Proceedings of Machine Learning Research 1:1–19, 2022 Algorithmic Fairness through the Lens of Causality and Privacy

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

Marc Juarez* MARC.JUAREZ@ED.AC.UK
University of Edinburgh

Aleksandra Korolova* KOROLOVA@PRINCETON.EDU
Princeton University

Abstract

As in traditional machine learning models, models trained with federated learning may exhibit disparate performance across demographic groups. Model holders must identify these disparities to mitigate undue harm to the groups. However, measuring a model's performance in a group requires access to information about group membership which, for privacy reasons, often has limited availability. We propose novel locally differentially private mechanisms to measure differences in performance across groups while protecting the privacy of group membership. To analyze the effectiveness of the mechanisms, we bound their error in estimating a disparity when optimized for a given privacy budget. Our results show that the error rapidly decreases for realistic numbers of participating clients, demonstrating that, contrary to what prior work suggested, protecting privacy is not necessarily in conflict with identifying performance disparities of federated models.

Keywords: differential privacy, algorithmic fairness, federated learning

1. Introduction

Cross-device federated learning (DFL) has become a popular way to distribute the training of machine learning (ML) models across multiple devices. Currently, there are several large-scale deployments of DFL in the industry, such as Android GBoard's next-word prediction (Hard et al., 2018; Yang et al., 2018), and Siri's speaker identification (Granqvist et al., 2020). A key motivation for these deployments is the aspiration of training powerful models while ensuring privacy and data minimization.

In parallel, ML models have been shown to exhibit disparate performance across groups, often falling short for people from marginalized groups, in the domains of vision, natural language processing, and healthcare (Buolamwini and Gebru, 2018; Sap et al., 2019; Celi et al., 2022; Mehrabi et al., 2021); and, more recently, these performance disparities have also been observed in DFL (Yuan et al., 2020; Xu et al., 2021a,b).

Performance disparities may be harmful beyond merely the individual's experience of worse quality of service (Crawford, 2017). A greater false positive rate of Alexa's wake-word detection on a group may lead to over-surveillance of that group, as a false activation may record unrelated speech and send it to the cloud for further processing (Vitaladevuni, 2020). In DFL applications to the security domain (Hosseini et al., 2020), a performance disparity may lead to a lower security level for certain groups. Overall, DFL has enormous traction in

* Most of their work was done while at the University of Southern California.

© 2022 M. Juarez & A. Korolova.

industry and academia and, if it were to become a de-facto data minimization or privacy standard for much of ML, as current trends suggest, an unknown performance disparity dependent on a demographic group could have tremendous negative consequences in many application domains.

The challenge in detecting and mitigating disparate performance of a DFL model is that access to information about the attributes related to group membership is often limited or noisy (Veale and Binns, 2017; Bogen et al., 2020). Regulations, such as the GDPR in the EU, mandate that protected attributes, such as gender or race, be collected only under appropriate privacy protections and explicit informed consent. In addition, without adequate privacy protection, volunteers who are members of stigmatized groups are more likely to provide a false group membership, which can add noise to the collected attributes.

We reconcile the seemingly incompatible goals of ensuring group membership privacy and mitigating performance disparity via local differential privacy (LDP). We propose novel LDP mechanisms that allow us to measure performance disparities while protecting the privacy of the attributes that define group membership. To compare the mechanisms over a range of privacy levels, number of clients, and group sizes, we characterize the measurement error they induce as a function of the privacy budget, and find budget allocations that minimize the error under a privacy constraint.

Our theoretical analysis shows that the mechanisms ensure strong privacy guarantees while the measurement error is relatively low for typical numbers of clients in a DFL setting. With our tools, the aggregator of the models, or even a regulatory agency could identify cases of performance disparity in applications where such disparity is undesirable or harmful for some of the groups.

2. Background

Differential Privacy (DP) We argue that the local model of DP is better suited for cross-device federated learning (DFL), as it is unclear who would play the role of a central curator in the existing deployments of DFL and the large-scale of these deployments can attenuate the privacy-induced error of an LDP mechanism.

Definition 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). For a binary protected attribute, RR returns the true value with probability a and returns the opposite value otherwise. Generalized RR (GRR) extends RR to a non-binary protected attribute by uniformly distributing the probability of giving a different value (Wang et al., 2017).

Definition 2 (The GRR mechanism) *For $x \in \{0, \ldots, d\}$, $d \geq 1$, and $a \in [\frac{1}{2}, 1]$, the GRR mechanism, $\mathcal{M}_{\text{GRR}}$, is defined by*

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

Federated Learning (FL) FL allows many clients, each with its own dataset, to train a model on the union of all datasets, without any dataset having to leave its client's device. The training is distributed over the clients, who train local models on their datasets. The clients then share the local models' parameters with a central *aggregator*, who averages them to obtain the parameters of the *global model*.

