“You Can't Fix What You Can't Measure”

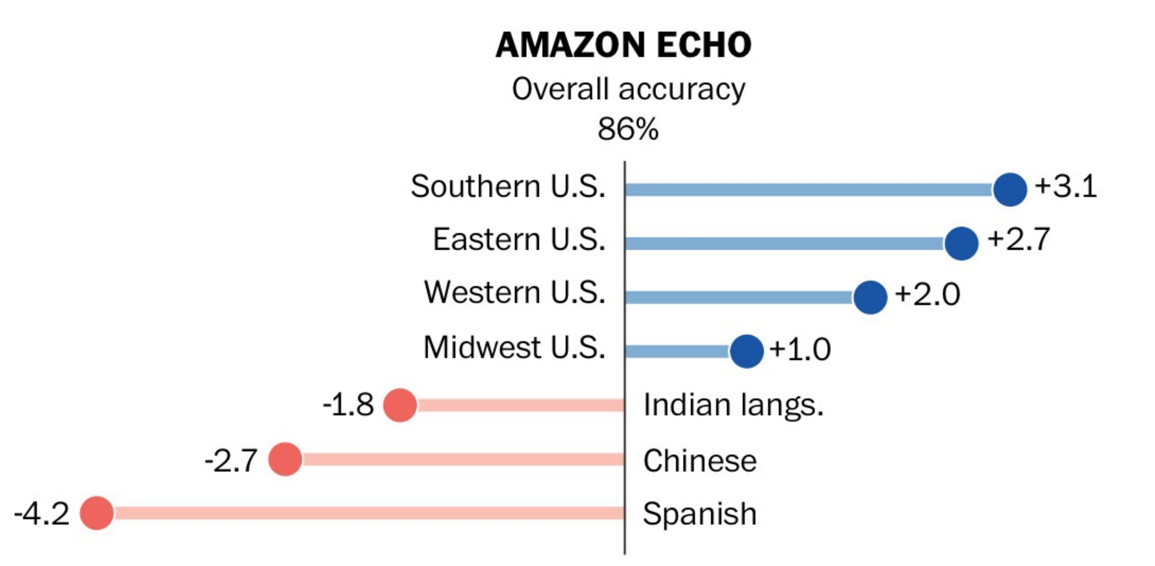
Privately Measuring Demographic Performance Disparities in Federated Learning

Marc Juarez (marc.juarez@ed.ac.uk) and Aleksandra Korolova (korolova@princeton.edu)

Introduction

Disparate performance of machine learning models across demographic groups can lead to disparate impact.

Example: *when waking up Amazon Echo, False Positive samples are sent to the cloud for further processing and may contain background speech. If a group has a higher False Positive Rate (FPR), it is more exposed to surveillance.*



Disparate accuracy of Amazon Echo's speech recognition across demographic groups. The Washington Post.

However, **access to the group membership attributes** that are needed to identify a performance disparity (e.g., ethnicity) is often **unavailable for privacy reasons.**

Objective & Impact

Prior works **overlook the measurement of the disparities**: they focus on correcting them once they know they exist. But enabling practical measurements of the disparities is the first step toward identifying and fixing the issues—**You Can't Fix What you Can't Measure!**

Definition (*Performance Gap*): the absolute difference between group averages of a performance metric (e.g., FPR):

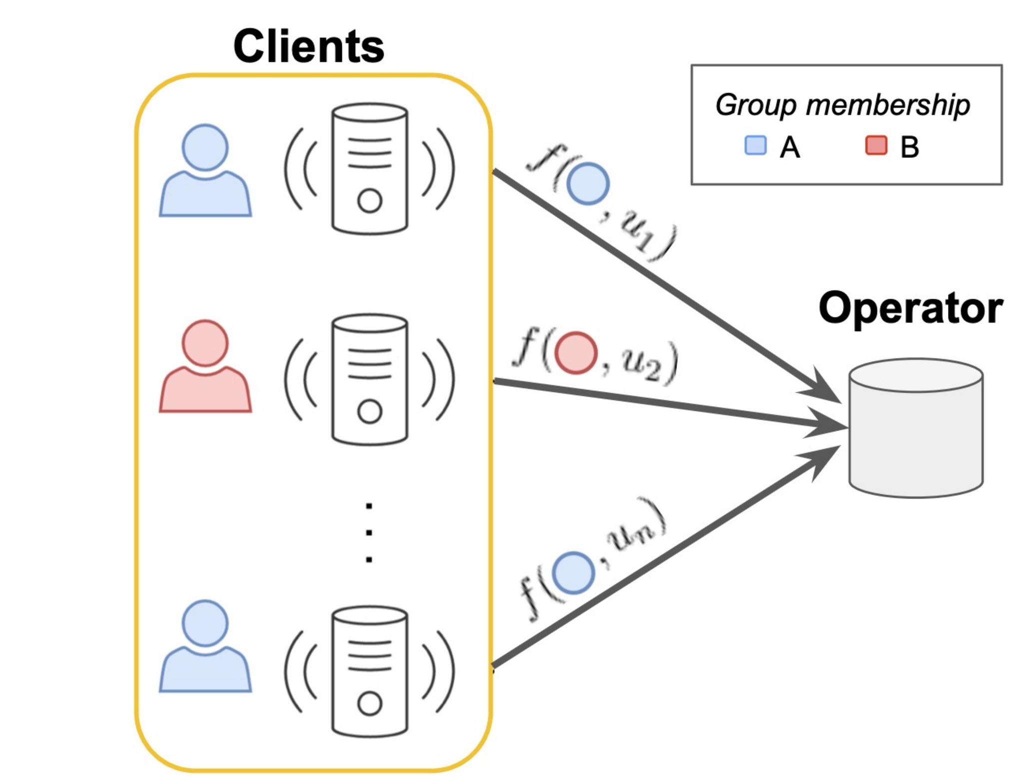
$$\Delta m := |\overline{FPR_A} - \overline{FPR_B}|$$

Problem statement: how can we measure the Performance Gap, **while protecting the privacy** of the group membership attributes?

