FL in Alexa would allow access to higher-quality data and, potentially, more accurate wake-word detection [64]. The challenge is that Alexa devices do not have local data to train a wake-word detection model. To obtain such data, the clients could use cues to label voice samples [64]. For example, raising the tone or pushing a button on the device could indicate a false negative, while false positives could be detected if the speech is followed by silence or non-command language.

Although there is limited research on the potential biases of wake-word detection models specifically, performance disparities by gender and ethnicity have been found across many existing speech recognition models [5, 62], including Alexa's model [38]. Given that wake-word detection uses speech recognition techniques, these findings suggest that similar disparities exist in wake-word detection models.

A performance disparity in wake-word detection can lead to an undue privacy violation for members of some of the demographic groups. For example, if the model exhibits higher FPRs for a particular group, unrelated speech might be uploaded to the cloud for further processing more often than for other groups, exposing that group to more surveillance. This concern is especially relevant given the privacy aspirations of FL: FL aims to minimize the data that clients have to share with the aggregator, but a performance disparity may lead to disparate privacy and surveillance risk across groups. In such a setting, our mechanisms can detect and help reduce the gap between group average FPRs.

In practice, to obtain the group membership values required by our mechanisms, Amazon may introduce a user interface that can prompt a user for their opt-in consent to their sensitive attribute's collection with guarantees of local differential privacy. This strategy was pioneered by RAPPOR [27], and is now successfully deployed by Apple in their collection of user health and typing data [32, 61]. This approach has also been recently proposed by Meta to measure performance disparities [2]. To obtain the performance values, the clients would set aside some of their data to evaluate the global model's performance.

When the entity interested in measuring a potential disparity is a third-party, such as a regulator, the entity could collaborate with the infrastructure provider, or implement a third-party software to request the user's sensitive attribute (with local differential privacy) and collect data to evaluate the global model—note that, in cross-device FL, users always have query access to the global model. However, much like any crowdsourced and non-collaborative approach, it may have the challenge of distributing the software to a representative sample of users.

Limiting exposure. In practice, the clients would need to agree on a secure channel (e.g., TLS) with the entity performing the measurement, to limit the exposure of the protected sensitive attributes to third parties and network eavesdroppers.

Repeated measurements. The privacy budgets that we provide in our analyses are for a single measurement. However, the operator of the mechanisms would likely run at least two measurements: one to identify a potential problem and another after applying a mitigation strategy. In such cases, as in other deployed LDP mechanisms [27], recurring clients in repeated measurements would need to memoize their protected values to maintain the theoretical privacy guarantees (e.g. [27]. In addition, the privacy budget the operator would need to spend would be the privacy budget for a single measurement multiplied by the number of measurements [25].

Developing mechanisms that monitor the performance gap is beyond the scope of this work. The current situation is that performance disparities in deployed systems are unknown due to a lack of private measurement mechanisms and the most pressing question is to understand their magnitude and prevalence. Our mechanisms can safely expose these disparities and raise awareness about the performance gap, a necessary first step in mitigating it.

6.3 Mitigation Strategies

Prior works on mitigating biases in machine learning models consider different (un)fairness definitions and thus may not be effective in reducing the performance gap. Our preliminary evaluation of existing post-processing mitigation techniques [36, 42] shows that, indeed, they only lead to a modest reduction in the gap. Instead, future work should focus on developing mitigation techniques specifically designed to reduce the performance gap.

Similarly to Jagielski et al., these mitigation techniques could leverage the distributions of protected values to guide the corrective interventions at the clients. In particular, simple interventions

could have the clients tune their decision thresholds to reduce the performance gap. For some applications, the clients could adjust their performance value by penalizing other performance metrics, that may not be as critical to that specific application. Going back to the wake-word detection example, the disadvantaged group could sacrifice some true positives, to reduce the FPR, as false positives are more important than true positives for protecting privacy and against unnecessary surveillance.

Furthermore, once the performance disparity has been identified, the FL infrastructure provider is not limited to the information provided by our mechanisms to address it. For example, a concrete action upon observing a performance disparity is to collect more data for the clients in the worst-performing group (e.g., by recruiting individuals from that group to contribute their data for a financial compensation), which can alleviate the disparity when the cause of the gap is a lack of training data.

6.4 Future Work

The limitations of our approach are shared by much of the LDP literature: we make observations on the sample that participates in the measurement, not about the whole population; our utility bounds are worst-case, and thus are overly pessimistic; and our measurements are focused on one particular statistic rather than on a characterization of the performance distributions.