There are different types of FL depending on who the clients are. We focus on cross-device FL (DFL), where clients run on different devices (*e.g.*, smartphones) and the training data usually belongs to the same user. We focus on DFL as it is becoming increasingly popular in the Big Tech industry (Yang et al., 2018; Granqvist et al., 2020).

3. Problem Statement

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 the disparate performance across groups of clients with the groups defined by a demographic attribute, such as sex or race.

Formally, an attribute is a set $\mathcal{A} = \{0, \ldots, d\}$, with $d \geq 1$, that induces a partition of the clients, $\mathcal{P} = \{G_0, \ldots, G_d\}$. We consider K clients and denote (g_k, v_k) the values of the attribute ($g_k \in \mathcal{A}$) and the performance of the model for client k. The client obtains v_k by evaluating the global model on a fraction of their data. The choice of an appropriate performance metric depends on various factors, such as the learning task and the potential harms in a particular application.

Definition 3 (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 any two groups, we measure the absolute difference between the mean performances of the global model on the groups.

Definition 4 (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 (un)fairness is in contrast with traditional fairness definitions that measure model performance on individual predictions. Previous definitions are suitable for scenarios where data points represent people and each single prediction concerns an individual (*e.g.*, credit score prediction). However, these definitions are not suited for the notion of (un)fairness that we consider. In the typical DFL setting, the global model is distributed to the clients for use on their *own* data and they are not necessarily the subjects of the predictions. Our concern is that the disparate performance of the global model across groups of users can lead to a disparate impact; the performance gap captures this disparity across groups.

Adversary model. We assume that the entity performing the measurements of the performance gap is the FL aggregator, 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 (McMahan and Ramage, 2017), we assume that the aggregator uses Secure Aggregation (Bonawitz et al., 2017). Therefore, the aggregator cannot infer any information from the FL updates about the users' protected attributes.

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 must apply an LDP mechanism, $\mathcal{M}$, on their group-value tuples before sending them to the aggregator. We denote the perturbed tuples $(g'_k, v'_k) := \mathcal{M}(g_k, v_k)$. From a privacy perspective, we assume that the aggregator follows the protocol as intended but may try to learn g_k or v_k from (g'_k, v'_k).

In this work, 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 tradeoffs they impose in terms of their privacy guarantees and the error they induce in a measurement. *We hypothesize that the number of clients of current DFL deployments is sufficient to allow for low-error and high-privacy measurements with our mechanisms.*

4. Measuring the Performance Gap

Since the performance gap is the absolute difference of two group mean performances, we first tackle the problem of private *group* mean estimation and then show that the privacy guarantees hold when combining them into a performance gap estimation.

A major distinction between *group* mean estimation and *population* mean estimation in the literature (Brown et al., 2021; Asi et al., 2022) is that, in estimating a group mean performance, the performance may be correlated with the group—especially, if there is a gap. In that case, the adversary would learn group information from observing the performance.

A successful mechanism must protect both group and performance values. A naïve approach is to protect both independently, but that would destroy the necessary information to measure the gap. Thus, our mechanisms are designed to preserve the *overall* aggregate correlation between the group and the performance values, while preventing inference of the group that an individual client belongs to from the perturbed tuples.

All our mechanisms use $\mathcal{M}_{\text{GRR}}$ 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, as they can always claim that the mechanism assigned them to a different group.

We present two mechanisms that differ by how they perturb the performance values:

The $\mathcal{M}_{\text{R}}$ mechanism. After perturbing the group (line 1 in Algorithm 1), $\mathcal{M}_{\text{R}}$ discretizes the value with Harmony's discretization (Nguyên et al., 2016) (lines 3–5) and then applies GRR on the discretized value (line 6). The Harmony discretization allows for unbiased estimates of the expected value from the discretized values. The mechanism sets to zero the performance value of the clients who flip their group (line 2). This is to ensure that they do not contribute to the other group's mean performance value (Gu et al., 2020).

Algorithm 1: Pseudocode of the privacy mechanism: $\mathcal{M}_{\mathrm{R}}$.

Input: The client's group $g \in \{0, 1\}$ and performance value $v \in [-1, 1]$. The privacy budgets $\epsilon_1, \epsilon_2 \in [0, +\infty)$.

Output: The perturbed tuple (g', v'), $g' \in \{0, 1\}$ and $v' \in \{-1, 1\}$.

