

DPBalance: Efficient and Fair Privacy Budget Scheduling for Federated Learning as a Service

Yu Liu*, Zibo Wang*, Yifei Zhu*[‡], and Chen Chen[§]

*UM-SJTU Joint Institute, Shanghai Jiao Tong University

[‡]Cooperative Medianet Innovation Center (CMIC), Shanghai Jiao Tong University

[§]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University

Email: yu_liu@sjtu.edu.cn, wangzibo@sjtu.edu.cn, yifei.zhu@sjtu.edu.cn, chen-chen@sjtu.edu.cn

Abstract—Federated learning (FL) has emerged as a prevalent distributed machine learning scheme that enables collaborative model training without aggregating raw data. Cloud service providers further embrace Federated Learning as a Service (FLaaS), allowing data analysts to execute their FL training pipelines over differentially-protected data. Due to the intrinsic properties of differential privacy, the enforced privacy level on data blocks can be viewed as a privacy budget that requires careful scheduling to cater to diverse training pipelines. Existing privacy budget scheduling studies prioritize either efficiency or fairness individually. In this paper, we propose DPBalance, a novel privacy budget scheduling mechanism that jointly optimizes both efficiency and fairness. We first develop a comprehensive utility function incorporating data analyst-level dominant shares and FL-specific performance metrics. A sequential allocation mechanism is then designed using the Lagrange multiplier method and effective greedy heuristics. We theoretically prove that DPBalance satisfies Pareto Efficiency, Sharing Incentive, Envy-Freeness, and Weak Strategy Proofness. We also theoretically prove the existence of a fairness-efficiency tradeoff in privacy budgeting. Extensive experiments demonstrate that DPBalance outperforms state-of-the-art solutions, achieving an average efficiency improvement of $1.44\times \sim 3.49\times$, and an average fairness improvement of $1.37\times \sim 24.32\times$.

Index Terms—Differential privacy, federated learning, resource allocation, privacy budget, fairness, efficiency

I. INTRODUCTION

The proliferation of data on edge devices and the advancement of machine learning models drive the wide deployment of machine learning services. However, existing model training paradigm has raised severe privacy concerns since sensitive information needs to be uploaded to central servers to be analyzed. In response to this, privacy protection laws like the General Data Protection Regulation of EU (GDPR) [1] have been established worldwide, which further drives the emergence of federated learning (FL). FL is a privacy-preserving distributed machine learning scheme where edge devices collaboratively train a global model while keeping their data local. It has been widely studied and deployed

nowadays in applications such as Google keyboard [2]–[4], hotword detection [5], and medical research [6].

With the advent of FL techniques, there has been a notable rise in distributed machine learning platforms like FedML [7], FATE [8] that offer a cloud-based “federated learning as a service” (FLaaS) [9]. With the lifetime of a federated model training process being represented as a pipeline, these platforms facilitate training of diverse pipelines from data analysts using crowdsourced data owners. To further protect and manage privacy loss, differential privacy (DP) is adopted in the FLaaS platform [10] [11]. It ensures the output of FL pipeline would not change much for any neighbor input datasets. DP is enforced by adding random noise to intermediate data like gradient [12] of FL training. With the help of DP mechanisms, data owners can specify a privacy level on their continuously generated data blocks. Each data analyst submits multiple FL pipelines with different training demands on the amount of data and its privacy level.

According to the definition of DP, the training of a FL model on a data block introduces a certain level of privacy loss. Furthermore, this privacy loss is additive when data is utilized for multiple times. As a result, privacy in FLaaS platforms can be perceived as a consumable and quantifiable resource. The privacy level set by each crowdsourced data owner can be regarded as a privacy resource budget. Consequently, the efficient scheduling of FL pipelines to train on crowdsourced private data within the privacy budget becomes a critical challenge for FLaaS platforms.

However, while privacy can be treated as a resource, traditional resource allocation methods in distributed computing cannot be directly applied. The primary distinction lies in the nature of privacy resources, which are characterized as *non-replenishable* and *continuously generated* as new data is collected. Unlike traditional resources such as CPU and memory which can be released for subsequent tasks after prior usage, privacy resources are not recoverable once their budget has been depleted. Furthermore, allocating available resources solely based on demand, a strategy that may work for CPU/memory allocation, may be however inefficient or unfair for privacy resources.

Existing works in privacy budget scheduling solely consider either efficiency [13]–[17] or fairness [18]. So far there is no privacy budget scheduling mechanism that jointly considers

This work is supported by the National Natural Science Foundation of China (Grant No. 62302292), the National Key R&D Program of China (Grant No. 2023YFB2704400), and the Fundamental Research Funds for the Central Universities. The work of C. Chen is supported by National Natural Science Foundation of China (Grant No. 62202300) and Shanghai Pujiang Program (Grant No. 22PJ1404600). Yifei Zhu and Chen Chen are corresponding authors.

them both. In fact, efficiency and fairness of an allocation scheme are complicatedly affected by many factors. The FL platform may hope to let the pipeline await to obtain a boarder view of the allocation market to achieve a better performance in system throughput and fairness. However, requiring the pipelines to wait for a long time increases the latency, which hurts the final efficiency from a different perspective. In addition, naively allocating the majority of resources to those low-demand pipelines to maximize the number of completed pipelines may make the allocation among data analysts imbalanced, which hurts fairness. Therefore, it is challenging to derive an allocation comprehensively taking efficiency and fairness into consideration in the context of FLaaS platform.