Extrapolating from client samples. As we explained above, when the infrastructure provider performs the measurements, they may deploy the mechanisms to all the clients and thus measure the performance gap of the population. However, in some scenarios the participating clients may be a small sample of the population and, thus, it may be necessary to extrapolate. Our mechanisms provide estimates for the performance gap of the clients that participate in the measurement but do not attempt to extrapolate to the performance gap of the client population. To extrapolate, one can aim to adapt techniques for differentially private hypothesis testing [59] to our mechanisms. Moreover, techniques of private hypothesis testing can be useful to validate the hypothesis that the group size ratio measured by the $\mathcal{M}_{\text{RR}}$ part of the mechanisms is equal to the one of a test distribution.

Sharper bounds. Even though the Chebyshev bounds provide evidence of the effectiveness of the mechanisms for measuring the performance gap in the DFL setting, our experiments show that the bounds overestimate the error of the measurements. We believe that deriving sharper bounds may be possible and such sharper bounds would allow a more precise estimation of the privacy budget. A more precise estimation of the budget is especially important in the cases when multiple measurements are needed. Future work may also improve the bounds of the mechanisms by including a semi-trusted intermediary who shuffles the data (e.g., an independent third party). Researchers have shown that by adding such an intermediary, the mechanisms can amplify the privacy guarantees provided by LDP at the same measurement accuracy level [6, 26].

Dispersion of the Performance Distribution. Our definition of fairness conveys information about the central value of the performance distributions, but is not informative about its dispersion. It can happen that we do not observe a gap in mean performance but that the performance in one group is much more dispersed than in the other, resulting in some clients experiencing a significantly worse performance than measured by the mechanisms. It is a limitation of all LDP mechanisms in the literature to focus on one descriptive statistic (e.g., the mean [13]) rather than many statistics. Recent work has begun to address this issue [28, 44]. Therefore, we are optimistic that future work will be able to develop private mechanisms to also measure dispersion (or otherwise provide a richer understanding) of the performance gap.

7 CONCLUSION

With federated learning gaining traction in industry and academia, there is a growing concern that models trained with federated learning will exhibit disparate performance across demographic groups, leading to harms ranging from a mere inconvenience to disparate impact, such as increased surveillance and lower online security for some of the groups. We propose considering the performance gap between demographic groups as a notion of (un)fairness in the DFL setting, and argue that the ability to measure it is crucial towards addressing such harms. However, especially under the privacy aspirations of federated learning, lack of demographic data hinders the applicability of existing techniques to measure performance disparities in ML models to the DFL setting. This poses an obstacle to identifying and mitigating the harms; as Roy Austin, Facebook's VP of Civil Rights, puts it: "we can't address what we can't measure" [3].

To address the legal, societal, and individual concerns related to the privacy of demographic data and fulfill the privacy aspirations of federated learning, we propose locally differentially private mechanisms that estimate the performance gap across demographic groups while protecting the privacy of the group membership and potentially correlated data such as model performance. Our theoretical and experimental results show that the mechanisms ensure strong privacy guarantees while performing relatively accurate performance gap measurements when relying on realistic numbers of clients in the DFL setting and reasonable privacy parameters. Our insight is that the large scale of existing DFL deployments offers a unique opportunity to measure and expose the potential disparities while guaranteeing strong privacy to users who choose to participate. Our mechanisms have low computational and communication costs, and are practically feasible to implement. Thus, they provide a way to reconcile the dual goals of protecting privacy while measuring the disparities, and provide a path forward for enabling aggregators and independent parties to mitigate potential performance gaps in current and future DFL deployments.

ACKNOWLEDGMENTS

This work has been supported in part by USC + Amazon Center on Secure & Trusted ML, and NSF Awards #1943584, #1916153, and #1956435.

REFERENCES

[1] Shimaa Ahmed, Ilia Shumailov, Nicolas Papernot, and Kassem Fawaz. 2021. Towards More Robust Keyword Spotting for Voice Assistants. In *Proceedings of the 31st USENIX Security Symposium*.

[2] Rachad Alao, Miranda Bogen, Jingang Miao, Ilya Mironov, and Jonathan Tannen. 2021. How Meta is working to assess fairness in relation to race in the U.S. across its products and systems. Technical Report https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems.

[3] Roy L. Austin. 2021. Race Data Measurement and Meta's Commitment to Fair and Inclusive Products. Meta blog. https://about.fb.com/news/2021/11/inclusive-products-through-race-data-measurement. Accessed: 2022-04-20.

[4] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. 2021. Evaluating Fairness of Machine Learning Models Under Uncertain and Incomplete Information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 206–214.

[5] Joan P. Bajorek. 2019. Voice Recognition Still Has Significant Race and Gender Biases. Harvard Business Review. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases. Accessed: 2022-04-20.