$g' \leftarrow \mathcal{M}_{\mathrm{GRR}}(g; d, \frac{e^{\epsilon_1}}{e^{\epsilon_1}+d-1})$ `// Perturb the group.`

if $g \neq g'$ **then**

  $v \leftarrow 0$ `// The performance distribution of the other group is unknown.`

end

Draw $B \sim \text{Bernoulli}(\frac{1+v}{2})$

$v' \leftarrow 2 * \mathcal{M}_{\mathrm{GRR}}(B; 2, \frac{e^{\epsilon_2}}{1+e^{\epsilon_2}}) - 1$ `// Perturb the value and transform.`

return (g', v')

The $\mathcal{M}_{\mathrm{L}}$ mechanism. $\mathcal{M}_{\mathrm{L}}$ is identical to $\mathcal{M}_{\mathrm{R}}$ except that instead of applying GRR to perturb the values, it uses the Laplace mechanism. 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. Thus, $\mathcal{M}_{\mathrm{L}}$ exposes a parameter k, $0 < k \leq 2$, that allows to fine-tune the scale of the Laplace distribution of the noise for the clients whose group was perturbed. In addition, the value of the clients that switch to another group is set to zero, such that, like in $\mathcal{M}_{\mathrm{R}}$, they do not contribute to that group's mean value.

Because the mechanisms are the composition of the GRR and the Laplace mechanisms, they have two privacy budget parameters: ϵ_1 and ϵ_2, the privacy budget to protect the group and the performance values, respectively.

One of our main results is that the mechanisms achieve ϵ-LDP for an overall privacy budget ϵ.

Theorem 5 *$\mathcal{M}_{\mathrm{R}}$ is ϵ-LDP with $\epsilon = \max\{\epsilon_1, \epsilon_2\}$.*

Proof See Appendix A.1. ■

Theorem 6 *The mechanism $\mathcal{M}_{\mathrm{L}}$ is ϵ-LDP with*

$$\epsilon = \max\left\{\epsilon_2,\ \ln(\frac{2}{k}) + \frac{\epsilon_2}{2} - \epsilon_1,\ \ln(\frac{k}{2}) + \frac{\epsilon_2}{k} + \epsilon_1\right\} \tag{1}$$

Proof See Appendix A.2. ■

These bounds are tighter than the ones obtained with the basic theorem on sequential composition of DP mechanisms (McSherry, 2009). 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 tradeoff.

5. Performance Evaluation

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. 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.

The estimators that the operator should use to estimate m_G from the perturbed group-performance tuples are as follows.

Definition 7 (Estimators of m_G) *We define the following estimators for the mechanisms:*

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

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

We have proven that the estimators are unbiased, and have obtained closed-form expressions of their variance (see Appendix A.4), hence their MSE.

Proposition 8 *The estimators of the mechanisms are unbiased: $\mathbb{E}[\hat{m}_G^L] = \mathbb{E}[\hat{m}_G^R] = m_G$.*

Proof See Appendix A.3 ■

Theorem 9 shows that the MSE of $\Delta\tilde{m}^*$ is the sum of the MSEs of the group mean value estimates.

Theorem 9 *$\Delta\tilde{m}^*$ is an unbiased estimator of Δm for two groups $G, \bar{G} \in \mathcal{P}$ with MSE:*

$$MSE\,[\Delta\tilde{m}^*] = MSE[\tilde{m}_G^*] + MSE[\tilde{m}_{\bar{G}}^*].$$

Proof See Appendix A.5. ■

The intuition behind Theorem 9 is that even though $\tilde{m}_G^*$ and $\tilde{m}_{\bar{G}}^*$ are not independent, the errors are uncorrelated, and thus they add up. Therefore, we can obtain a closed-form expression of the MSE of Δm by adding the closed-form expressions of the group MSEs.

MSE for a fixed privacy budget As shown by Theorem 5 and Theorem 6, the privacy budgets ϵ_1 and ϵ_2 have a different impact on the LDP bound of the mechanism. To draw a fair comparison between the mechanisms, we need to find the parameters that minimize the error on utility for a fixed overall privacy budget ϵ. With the closed-form expression of the MSE of $\Delta\tilde{m}^*$, we can compare the estimators for specific values of ϵ_1 and ϵ_2. It is unclear a priori how to divide a fixed privacy budget into the mechanism's group and performance components to maximize utility. Our approach is to find an allocation that minimizes the MSE of the estimators, for the total privacy budget of the mechanism (ϵ).

For $\mathcal{M}_{\mathrm{R}}$, the optimal allocation is $\epsilon_1 = \epsilon_2 = \epsilon$; for $\mathcal{M}_{\mathrm{L}}$ with $k = 2$, it is $\epsilon_2 = \epsilon$, and $\epsilon_1 = \frac{\epsilon}{2}$. In Appendix B, we find the optimal allocation for $\mathcal{M}_{\mathrm{L}}$ with k as a parameter.

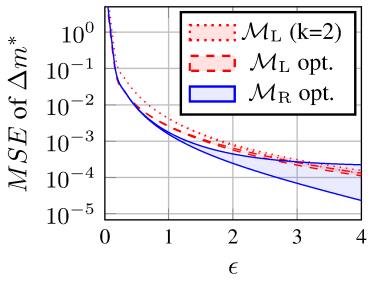

Figure 1: Upper and lower bounds of estimator MSE for different privacy budgets ϵ. We have set $n_G = n_{\bar{G}} = 10^4$.