In this paper, we present the first privacy budget scheduling mechanism, DPBalance, that jointly considers efficiency and fairness. We first develop a comprehensive utility function incorporating dominant shares across different FL pipelines to calculate data analyst-level fairness and platform-level efficiency. To capture the characteristics of FLaaS training, FL-specific performance metrics, such as the matching degree of data and FL models, are further integrated into the utility model. A sequential allocation mechanism is then designed using the Lagrange multiplier method and effective greedy heuristics. From the perspective of fairness, we theoretically prove that our solution can simultaneously satisfy four key economic properties. From the perspective of efficiency, the most efficient allocation result under the privacy preference is derived by maximizing platform’s utility function. In summary, the contributions of our work are as follows.

- We present the first privacy budget scheduling mechanism, DPBalance, that jointly considers efficiency and fairness in the context of FLaaS model training.
- We design a sequential allocation algorithm using the Lagrange multiplier method and effective greedy heuristics so that data analyst-level fairness and platform-level efficiency can be maximized.
- We theoretically prove that DPBalance satisfies four essential economic properties: Pareto Efficiency, Sharing Incentive, Envy-Freeness, and Weak Strategy Proofness.
- We discover and theoretically prove the existence of a tradeoff between fairness and efficiency under practical conditions.
- Extensive experiments show that DPBalance outperforms other state-of-the-art budget scheduling baselines by $1.44\times \sim 3.49\times$ in efficiency and $1.37\times \sim 24.32\times$ in fairness on average.

The remainder of this paper is constructed as follows: Section II summarizes the related work. Background and motivation are presented in Section III. In Section IV, we present the system model and problem formulation of privacy budget scheduling. Algorithm design and theoretical analysis are presented in Section V. The evaluation part is presented in Section VI, followed by the conclusion in Section VII.

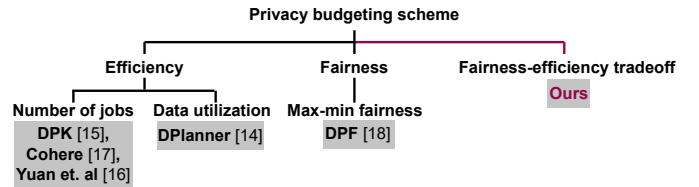


Fig. 1. Taxonomy of existing privacy budgeting studies and ours (DPBalance).

II. RELATED WORK

A. Resource allocation in FL

Existing resource allocation studies in FL mainly focus on allocating traditional resources, like energy consumption [19], bandwidth [20] or computation resources [21], [22] to train a single model. In FEDL [19], the uplink transmission rate is allocated to minimize energy cost and training time. In [20], a joint device scheduling and resource allocation policy is proposed to maximize model accuracy with a limited training time budget and latency threshold. In q-FEL [21], fairness in accuracy across different users is considered when allocating edge devices. In [22], a deep learning-based auction mechanism is proposed to infer the valuation of each cluster head’s service to allocate clusters’ computation resources. However, none of these works consider privacy as resources and study privacy allocation problems in the emerging FLaaS setting, where multiple FL models need to be trained on top of privacy-preserved data owners.

B. Privacy budgeting for distributed model training

Current studies on privacy budget scheduling can mainly be categorized into two types: efficiency-oriented [13]–[17] and fairness-oriented approaches [18]. In [14], privacy budget is allocated to maximize the overall quality of query results with an online budgeting algorithm. In [15]–[17], privacy budget allocation problem is formulated to maximize the number of allocated pipelines, and adopt heuristic algorithms [15], [16], or solver [17] to resolve it. DPF [18] focuses on the fairness in allocation and is designed to ensure first N pipelines’ max-min fairness at the pipeline level. Different from these studies, we propose the first work to consider both fairness and efficiency in privacy budget scheduling. We further present a comprehensive comparison of our work to others in Fig. 1.

III. BACKGROUND AND MOTIVATION

A. Why privacy can be allocated?

In this subsection, we introduce bounded property, composition of privacy to illustrate why privacy can be regarded as resources to be allocated.

To show the bounded property, we introduce the state-of-the-art DP, Rényi Different Privacy [23] [24], and the resulting privacy loss from it. We first give the formal definition of Rényi divergence:

Definition 1 (Rényi Divergence). *Let Γ and Υ be two distributions, Rényi divergence \mathcal{RD}_α of order α is defined as*

$$\mathcal{RD}_\alpha(\Gamma||\Upsilon) \triangleq \frac{1}{\alpha-1} \log E_{x \sim \Upsilon} \left(\frac{\Gamma(x)}{\Upsilon(x)} \right), \quad (1)$$

where $\Gamma(x)$ and $\Upsilon(x)$ are densities of Γ and Υ at point x .

Then, Rényi Different Privacy is defined as follows:

Definition 2 ((α, ϵ) -Rényi Different Privacy (RDP)). *Let M be the random mechanism processing on two adjacent datasets D and D' that only differ in one record. (α, ϵ) -Rényi Different Privacy holds that*

$$\mathcal{RD}_\alpha(M(D)||M(D')) \leq \epsilon, \quad (2)$$

where Rényi divergence $\mathcal{RD}_\alpha(M(D)||M(D'))$ is equivalent to the privacy loss of dataset D .