[6] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. 2019. The privacy blanket of the shuffle model. In *Annual International Cryptology Conference*. Springer, 638–667.

[7] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).

[8] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[9] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 492–500.

[10] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical Secure Aggregation for Privacy Preserving Machine Learning. Cryptology ePrint Archive, Report 2017/281. https://ia.cr/2017/281.

[11] Kallista Bonawitz, Peter Kairouz, Brendan McMahan, and Daniel Ramage. 2021. Federated Learning and Privacy: Building privacy-preserving systems for machine learning and data science on decentralized data. *Queue* 19, 5 (2021), 87–114.

[12] Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. 2018. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics* 112 (2018), 59–67.

[13] Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. 2021. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems* 34 (2021).

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[15] William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. 2022. Adaptive Sampling Strategies to Construct Equitable Training Datasets. *arXiv preprint arXiv:2202.01327* (2022).

[16] Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, et al. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health* 1, 3 (2022), e0000022.

[17] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2021. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*. PMLR, 1349–1361.

[18] Olivia Choudhury, Aris Gkoulalas-Divanis, Theodoros Salonidis, Issa Sylla, Yoonyoung Park, Grace Hsu, and Amar Das. 2019. Differential Privacy-enabled Federated Learning for Sensitive Health Data. *CoRR* abs/1910.02578 (2019). arXiv:1910.02578 http://arxiv.org/abs/1910.02578

[19] Kate Crawford. 2017. The Trouble with Bias. NIPS Keynote (2017) https://www.youtube.com/watch?v=fMym_BKWQzk. (2017).

[20] Arielle Elyse Czalbowski. 2015. *A cruel angel's thesis: a quantitative study of online privacy values dependent on social factors*. Ph. D. Dissertation. Iowa State University.

[21] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Multiaccurate Proxies for Downstream Fairness. *arXiv e-prints* (2021), arXiv–2107.

[22] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 429–438.

[23] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2014. Privacy aware learning. *Journal of the ACM (JACM)* 61, 6 (2014), 1–57.

[24] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.

[25] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407.

[26] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. 2019. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2468–2479.

[27] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. 1054–1067.

[28] Georgina Evans, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta. 2020. Statistically valid inferences from privacy protected data. *URL: GaryKing. org/dp* (2020).

[29] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. 2021. FairFed: Enabling Group Fairness in Federated Learning. *arXiv preprint arXiv:2110.00857* (2021).

[30] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[31] Filip Granqvist, Matt Seigel, Rogier van Dalen, Áine Cahill, Stephen Shum, and Matthias Paulik. 2020. Improving On-Device Speaker Verification Using Federated Learning with Privacy. In *Proceedings of the INTERSPEECH conference*. 4328–4332. https://doi.org/10.21437/Interspeech.2020-2944

[32] Andy Greenberg. 2016. Apple's 'Differential Privacy' Is About Collecting Your Data—But Not Your Data. https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/ Wired. [Accessed: 2022-05-2].

[33] Xiaolan Gu, Ming Li, Yueqiang Cheng, Li Xiong, and Yang Cao. 2020. {PCKV}: Locally Differentially Private Correlated Key-Value Data Collection with Optimized Utility. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 967–984.

[34] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *arXiv preprint arXiv:1806.11212* (2018).

[35] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604* (2018).

[36] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[37] Florian Hartmann, Sunah Suh, Arkadiusz Komarzewski, Tim D Smith, and Ilana Segall. 2019. Federated learning for ranking browser history suggestions. *arXiv preprint arXiv:1911.11807* (2019).

[38] Drew Harwell. 2018. The Accent Gap. The Washington Post. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/. Accessed: 2022-04-20.

[39] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[40] Hossein Hosseini, Sungrack Yun, Hyunsin Park, Christos Louizos, Joseph Soriaga, and Max Welling. 2020. Federated Learning of User Authentication Models. *arXiv preprint arXiv:2007.04618* (2020).

[41] Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *arXiv preprint arXiv:2002.10361* (2020).

[42] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 3000–3008.

[43] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[44] Vishesh Karwa and Salil Vadhan. 2017. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908* (2017).

[45] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[46] Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. 2018. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*. PMLR, 2630–2639.

[47] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *34th Conference on Neural Information Processing Systems*. Curran Associates, Inc.

[48] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. 2019. Noise-tolerant fair classification. *Advances in Neural Information Processing Systems* 32 (2019), 294–306.

[49] Brendan McMahan and Daniel Ramage. 2017. Federated Learning: Collaborative Machine Learning without Centralized Training Data. Google AI Blog. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html. Accessed: 2022-04-20.