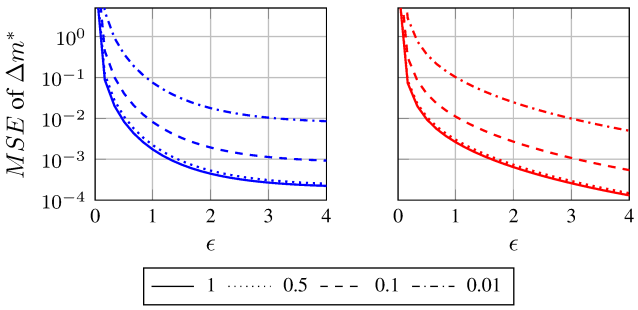

Figure 2: Upper bound of the MSE of $\mathcal{M}_{\mathrm{R}}$ opt. (left) and $\mathcal{M}_{\mathrm{L}}$ with $k = 2$ (right) for $K = 2 \cdot 10^4$ and different group ratios $n_G/n_{\bar{G}}$.

Comparison of the mechanisms. In Fig. 1, we plot the MSE of the performance gap estimator for two groups G and $\bar{G}$ of the same size. $\mathcal{M}_{\mathrm{L}}$ opt. and $\mathcal{M}_{\mathrm{R}}$ opt. are the mechanisms with optimal parameters. Since the specific set of v_k's has an impact on the MSE, in the graph we show the lower and upper bounds for each mechanism, which enclose a (colored) region of MSEs.

We observe that $\mathcal{M}_{\mathrm{L}}$ opt. achieves lower MSE than $\mathcal{M}_{\mathrm{L}}$ with $k = 2$ for the range of ϵ that we consider. $\mathcal{M}_{\mathrm{R}}$ opt.'s MSE is lower than $\mathcal{M}_{\mathrm{L}}$ opt. only when $0.31 \lessapprox \epsilon \lessapprox 2.6$ but, for $\epsilon \gtrapprox 2.6$, the upper bound of $\mathcal{M}_{\mathrm{R}}$ opt.'s MSE is larger than the upper bound of $\mathcal{M}_{\mathrm{L}}$ opt.'s and the gap between the two rapidly widens for larger privacy budgets. As a consequence, there is no overall best mechanism and the operator should select the one that best suits their budget.

Unbalanced Groups We also study the impact that an unbalance of group sizes has on the MSE of the estimators. Fig. 2 depicts the upper bounds of the MSEs, $\mathcal{M}_{\mathrm{R}}$ opt. (left) and $\mathcal{M}_{\mathrm{L}}$ with $k = 2$ (right), for two groups with different size ratios. We observe that the mechanisms can cope with relatively small groups, but the MSE rapidly grows for minorities that account for less than 1% of the clients. Consequently, the mechanisms would incur in high MSE with protected attributes that include small minorities, but would maintain low MSEs for protected attributes like gender and race.

We have implicitly assumed that the entity performing the measurements knows n. Fig. 2 also quantifies the difference of the mechanisms' MSE when the operator incorrectly estimates the group fractions. 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 to a large extent.

Precision Relative to the Number of Clients Our mechanisms allow for high-precision measurements with realistic numbers of clients in the DFL setting. Using Theorem 9, we obtain a Chebyshev concentration bound and numerically find the privacy budget ϵ such that $|\Delta\tilde{m}^*_G - \Delta m| < \alpha$, for a small $\alpha > 0$, with high probability.

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.

	$\mathcal{M}_{\mathbf{R}}$ opt.			$\mathcal{M}_{\mathbf{L}}$ opt.		
K	$\alpha = 10^{-1}$	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$	$\alpha = 10^{-1}$	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$
10^5	1.86	-	-	2.56	17.89	178.89
10^6	0.63	-	-	0.71	6.32	56.57
10^7	0.23	1.86	-	0.21	2.56	17.89
10^8	0.08	0.63	-	0.07	0.71	6.32
10^9	0.02	0.23	1.86	0.02	0.21	2.56

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.

Due to $\mathcal{M}_{\mathrm{R}}$'s discretization step, the maximum estimator variance (attained when all the performance values are zero) tends to a constant (inversely proportional to group size). Thus, there is no budget that allows to achieve α for those cells marked with "-".

These results show that the required budgets to achieve an error of less than one percentage point and less of one tenth of a percentage point are reasonable for $K \geq 10^7$ and $K \geq 10^9$ clients, respectively. Even though these may look like large numbers of clients, current DFL deployments have this many, and even more, clients. For example, in 2018 Apple reported a total of half a billion active Siri clients (Apple press team, 2018) and, in the same year, Gboard surpassed 1 billion installs (Google Play Store, 2018).