We call the ϵ -bounded privacy loss as the bounded property. When α approaches to ∞ , (α, ϵ) -RDP is equivalent to the basic $(\epsilon, 0)$ -DP. RDP is widely applied in the privacy budgeting solutions [15]–[18] for its advantages over basic DP: 1) it bounds privacy loss tighter when achieving similar performance in FL; 2) it permits more convenient composition of multiple different DP mechanisms including Gaussian mechanism and Laplacian mechanism, and scales better.

As different data analyst requires different amount of data, we partition growing dataset of each FL device into data blocks based on time, namely $\mathcal{S} = \{d_k | k \in [1, \infty)\}$, where \mathcal{S} is the data block set. In this way, each data analyst can specify different number of data blocks. To make each data block reusable and accountable for privacy loss each time being used, it's essential to introduce parallel composition and sequential composition, which depict the additive property of privacy.

Definition 3 (Parallel composition). *We denote each data block i 's privacy loss as ϵ_i . Then, the privacy loss for the whole growing dataset is $\max \epsilon_i, \forall i \in [1, \infty)$.*

Definition 4 (Sequential composition). *Let the total privacy budget of FL device be ϵ_g . We assume it has been used K times by different FL pipelines, and each time it is consumed ϵ_i privacy budget. Then, the privacy loss for this data block can be $\sum_{i=1}^K \epsilon_i$.*

Intuitively, parallel composition guarantees the privacy loss of dataset is bounded by the maximum privacy loss of its single privacy loss. Sequential composition depicts that privacy loss is additive for a single data block. It is straightforward to prove that these two properties hold perfectly for (α, ϵ) -RDP.

As privacy is limited due to the bounded property and additivity due to the composition property, it can be regarded as a special resource to be allocated.

B. Effects of bad scheduling

As mentioned, privacy can be considered as an important resource to be allocated. When the limited resource is allocated to multiple pipelines, efficiency and fairness of the allocation becomes a critical metric [13], [21], [25]–[27].

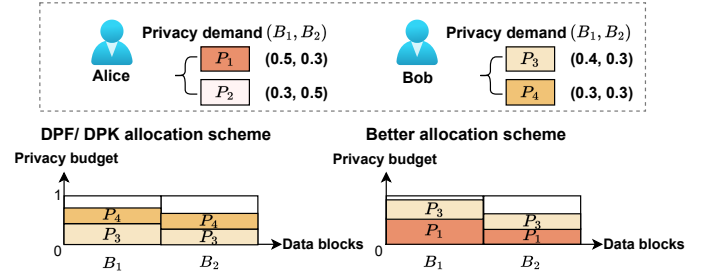


Fig. 2. Allocation results under different schemes for a particular time slot.

Consider that data blocks are constantly contributed by edge devices and data analysts constantly coming in with their pipelines. Each pipeline demands certain data blocks with privacy demands. Privacy demand can be regarded as the amount of privacy loss of a certain data block specified by a data analyst. In our context, dominant share is defined as the maximum privacy budget among all required data blocks of a training demand from a pipeline or data analyst. Training demand includes number of data blocks and privacy demand on each data block.

Definition 5 (Pipeline's maximum share). *A pipeline's maximum share of all the required data block's privacy resources is defined as*

$$\mu_{ij} = \max_k \gamma_{ij}^{<k>}, \quad (3)$$

where $\gamma_{ij}^{<k>}$ is the normalized privacy demand of pipeline j from data analyst i for data block k .

Thus, this pipeline's dominant share can be represented as $\mu_{ij} x_{ij}$. x_{ij} is the ratio of allocated resource to privacy demand of pipeline j from data analyst i , which can also be considered as allocated result to pipeline j .

Definition 6 (Data analyst's maximum share). *Data analyst i 's maximum share of all the required data block's privacy resource is defined as*

$$\mu_i = \max_k \gamma_i^{<k>}, \quad (4)$$

where $\gamma_i^{<k>}$ is the normalized privacy demand that data analyst i wants for data block k on all selected edge devices, namely, $\gamma_i^{<k>} = \sum_{j=0}^{n_i} \gamma_{ij}^{<k>} x_{ij}$.

Thus, data analyst i 's dominant share is represented as $\mu_i x_i$, which indicates the degree to its demand was satisfied. x_i is the ratio of allocated resource to privacy demand of data analyst i , which can be considered as allocation to data analyst i .

