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

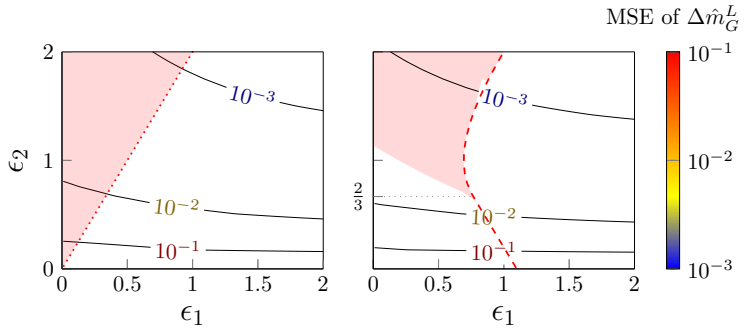

Figure 3: Contour plot of the MSE of $\Delta\hat{m}^L$ for $k = 2$ (left) and $k = \frac{2}{3}$ (right), as a function of ϵ_1 and ϵ_2. The colored area is the region where the parameters satisfy ϵ-LDP for $\epsilon = \epsilon_2$. The curves represent the optimal allocations when $k = 2$ (dotted) and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

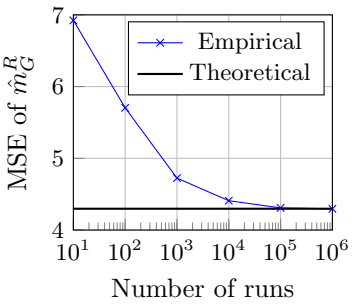

Figure 4: The theoretical upper bound of the MSE of $\hat{m}_G^R$ as derived from Theorem 9, and its empirical MSE over different runs of $\mathcal{M}_{\mathrm{R}}$, for $n_G = n_{\bar{G}} = 10$.

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.**	
$\boldsymbol{\epsilon}$	$\lvert\boldsymbol{\Delta}\hat{\boldsymbol{m}}^{\boldsymbol{R}} - \boldsymbol{\Delta m}\rvert$	$\boldsymbol{\alpha}$	$\lvert\boldsymbol{\Delta}\hat{\boldsymbol{m}}^{\boldsymbol{L}} - \boldsymbol{\Delta m}\rvert$	$\boldsymbol{\alpha}$
0.01	0.1241($\pm$0.1410)	1.2586	0.0504($\pm$0.0337)	1.0525
0.10	0.0082($\pm$0.0059)	0.1206	0.0046($\pm$0.0040)	0.1060
1.00	0.0008($\pm$0.0006)	0.0094	0.0008($\pm$0.0005)	0.0118
10.00	0.0001($\pm$0.0000)	0.0032	0.0001($\pm$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.