We have also evaluated these bounds on a real-world dataset of performances of a simulated FL deployment (Appendix D). Our results show that the bounds not only hold but that they are overly conservative: in practice, the operators of the mechanisms would be able to ensure the same privacy level by spending less privacy budget. The Chebyshev bounds are known to be loose because they do not make assumptions about the underlying distributions; we leave finding tighter bounds for future work.

6. Related Work

In the ML literature, Veale and Binns (2017) first noted the legal, institutional, and commercial deterrents against collecting demographic data. To address the lack of demographic data, they envisioned privacy-preserving protocols that rely on a third-party to detect and mitigate discrimination.

Researchers materialized these protocols with DP and Secure Multi-Party Computation (SMPC) (Jagielski et al., 2019; Kilbertus et al., 2018; Alao et al., 2021). Jagielski et al. (2019) proposed DP versions of existing post- and in-processing techniques to train classifiers that satisfy the Equalized Odds constraint. In contrast, our work defines the notion of performance gap as it is more suitable for the DFL setting. In addition, as most of related

work, their focus is on mitigating the disparity assuming that it has occurred, while we focus on the measurement of the disparity rather than its mitigation.

In contrast to SMPC, our mechanisms have low computational and bandwidth costs, and are robust to client dropouts. Moreover, SMPC and DP provide different privacy guarantees; in particular, SMPC does not limit the information about individual group membership that the aggregator can infer from the measurements.

Another difference between our work and the prior works is the involvement of the model holder towards the goal of identifying disparities. 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.

Prior work on LDP mechanisms to protect sensitive attributes is too extensive to be covered in detail. Recent work has made progress on designing mechanisms for private mean estimation on the theoretical (Nguyên et al., 2016; Asi et al., 2022) and practical fronts (Gu et al., 2020; Ye et al., 2019). However, this literature does not consider the perturbation of performance values for a performance gap measurement, and therefore, is focused on slightly different privacy vs. utility tradeoffs than we are.

7. Discussion and Conclusion

With FL gaining traction in industry and academia, there is a growing concern that models trained with it 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 DFL models. This poses an obstacle to mitigating the harms; as Roy Austin (2021), Facebook's VP of Civil Rights, puts it: "we can't address what we can't measure."

To address the legal, societal, and individual concerns related to the privacy of demographic data, we propose locally differentially private mechanisms that estimate the performance gap 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 precise 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 the participants.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. We are also grateful to Shuang Wu, Nedelina Teneva, and Basileal Imana for their valuable input during the development of this work. This work has been supported in part by USC + Amazon Center on Secure & Trusted ML, and NSF Awards #1943584, #1916153, and #1956435.

References

Rachad Alao, Miranda Bogen, Jingang Miao, Ilya Mironov, and Jonathan Tannen. 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, 2021.

Apple press team. Homepod arrives february 9, available to order this friday, 1 2018. URL https://www.apple.com/newsroom/2018/01/homepod-arrives-february-9-available-to-order-this-friday/. Apple Newsroom. [Accessed: 2022-05-5].

Hilal Asi, Vitaly Feldman, and Kunal Talwar. Optimal algorithms for mean estimation under local differential privacy, 2022.

Roy L. Austin. 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, 2021. Accessed: 2022-04-20.

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

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy preserving machine learning. Cryptology ePrint Archive, Report 2017/281, 2017. https://ia.cr/2017/281.

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

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

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. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1(3):e0000022, 2022.

Kate Crawford. The trouble with bias. *NIPS keynote*, 2017.

Google Play Store. Gboard – the Google Keyboard. https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin, 2018. Accessed: 2022-04-05.

Filip Granqvist, Matt Seigel, Rogier van Dalen, Áine Cahill, Stephen Shum, and Matthias Paulik. Improving On-Device Speaker Verification Using Federated Learning with Privacy.

In *Proceedings of the INTERSPEECH conference*, pages 4328–4332, 2020. doi: 10.21437/Interspeech.2020-2944.

Xiaolan Gu, Ming Li, Yueqiang Cheng, Li Xiong, and Yang Cao. {PCKV}: Locally differentially private correlated key-value data collection with optimized utility. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 967–984, 2020.

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

Hossein Hosseini, Sungrack Yun, Hyunsin Park, Christos Louizos, Joseph Soriaga, and Max Welling. Federated learning of user authentication models. *arXiv preprint arXiv:2007.04618*, 2020.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3000–3008. PMLR, 09–15 Jun 2019.

Marc Juarez and Aleksandra Korolova. ldp-measurements-fl. https://github.com/mjuarezm/ldp-measurements-fl, 2022.

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

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

Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6):1–35, 2021.

Thông T Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. Collecting and analyzing data from smart device users with local differential privacy. *arXiv preprint arXiv:1606.05053*, 2016.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019.

Roberto Luis Shinmoto Torres, Renuka Visvanathan, Stephen Hoskins, Anton Van den Hengel, and Damith C Ranasinghe. Effectiveness of a batteryless and wireless wearable sensor system for identifying bed and chair exits in healthy older people. *Sensors*, 16(4): 546, 2016.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

Shiv Vitaladevuni. Federated learning applications in alexa, 7 2020. URL 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].

Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745, Vancouver, BC, August 2017. USENIX Association. ISBN 978-1-931971-40-9. URL https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao.

Xiaohang Xu, Hao Peng, Md Zakirul Alam Bhuiyan, Zhifeng Hao, Lianzhong Liu, Lichao Sun, and Lifang He. Privacy-preserving federated depression detection from multi-source mobile health data. *IEEE Transactions on Industrial Informatics*, 2021a.

Xiaohang Xu, Hao Peng, Lichao Sun, Md Zakirul Alam Bhuiyan, Lianzhong Liu, and Lifang He. Fedmood: Federated learning on mobile health data for mood detection. *arXiv preprint arXiv:2102.09342*, 2021b.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Qingqing Ye, Haibo Hu, Xiaofeng Meng, and Huadi Zheng. Privkv: Key-value data collection with local differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 317–331. IEEE, 2019.

Binhang Yuan, Song Ge, and Wenhui Xing. A federated learning framework for healthcare iot devices. *arXiv preprint arXiv:2005.05083*, 2020.

Appendix A. Proofs

A.1. Proof of Theorem 5

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 $x = (g,\ v)$,

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

We prove it for $d = 2$ as that is what we use in most of our evaluation, and leave the case $d > 2$ for future work.

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

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

and a lower bound is

$$\Pr[y \mid x] \geq a(1-b) \tag{3}$$

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 may (or may not) differ in value, we consider two cases: $g' = g$ and $g' \neq g$ (where $g = g_0 = g_1$).

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

$$\frac{\Pr[y \mid x_0]}{\Pr[y \mid x_1]} \leq \frac{ab}{a(1-b)} = e^{\epsilon_2} \tag{4}$$

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

$$\frac{\Pr[y \mid x_0]}{\Pr[y \mid x_1]} = 1 \leq e^{\epsilon_2},\ \text{as } \epsilon_2 \in [0, +\infty) \tag{5}$$

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, we again break down the analysis into two cases: $g' = g_0 \neq g_1$ and $g' = g_1 \neq g_0$.

Case 1: $g' = g_0 \neq g_1$**.** Using the upper bound and taking $e_2 = 0$ as the minimum value for the denominator, we obtain:

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

Case 2: $g' = g_1 \neq g_0$**.** Using the lower bound and that $1 \leq e^{\epsilon_2}$, we have:

$$\frac{\Pr[y \mid x_0]}{\Pr[y \mid x_1]} \leq \frac{1-a}{2a(1-b)} = \frac{1+e^{\epsilon_2}}{2e^{\epsilon_1}} \leq \frac{2e^{\epsilon_2}}{2e^{\epsilon_1}} = e^{\epsilon_2-\epsilon_1} \tag{7}$$

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

A.2. Proof of Theorem 6

Proof This proof is for $k = 2$. 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. Because $\mathcal{M}_{\text{L}}$ perturbs the values with Laplacian noise, we have that for an arbitrary input $x = (g,\ v)$,

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

This is because when the mechanism preserves the group, $v' = v+Y$ where $Y \sim \text{Lap}(0, \frac{2}{\epsilon_2})$, 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 therefore in that case it is the probability of sampling v' from $\text{Lap}(0, \frac{2}{\epsilon_2})$.

As in the proof of Theorem 5, 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: $g' = g$**.**

$$\frac{\Pr[y \mid x_0]}{\Pr[y \mid x_1]} = \frac{f_{\text{Lap}(0,\frac{2}{\epsilon_2})}(v'-v_0)}{f_{\text{Lap}(0,\frac{2}{\epsilon_2})}(v'-v_1)} = e^{\epsilon_2(\frac{|v'-v_1|}{2}-\frac{|v'-v_0|}{2})} \leq e^{\epsilon_2} \tag{8}$$

Case 2: $g' \neq g$**.**

$$\frac{\Pr[y \mid x_0]}{\Pr[y \mid x_1]} = \frac{f_{\text{Lap}(0,\frac{2}{\epsilon_2})}(v')}{f_{\text{Lap}(0,\frac{2}{\epsilon_2})}(v')} = 1 \tag{9}$$

If x_0 and x_1 differ in group, we again consider two cases: $g' = g_0 \neq g_1$ and $g' = g_1 \neq g_0$.

Case 1: $g' = g_0 \neq g_1$**.**

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

The last inequality follows from $\frac{|v'|}{2} - \frac{|v'-v_0|}{2} \leq \frac{1}{2}$.