Objective: design **Local Differential Privacy** (LDP) mechanisms to measure the Performance Gap.

Approach

We must protect both the **performance** and **group membership** information, as they are correlated. However, **preserving the overall correlation is necessary to ensure high-accuracy** measurements of the Gap.



f is one of our LDP mechanisms. The clients use f to protect the group membership and the performance information.

We design two novel families of LDP mechanisms by composing LDP primitives:

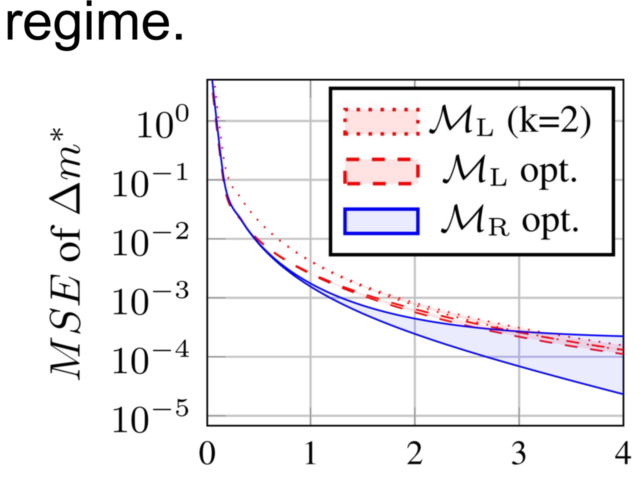
$\mathcal{M}_{\mathrm{R}}$ and $\mathcal{M}_{\mathrm{L}}$

Theoretical evaluation: bound the error of the mechanisms as a function of privacy (ϵ).

Comparison between M's

RQ1: *best method given a privacy budget?*

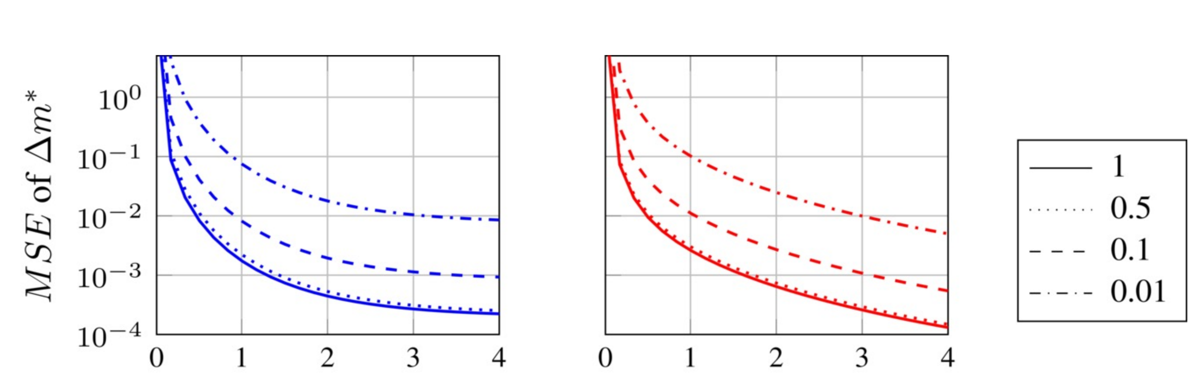
The MSE is small for typical privacy budgets. No mechanism is optimal: it depends on the privacy regime.



Upper and lower bounds of the estimators' MSE for different overall privacy budgets.

RQ2: *effect of group size ratio?*

For common group ratios: 1:1, 1:2, and 1:10 (e.g., race, sex), the mechanisms maintain a small MSE.



Upper bound of the MSE of M_R (left) and M_L (right) for different group ratios.

Federated Learning

Cross-device Federated Learning (CFL) is a popular machine learning paradigm. Because CFL **aspires to provide data privacy**, the challenges of protecting sensitive attributes are even **more relevant** than in other settings.

RQ3: *can existing CFL deployments afford the privacy budget required by the mechanisms?*

We show that the size of current CFL deployments (e.g., by Apple and Google) **allows for accurate measurements** of the Performance Gap **even under the strong privacy guarantees** of LDP.

	$\mathcal{M}_{\mathrm{R}}$			$\mathcal{M}_{\mathrm{L}}$		
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

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

K for existing CFL deployments:

- 10^8 active Siri clients[1].
- 10^9 install of Gboard in Android[2].

[1]Apple Newsroom, 2018
[2]Google Play Store, 2021

Conclusion

We explore the space of LDP-based solutions to measure the disparate performance of machine learning models **while preserving the privacy of the group membership information**.

Specifically, the sheer number of clients in **CFL offers a unique opportunity** to measure performance disparities, thus **raising awareness** of new issues and driving work towards fixing them.

We believe our work paves the way for **service providers, regulatory agencies, or even coalitions of users** to make measurements of the Performance Gap and **uncover existing performance disparities** in deployed models.