[50] Brendan McMahan and Abhradeep Thakurta. 2022. Federated Learning with Formal Differential Privacy Guarantees. Google AI Blog. https://ai.googleblog.com/2022/02/federated-learning-with-formal.html. Accessed: 2022-04-20.
[51] Frank D McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. 19–30.
[52] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
[53] Thông T Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. 2016. Collecting and analyzing data from smart device users with local differential privacy. *arXiv preprint arXiv:1606.05053* (2016).
[54] Manisha Padala, Sankarshan Damle, and Sujit Gujar. 2021. Federated Learning Meets Fairness and Differential Privacy. *arXiv preprint arXiv:2108.09932* (2021).
[55] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231* (2018).
[56] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ digital medicine* 3, 1 (2020), 1–7.
[57] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014), 4349–4357.
[58] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.
[59] Or Sheffet. 2018. Locally private hypothesis testing. In *International Conference on Machine Learning*. PMLR, 4605–4614.
[60] Roberto Luis Shinmoto Torres, Renuka Visvanathan, Stephen Hoskins, Anton Van den Hengel, and Damith C Ranasinghe. 2016. Effectiveness of a batteryless and wireless wearable sensor system for identifying bed and chair exits in healthy older people. *Sensors* 16, 4 (2016), 546.
[61] Differential Privacy Team. 2017. Learning with Privacy at Scale. https://machinelearning.apple.com/research/learning-with-privacy-at-scale Apple Machine Learning Research Blog. [Accessed: 2022-05-2].
[62] Wiebke Toussaint and Aaron Yi Ding. 2022. Bias in Automated Speaker Recognition. In *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency*.
[63] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 2053951717743530.
[64] Shiv Vitaladevuni. 2020. Federated Learning Applications in Alexa. http://federated-learning.org/fl-icml-2020/#k3 Keynote session by Shiv Vitaladevuni (Amazon Research) at the ICML'20 workshop "Federated Learning for User Privacy and Data Confidentiality" . [Accessed: 2022-04-20].
[65] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally Differentially Private Protocols for Frequency Estimation. In *26th USENIX Security Symposium (USENIX Security 17)*. USENIX Association, Vancouver, BC, 729–745. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao
[66] Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 309 (1965), 63–69.
[67] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. 2021. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research* 5, 1 (2021), 1–19.
[68] Xiaohang Xu, Hao Peng, Md Zakirul Alam Bhuiyan, Zhifeng Hao, Lianzhong Liu, Lichao Sun, and Lifang He. 2021. Privacy-Preserving Federated Depression Detection from Multi-Source Mobile Health Data. *IEEE Transactions on Industrial Informatics* (2021).
[69] Xiaohang Xu, Hao Peng, Lichao Sun, Md Zakirul Alam Bhuiyan, Lianzhong Liu, and Lifang He. 2021. FedMood: Federated Learning on Mobile Health Data for Mood Detection. *arXiv preprint arXiv:2102.09342* (2021).
[70] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903* (2018).
[71] Qingqing Ye, Haibo Hu, Xiaofeng Meng, and Huadi Zheng. 2019. PrivKV: Key-value data collection with local differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 317–331.
[72] Binhang Yuan, Song Ge, and Wenhui Xing. 2020. A federated learning framework for healthcare iot devices. *arXiv preprint arXiv:2005.05083* (2020).
[73] Xubo Yue, Maher Nouiehed, and Raed Al Kontar. 2021. GIFAIR-FL: An Approach for Group and Individual Fairness in Federated Learning. *arXiv preprint arXiv:2108.02741* (2021).
[74] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.
[75] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

A EMPIRICAL VALIDATION

We have run experiments to validate the correctness of our expressions of the variance of the estimators. In the experiments, we initialize two groups with 10 clients each with fixed performance values. Then, we run the mechanisms a number of times to obtain sets of perturbed tuples and calculate the performance gap estimates. The empirical MSE is the average of the squared differences between these estimates and the true performance gap. We plot the empirical and theoretical MSE for mechanism $\mathcal{M}_{\mathrm{R}}$ in Fig. 4. We observe that, as we increase the number of runs, the empirical MSE converges to the theoretical MSE, validating our results.

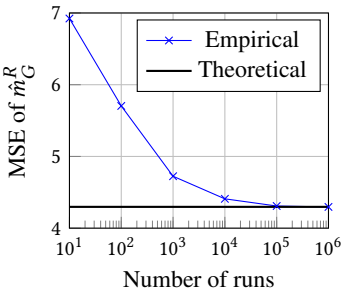

Figure 4: The theoretical upper bound of the MSE of $\hat{m}_G^R$ as derived from Theorem 5.3, and its empirical MSE over different runs of $\mathcal{M}_{\mathrm{R}}$, for $n_G = n_{\bar{G}} = 10$.