Case 2: $g' = g_1 \neq g_0$**.**

$$\frac{\Pr[y \mid x_0]}{\Pr[y \mid x_1]} = e^{\epsilon_2(\frac{|v'-v_1|}{2}-\frac{|v'|}{2})-\epsilon_1} \leq e^{\frac{\epsilon_2}{2}-\epsilon_1} \tag{10}$$

The last inequality follows from the triangle inequality: $\frac{|v'-v_1|}{2} - \frac{|v'|}{2} \leq \frac{|v_1|}{2} \leq \frac{1}{2}$.

Finally, combining all the inequalities above, we obtain the ϵ in the bound of the probability ratio

$$\epsilon = \max\left\{\epsilon_2,\ \frac{\epsilon_2}{2} - \epsilon_1,\ \frac{\epsilon_2}{2} + \epsilon_1\right\} = \max\left\{\epsilon_2, \frac{\epsilon_2}{2} + \epsilon_1\right\}$$

■

Thus, the optimal budget allocation for mechanism $\mathcal{M}_{\text{L}}$ with $k = 2$ is $\epsilon_2 = \epsilon$ and $\epsilon_1 = \frac{\epsilon}{2}$.

A.3. Proof of Theorem 8

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

$$V_i = B_i(v_i + Y_i), \quad i = 1, \ldots, \ n, \tag{11}$$

$$\bar{V}_j = \bar{B}_j(0 + \bar{Y}_j) = \bar{B}_j\bar{Y}_j, \quad j = 1, \ldots, \ K - n \tag{12}$$

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

$$\mathbb{E}[\hat{m}_G^L] = \frac{1}{a\,n}\left(\sum_{i=1}^{n}\mathbb{E}\left[V_i\right] + \sum_{j=1}^{K-n}\mathbb{E}[\bar{V}_j]\right) \qquad \text{Linearity of } \mathbb{E} \tag{13}$$

$$= \frac{1}{a\,n}\sum_{i=1}^{n}\mathbb{E}[B_i(v_i + Y_i)] \qquad \mathbb{E}[\bar{V}_j] = 0 \tag{14}$$

$$= \frac{1}{a\,n}\sum_{i=1}^{n}\mathbb{E}[B_i](v_i + \mathbb{E}[Y_i]) \qquad \text{Mutual independence} \tag{15}$$

$$= \frac{1}{a\,n}\sum_{i=1}^{n}\mathbb{E}[B_i]v_i \qquad \mathbb{E}[Y_i] = 0 \tag{16}$$

$$= \frac{a}{a\,n}\sum_{i=1}^{n}v_i \qquad \mathbb{E}[B_i] = a \tag{17}$$

$$= m_G \tag{18}$$

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

A.4. Closed-form expressions of Variance

Using the probabilistic model defined in Appendix A.3, we can write the variance of the estimator $\hat{m}_G^L$ as

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

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 closed expression of the variance:

$$\mathrm{Var}[\hat{m}_G^L] = \frac{1}{n}\left(\nu^2 e^{-\epsilon_1} + \left(1 + e^{-\epsilon_1}\right)\left(\sigma_L^2 + \frac{K-n}{n}\bar{\sigma}_L^2 e^{-\epsilon_1}\right)\right) \tag{19}$$

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 (functions of ϵ_2), for clients who do not swap and those who do, respectively. The lower and upper bounds shown in Fig. 1 are taken using that $0 \leq \nu^2 \leq 1$.

The closed-form expression of $\hat{m}_G^R$'s variance can be obtained similarly, and is

$$\mathrm{Var}[\hat{m}_G^R] = \frac{1}{a(2b-1)^2 n}\left(1 - a(2b-1)^2\nu^2 + \frac{K-n}{n}\frac{1-a}{a}\right) \tag{20}$$

Recall that a and b are functions of the privacy budgets.

A.5. Proof outline for Theorem 9

First, we prove the unbiasedness of $\Delta\hat{m}^*$. Due to Theorem 8 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 $\mathcal{M}_{\mathrm{R}}$ is similar, as the expected value of clients with the group perturbed is zero.

Appendix B. Allocating the privacy budget for the $\mathcal{M}_{\mathrm{L}}$ mechanism

In Eq. (19), 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. (1), 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. (1).

If we take $\epsilon_2 = \epsilon$ in Eq. (1) of Theorem 6, we obtain bounds for ϵ_1

$$\ln(\frac{2}{k}) - \frac{\epsilon}{2} \leq \epsilon_1 \leq \ln(\frac{2}{k}) + \frac{\epsilon}{2}\,\lambda(k), \tag{21}$$

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. (21), a function of k, and find that $k = \epsilon$ is a maximum for a constant ϵ. However, for $0 < k < 2/3$, Eq. (21) 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. (21).