To illustrate the shortcomings of bad scheduling, we give an example in Fig. 2. Consider two data analysts named Alice and Bob, two data blocks B_1, B_2 whose total privacy budget is 1.0. Alice has two pipelines (P_1, P_2). Pipeline P_1 demands $(0.5, 0.3)$ for (B_1, B_2) , where the normalized privacy demand $(\gamma^{<1>}, \gamma^{<2>})$ for (B_1, B_2) is $(0.5, 0.3)$ and maximum share μ is 0.5. Pipeline P_2 demands $(0.3, 0.5)$ for (B_1, B_2) , where the normalized privacy demand $(\gamma^{<1>}, \gamma^{<2>})$ for (B_1, B_2)

is $(0.3, 0.5)$ and maximum share μ is 0.5. Bob also has two pipelines (P_3, P_4) . Pipeline P_3 demands $(0.4, 0.3)$ for (B_1, B_2) , where the normalized privacy demand $(\gamma^{<1>}, \gamma^{<2>})$ for (B_1, B_2) is $(0.4, 0.3)$ and maximum share μ is 0.4. Pipeline P_4 demands $(0.3, 0.3)$ for (B_1, B_2) , where the normalized privacy demand $(\gamma^{<1>}, \gamma^{<2>})$ for (B_1, B_2) is $(0.3, 0.3)$ and maximum share μ is 0.3.

In this example, DPK [15], which is the state-of-the-art privacy scheduling method focusing on efficiency and allocating resources to the pipeline with the lowest weight-to-demand ratio, allocates pipeline P_3 and P_4 shown in the left side of Fig. 2. In addition, DPF [18], which is the state-of-the-art privacy scheduling method focusing on fairness and allocating resources to the pipeline with the smallest dominant share, gets the same allocation result as DPK. Efficiency can be evaluated with dominant share and leftover privacy resources. Based on the definition in (3), the overall dominant share of DPF and DPK solution is 0.7. From the perspective of leftover privacy resources, DPF and DPK waste $0.3 B_1$ and $0.4 B_2$. We can figure out another solution shown on the right side of Fig. 2, pipeline P_1 from Alice and pipeline P_3 from Bob can be satisfied at the same time. Furthermore, this allocation result leaves less unused resource, as it leaves $0.1 B_1$ and $0.4 B_2$ unused, and achieves 0.9 overall dominant share. Furthermore, it's much fairer because there is no analyst who occupies all the privacy resources.

IV. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we first introduce the threat model and the entities involved in our framework. Then, we formally formulate the privacy resource allocation in FLaaS platforms as an optimization problem.

A. Threat model

In the FLaaS platform, the data analysts submit multiple pipelines and receive them after being trained by the private data held by vast data contributors. The trained pipelines exposed to data analysts or other entities, although does not explicitly contain the raw data, still leak sensitive information of the individual data contributors and are vulnerable to privacy threats such as membership inference attacks [28] or data reconstruction attacks [29]. To tackle the privacy threats, we apply the DP model, which restricts information leakage by introducing randomness to the outputs, to the trained pipelines [11]. In the FLaaS platform, one data block may be utilized in training multiple pipelines, where an adversary has an increased ability in performing attacks with an aggregated view of multiple pipelines. Therefore, we propose privacy budget scheduling mechanism in the FLaaS platform, where each data block satisfies (α, ϵ) -RDP over the whole execution after careful DP assignment on each pipeline.

B. System overview

The framework of DPBalance is depicted in Fig. 3 and it considers three entities: Data analysts, FLaaS platform, and data contributors. Data contributors can be mobile phones,

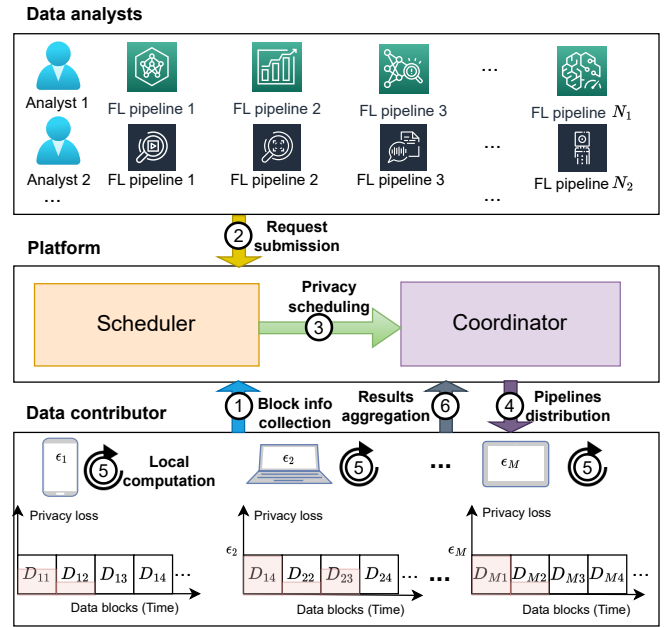


Fig. 3. Framework of DPBalance

tablets, or other edge devices with growing databases. In this system, Q edge devices join the FLaaS platform training and each participant has its own database partitioned in many data blocks by time. We denote the number of data blocks from all edge devices in total as K . We assume there is no edge device in and out and data on devices are continuously generated.

Our framework includes M data analysts, and each data analyst i consists of N_i pipelines. FL pipelines are associated with their training demands. A pipeline can only be successfully executed if its training demand is met. Pipelines are willing to get data blocks with more privacy loss as possible because it could improve the performance of trained pipelines. In addition, we assume data analysts are greedy and selfish. Greedy means the data analyst wants to satisfy more pipelines and selfish means the data analyst wouldn't return the resource back if its pipelines could benefit from it. The main objective of the FLaaS platform is to allocate a certain amount of data blocks efficiently and fairly under limited privacy resources. An unfair allocation scheme would discourage data analysts from participating and an inefficient allocation scheme would cause a waste of privacy resources.