Thus, the maximum ϵ_1 as a function of ϵ is

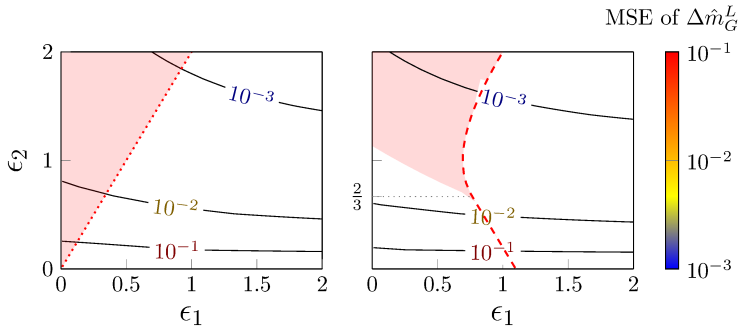

Figure 3: Contour plot of the MSE of $\Delta\hat{m}^L$ for $k = 2$ (left) and $k = \frac{2}{3}$ (right), as a function of ϵ_1 and ϵ_2. The colored area is the region where the parameters satisfy ϵ-LDP for $\epsilon = \epsilon_2$. The curves represent the optimal allocations when $k = 2$ (dotted) and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

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

Fig. 3 shows the allocations of the privacy budgets that satisfy the LDP constraint (colored area). The dashed and dotted borders of the areas show the allocations that minimize the MSE for a total privacy budget of $\epsilon = \epsilon_2 \in (0, 2]$ for $k = 2$ and k=2/3, respectively. A closer look at the MSE contour lines reveals that the mechanism with $k = 2/3$ achieves lower MSE values than for $k = 2$ when $\epsilon < 2/3$.

Appendix C. 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.

The source code for reproducing these experiments is publicly available (Juarez and Korolova, 2022).

Appendix D. Empirical Evaluation

We now describe the experiments to 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

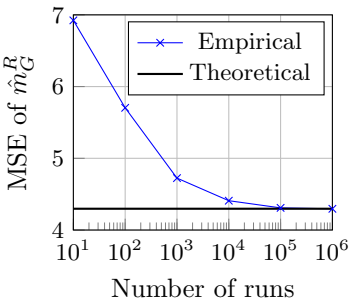

Figure 4: The theoretical upper bound of the MSE of $\hat{m}_G^R$ as derived from Theorem 9, and its empirical MSE over different runs of $\mathcal{M}_{\mathrm{R}}$, for $n_G = n_{\bar{G}} = 10$.

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 privacy risk assessment.

Data Generation Our data generation model is based on the activity detection dataset collected by Shinmoto Torres et al. (2016). The dataset comprises the sensor readings for 14 subjects who were instructed to perform a number of scripted daily activities in two different rooms. The features include the sensor's readings of time, accelerometer position, and radio signal's strength, frequency, and phase. The labels describe one of these activities: 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 a different FL client, with each client having an average of 864 samples. We stratify the data ensuring that all clients have the same data distribution between training and test sets (70% of the samples for training and 30% for testing).

We simulated the federated learning of a model by training a logistic regression model. 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 hold-out average test accuracy for 10 runs is 84.37%, with a false positive rate (FPR) of 10.69%, and a true positive rate (TPR) of 82.05% (all SD values are smaller than 1%). Then, we independently test the global model on each client's test set, resulting in two performance values for each client. We take the TPR and the FPR as performance metrics: the mean TPRs are 89.01% and 71.77% and the mean FPRs are 15.26% and 24.90% for males and females, respectively. We observe a significant performance gap on both metrics: $\Delta\mathrm{TPR} = 17.33\%$ and $\Delta\mathrm{FPR} = 9.63\%$.

Regression model implementation We implemented the evaluation of the logistic regression model with Python 3.7.6 and sklearn 0.22.1.

Table 2: Comparison of the Chebyshev bounds with the empirical mean error for 10 runs of the mechanisms on the synthetic dataset with $K = 10^7$ clients. 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 ($\boldsymbol{\alpha}$) for each mechanism.

	$\mathcal{M}_{\mathbf{R}}$ **opt.**		$\mathcal{M}_{\mathbf{L}}$ **opt.**	
$\boldsymbol{\epsilon}$	$\lvert\boldsymbol{\Delta\hat{m}^R - \Delta m}\rvert$	$\boldsymbol{\alpha}$	$\lvert\boldsymbol{\Delta\hat{m}^L - \Delta m}\rvert$	$\boldsymbol{\alpha}$
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

We use Elastic-Net loss (with a 0.99 L1 component) and SAGA as the algorithm to minimize it. To balance the classes, we adjust class weights inversely proportional to class frequency. To find these hyperparameters we do not optimize for best generalization performance, as we are interested in inducing an disparate performance between the groups.

We evaluated the model selection by 10 runs of hold-out cross-validation (70–30% as the random training–testing split). We fix the PRNG seed and release the source code included in the supplementary material.

We published the data and the source code to reproduce these experiments (Juarez and Korolova, 2022).

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 ν^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 generate enough data to represent a realistic DFL deployment.

Table 2 compares the empirical error with the 0.99-probability bounds (α) obtained with the procedure explained in the previous section, for a range of 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.