Next, we briefly introduce the workflow of our system in Fig. 3. In the training demand formulation part, the scheduler collects data analysts' general demand, including the number of demanding devices N , privacy budget $\vec{\epsilon}_j$ and time interval \vec{T}_j of demanding blocks, and randomly selects a set of N edge devices \vec{N}_j . After getting the privacy demands, the scheduler allocates privacy resources and sends results to the platform. Then, the platform distributes successfully allocated pipelines to the target devices for training and aggregates.

C. Privacy resource model

We partition each device's local data into many data blocks by time and partition granularity can be hour, day, or month. Each FL pipeline can specify the data blocks from a certain range of time. Every time the FL pipeline is trained on a local device, data blocks would lose part of their privacy. As time goes by, private data used for training FL pipelines would grow and get retired after exposing a certain amount of privacy. Data blocks can serve other FL pipelines before retirement.

We define the privacy budget of data contributor i 's whole dataset as ϵ_i^g and any privacy budget of data block j of it as ϵ_{ij}^g . We define device i 's privacy loss as ϵ_i^c and any privacy budget of data block j of it as ϵ_{ij}^c . We assume privacy loss should not exceed privacy budget ($\epsilon_i^c \leq \epsilon_i^g, \epsilon_{ij}^c \leq \epsilon_{ij}^g$). Furthermore, we assume different edge device i has different privacy budget ϵ_i^g . In addition, to keep (α, ϵ_i) -RDP for edge device i , data block j on edge device i holds the same privacy budget as the global budget held by each edge device ($\epsilon_{ij}^g = \epsilon_i^g$). Furthermore, we assume the privacy allocation follows one-or-more property.

Definition 7 (One-or-more). *A pipeline from a data analyst would be successfully allocated if and only if its allocated resource is equal to or more than its minimal demand. Formally speaking, it can be represented as follows:*

$$\mathcal{D}' = \varkappa \mathcal{D}, \varkappa \geq 1, \quad (5)$$

where \mathcal{D} is the privacy demand vector of this pipeline, and \mathcal{D}' is the vector representing the amount of privacy resource allocated to the pipeline.

This property is unique to privacy resources, while traditional resources (e.g. CPU, memory) do not have.

D. Data analyst's efficiency model

Before defining the platform's utility, we define the data analyst's efficiency first. We adopt dominant share, weighted by the data-pipeline matching degree and pipeline delay time as the efficiency of a data analyst. Pipeline delay time means how long a pipeline has been waiting since it proposes training demand. The data-pipeline matching degree means the importance of a certain data block to a certain pipeline measured by training loss [14] [30]. We assume longer delay time and lower training loss bring lower efficiency. We use the weighted average of the training losses l_{ij} of all data blocks on device i to represent the training loss l_i of this device.

$$l_i = \sum_{j=1}^n \frac{\mu_{ij}}{\sum_{j=1}^n \mu_{ij}} l_{ij}. \quad (6)$$

The data analyst's efficiency is defined as follows:

Definition 8 (Data analyst's efficiency). *Data Analyst's Efficiency is defined as data analyst i 's dominant share weighted by waiting time coefficient $T(t_i)$ and training loss l_i .*

$$U_i(x_i) = \mu_i x_i T(t_i) l_i, \quad (7)$$

where $T(t_i)$ can be any monotonic decreasing function of data analyst's waiting time t_i .

E. Platform's utility model

The platform's utility considers two perspectives. First, it pursues a higher overall efficiency for all data analysts. Second, it tries to balance the efficiency of the data analysts for fairness. Therefore, we present to model the platform's utility incorporating both efficiency and fairness.

For the efficiency, we define the following dominant efficiency to represent the platform's efficiency.

Definition 9 (Dominant efficiency). *Dominant efficiency is defined as the summation of all data analysts' efficiency*

$$\sum_{i=1}^m U_i(x_i). \quad (8)$$

For the fairness metric, we consider data analyst-level fairness, which means we consider fairness among data analysts. We give the formal definition of dominant fairness.

Definition 10 (Dominant fairness). *We define dominant fairness $f_\beta(x_i)$ from the perspective of dominant share $\mu_i x_i$, average delay time t_i , and weighted average training loss l_i .*

$$f_\beta(x_i) = \text{sgn}(1 - \beta) \left(\sum_{i=1}^m \left(\frac{U_i(x_i)}{\sum_{i=1}^m U_i(x_i)} \right)^{1-\beta} \right)^{\frac{1}{\beta}}, \quad (9)$$

where $\text{sgn}(\cdot)$ is the signum function and β can be considered as the fairness preference of platform manager.

After that, we combine the efficiency metric and fairness metric and formulate the platform's utility function as

$$\Psi_\lambda(x_i) = \text{sgn}(1 - \beta) \left(\sum_{i=1}^m \left(\frac{\mu_i x_i T(t_i) l_i}{\sum_{i=1}^m \mu_i x_i T(t_i) l_i} \right)^{1-\beta} \right)^{\frac{1}{\beta}} \left(\sum_{i=1}^m \mu_i x_i T(t_i) l_i \right)^\lambda, \quad (10)$$

where λ is the efficiency preference of the platform manager.

When $\lambda = |\frac{1-\beta}{\beta}|$, (10) can be transformed into

$$\max \text{sgn}(1 - \beta) \left(\sum_{i=1}^m \left(\mu_i x_i T(t_i) l_i \right)^{1-\beta} \right)^{\frac{1}{\beta}}, \quad (11)$$

which is equivalent to the famous α -fairness function [31]

$$\max \sum_{i=1}^m \frac{\left(\mu_i x_i T(t_i) l_i \right)^{1-\beta}}{1 - \beta}. \quad (12)$$

F. Problem formulation

From the perspective of the FLaaS platform, we aim at maximizing the platform's utility. Formally, we can formulate our problem as

$$\max \Psi_\lambda(x_i), \quad (13)$$

$$s.t. \sum_{i=1}^m \gamma_i^{<k>} x_i T(t_i) l_i \leq 1, \forall k \in [1, K], \quad (14)$$

$$\gamma_i^{<k>} = \sum_{j=0}^{n_i} \gamma_{ij}^{<k>} x_{ij}, \forall i \in [1, M], \forall k \in [1, K], \quad (15)$$

$$x_{ij} = 0 \text{ or } x_{ij} \geq 1. \quad (16)$$

This optimization problem is constrained by resource availability in (14), where $\gamma_i^{<k>}$ specifies the normalized demand that data analyst i wants for resource k . Constraint (15) represents the demand of data analyst i is composed of demands of all its pipelines. Constraint (16) is designed because of one-or-more property.

V. ALGORITHM DESIGN AND THEORETICAL ANALYSIS

A. Algorithm design

The original optimization problem is non-convex and discontinuous because constraint in (16) contains a discrete point $x_{ij} = 0$. Classical solutions for mixed integer optimization problems, including the branch and bound method, dynamic programming, and cutting plane method, are iterative and cannot provide a closed-form solution, which makes it hard to analyze the economic properties. Therefore, we decompose the original problem into two sub-problems. The first sub-problem is to allocate privacy resources among data analysts and is formulated as

$$\max \Psi_\lambda(x_i), \quad (17)$$

$$s.t. \sum_{i=1}^m \gamma_i^{<k>} x_i T(t_i) l_i \leq 1, \forall k \in [1, K], \quad (18)$$

$$x_i \geq 0. \quad (19)$$

The second sub-problem is to reallocate privacy resources among pipelines in one data analyst to cover more pipelines. It is formulated as

$$\max \sum_{j=1}^n \mu_{ij} x_{ij} T(t_{ij}) l_{ij}, \quad (20)$$

$$s.t. \gamma_i^{<k>} x_i = \sum_{j=0}^{n_i} \gamma_{ij}^{<k>} x_{ij}, \forall i \in [1, M], \quad (21)$$

$$x_{ij} = 0 \text{ or } x_{ij} \geq 1. \quad (22)$$

We then propose a novel sequential allocation algorithm to solve them. The detailed algorithm is presented in Algorithm. 1. After decomposition, the first optimization problem becomes convex and continuous. We first assemble pipelines' training demands to formulate data analyst i 's training demand and calculate its maximum share of all required resources μ_i using (3) (line 1). We then solve the optimization problem in (17) along using Lagrange multiplier method and output each data analyst's allocation result X (line 2).

We solve the second sub-problem in a greedy heuristic way, and the general policy is to return unused resources to

the resource pool (line 4). To satisfy the greedy and selfish characteristics of data analysts mentioned in Section IV and cover more pipelines in one data analyst, we temporarily relax the one-or-more property in (22) into integer constraint shown in (24). For each data analyst i , we get the successfully allocated pipelines set χ_i by solving problem in (23) (line 5).

$$\max \sum_j \Gamma(x_{ij}), \forall i \in [0, M], \quad (23)$$

where

$$\Gamma(x) = \begin{cases} 1, & x_{ij} \geq 1, \\ 0, & x_{ij} = 0. \end{cases} \quad (24)$$

This action is to allocate fewer resources to each analyst and select as many pipelines as possible. Secondly, to make full use of allocated resources for data analysts, we choose to set $x_{ij} \in \chi_i$ and maximize the second sub-problem in (20) using the commercial Gurobi solver [32] (line 6). Each data analyst i returns unused privacy resource back if there is privacy resource left or its γ_{ij} from pipeline j with minimal μ_{ij} Pareto-dominates than data analyst i 's share $\gamma_i x_i$ because of one-or-more property (line 7), where the latter means left resource can not meet the minimal requirement of any pipeline in the data analyst. This behavior is also a kind of economizing behavior, which can give resources to the data analyst who will come continuously in the future. Finally, the coordinator assembles all pipelines' allocation result and form the global allocation result X' (line 8).

We further use the allocation example in Fig. 2 to illustrate how our algorithm works. Firstly, we set fairness preference $\beta = 2.2$ and efficiency preference $\lambda = \frac{\beta-1}{\beta}$. Then, we assemble data analysts' training demands by adding up all its pipelines' demand. After that, we solve the optimization problem in (17) along with its constraints and get two data analysts' output. Alice gets (0.5, 0.5) for (B_1, B_2) , and Bob gets (0.5, 0.42) for (B_1, B_2) . As both data analysts are allocated with partial resources, they do not need to return resources to the resource pool. Then, by maximizing the optimization in (20) along with constraints, we can get the resource allocation of two data analysts to their pipelines. Alice allocates (0.5, 0.3) of (B_1, B_2) to pipeline P_1 . Bob allocates (0.5, 0.375) of (B_1, B_2) to pipeline P_3 . Finally, Alice returns 0.2 B_2 and Bob returns 0.045 B_2 back. In this way, our algorithm achieves 1.0 dominant efficiency and allocates 2.25 pipelines, while DPF and DPK only achieve 0.7 dominant efficiency and allocates 2 pipelines successfully.

B. Four key properties of fairness

In this subsection, we introduce four definitions of key properties for fairness, and prove under what conditions these properties can be satisfied.

Definition 11 (Pareto Efficiency (PE)). *Suppose that we have 2 vectors \mathbf{x} and \mathbf{y} , if $x_i \geq y_i$ for any index i and $x_j > y_j$ for some index j , we denote \mathbf{x} Pareto-dominates \mathbf{y} .*

Algorithm 1 Sequential allocation algorithm in DPBalance

Input:

The training demand of all the pipelines, $d_{ij} = [d_{ij}^{<1>}, d_{ij}^{<2>}, \dots, d_{ij}^{<K>}]$, $i \in [1, M]$, $j \in [1, N]$;

The delay time of all the data analysts' pipelines, $t = [t_1, t_2, \dots, t_M]$;

The training loss of all the data analysts' pipelines, $l = [l_1, l_2, \dots, l_M]$, where $l_i = [l_{i1}, l_{i2}, \dots, l_{iM}]$;

The hyper-parameter β and λ ;

Output:

Allocation result, $X' = [x_1, x_2, \dots, x_m]$, $x_i = [x_{i1}, x_{i2}, \dots, x_{iN}]$, $\forall i \in [1, M]$;

- 1: Calculate data analyst i 's maximum share of all the required resource μ_i and γ_{ij} using (3);
 - 2: Solve the first sub-problem in (17) along with constraints in (18) and (19), and output the data analyst i 's primary allocation result X ;
 - 3: **for** each data analyst i **do**
 - 4: Return allocated resource if γ_{ij} from pipeline j with lowest μ_{ij} Pareto-dominates $\gamma_i x_i$ and break;
 - 5: Maximize the optimization problem in (23) with constraints and get the successful allocated pipeline set χ_i ;
 - 6: Maximize optimization problem in (20) and form pipeline's allocation result x_{ij} ;
 - 7: Return unused privacy resources back.
 - 8: Assemble all the pipelines' allocation results from all data analysts and form the final allocation result X' ;
 - 9: **end for**
 - 10: **return** X' ;
-

In other words, data analyst cannot increase its utility without decreasing others' utility. Mathematically speaking, if demand x'_i Pareto-dominates x_i , then $\Psi(x'_i) > \Psi(x_i)$.

Theorem 1. *The solution for the optimization problem in (10) is PE if and only if when $\beta > 0$ and $|\lambda| \geq |\frac{1-\beta}{\beta}|$.*

Proof. See Appendix A in the technical report [33]. \square

Definition 12 (Sharing Incentive (SI)). *A sharing incentive allocation x_i to data analyst i satisfies the sharing incentive property if each data analyst i gains more utility than the evenly fair share result x'_i , i.e.,*

$$U_i(x_i) \geq U_i(x'_i).$$

In other words, there is no incentive for data analysts to ask servers to partition privacy resources equally.

Theorem 2. *The solution to the original optimization problem in (13) satisfies SI (or not) as follows.*

- (a) SI is satisfied when $\beta > 1$ and $\lambda = \frac{\beta-1}{\beta}$.
- (b) SI is not satisfied when $0 < \beta < 1$ and $\lambda = \frac{\beta-1}{\beta}$.
- (c) SI is satisfied when $\lambda = 0$ for any β .
- (d) SI is not satisfied when $\lambda = \infty$ for any β .

Proof. See Appendix B in the technical report [33]. \square

Next, we would like to define envy-freeness:

Definition 13 (Envy-Freeness (EF)). *This property holds if and only if any data analyst i would not envy other data analyst's allocation. Mathematically speaking, data analyst i and data analyst j get the allocation x_i and allocation x_j respectively in our allocation scheme. Suppose that data analyst i envies data analyst j 's allocation x_j , it's impossible that data analyst i gets more utility with data analyst j 's allocation result x_j .*

Theorem 3. *The solution to the original optimization problem in (13) satisfies EF (or not) as follows.*

- (a) EF is satisfied when $\beta > 1$ and $\lambda = \frac{\beta-1}{\beta}$.
- (b) EF is not satisfied when $0 < \beta < 1$ and $\lambda = \frac{\beta-1}{\beta}$.
- (c) EF is satisfied when $\lambda = 0$ for any β .
- (d) EF is not satisfied when $\lambda = \infty$ for any β .

Proof. See Appendix C in the technical report [33]. \square

Definition 14 (Strong Strategy Proofness (SSP)). *It depicts that a data analyst can not increase both its utility and non-dominant share by lying about its training demands.*

Definition 15 (Weak Strategy Proofness (WSP)). *It depicts that when a data analyst lies about its training demands, it may increase its utility but decrease its non-dominant share.*

Both definitions of Strategy Proofness (SP) mean data analyst has no incentive to lie about training demands. Under SSP, there would be no benefit from lying. Under WSP, lying about training demand may increase its utility, but it may also reduce non-dominant shares. Unlike the traditional resource allocation where training demands must be strictly proportional to each resource, getting extra non-dominant privacy resources may also help data analyst to finish more pipelines or increase some pipeline's accuracy as the data analyst's training demand is composed of pipelines' training demands.

Theorem 4. *The solution to the original optimization problem in (13) satisfies SP (or not) as follows.*

- (a) WSP is satisfied when $\beta > 1$ and $\lambda = \frac{\beta-1}{\beta}$.
- (b) Neither WSP nor SSP is satisfied when $0 < \beta < 1$ and $\lambda = \frac{\beta-1}{\beta}$.
- (c) SSP is satisfied when $\lambda = 0$ for any β .
- (d) SSP is satisfied when $\lambda = \infty$ for any β .

Proof. See Appendix D in the technical report [33]. \square

C. Fairness-efficiency tradeoff

In this subsection, we theoretically analyze the relationship between efficiency and fairness by examining the monotonicity of these two metrics. For simplicity, we take α -fairness function to discuss the tradeoff. As mentioned, when $\lambda = |\frac{1-\beta}{\beta}|$, the original optimization problem can be degenerated into the famous α -fairness function in (12), where α can be considered as the fairness measurement [34] [35] [36]. The larger β ($\beta = \alpha$) is, the fairer the utility is. When $\beta = 0$, it means our utility function in (12) only focuses on efficiency. When β approaches ∞ , our utility function becomes the fairest.

Theorem 5 (Efficiency non-increasing property). *Efficiency is non-increasing as β increases if and only if when*

$$\sum_{i=1}^{M-K} \tau_i \det \overline{A}_i \geq 0. \quad (25)$$

Proof. Let R be the $M \times K$ matrix that has full row rank. $M - K$ is the dimension of the null space. Let z_i be the basis of the null space of R , where $i = 1, 2, \dots, M - K$. Collect all z_i to form the matrix $Z = [z_1, z_2, \dots, z_{M-K}]$. Then, we let D be the Hessian matrix of our utility function in (12) and let $\mathbf{b} = \frac{\partial U}{\partial x \partial \beta}$. According to [34], we can get that

$$\frac{\partial E}{\partial \beta} = \mathbf{1}^T Z (Z^T D Z)^{-1} Z^T \mathbf{b}. \quad (26)$$

Let $A = Z^T D Z$, we can further infer that the conditions for that efficiency are non-increasing as α increases. \square

We find efficiency is non-increasing with fairness as shown in Theorem 5. In other words, there exists a tradeoff between efficiency and fairness when the conditions in Theorem 5 are satisfied. Furthermore, we discuss two scenarios that the tradeoff between efficiency and fairness doesn't exist. More specifically, we discuss the scenarios in which the allocation result can be the most efficient and fairest at the same time. As α is the measure of fairness, the fairest allocation scheme is max-min fairness when α approaches ∞ . Thus, the equal-weighted dominant share is the fairest allocation result. To get the most efficient allocation result, it's essential to ensure as many resource constraints are tight as possible. We divide it into two scenarios: $K < M$ and $K = M$.

Theorem 6 (Fairest and most efficient allocation scheme (I)). *The equal-weighted dominant share scheme makes $K = M$ constraints tight if and only when*

$$\sum_{i=1}^M \frac{\gamma_i^{<k>}}{\mu_i} = C, \quad (27)$$

for any data analyst i who requests K resources, and C is a constant.

Proof. See Appendix E in the technical report [33]. \square

Theorem 7 (Fairest and most efficient allocation scheme (II)). *The equal-weighted dominant share scheme makes $K < M$ constraints tight if and only if when at least one data analyst i 's dominant share $\mu_i x_i = 0$, which means at least one data analyst is allocated no dominant share.*

Proof. See Appendix F in the technical report [33]. \square

In the first scenario, DPBalance outputs equal share for each data analyst when each data analyst has the same weighted dominant share. In the second scenario, DPBalance outputs zero share for data analysts with zero weighted dominant share. Therefore, DPBalance covers these two special scenarios when conditions in Theorem 6 or Theorem 7 are satisfied.

VI. EVALUATION

In this section, we evaluate DPBalance regarding its efficiency and fairness. We aim to answer the following questions:

Q1: How does DPBalance compare to other baseline scheduling methods?

Q2: How does fairness preference β affect dominant efficiency and dominant fairness?

Q3: Can we empirically validate the existence of the trade-off between dominant efficiency and dominant fairness when the condition of Theorem 5 is satisfied?

Metrics: We use the following four metrics to measure the performance of DPBalance and baselines.

- *Cumulative efficiency* is the sum of the dominant efficiency defined in (8) of all rounds.
- *Cumulative fairness* is the sum of the dominant fairness defined in (9) of all rounds.
- *Round efficiency* is the dominant efficiency defined in (8) of current training round.
- *Round fairness* is the dominant fairness defined in (9) of current training round.

Round efficiency and round fairness are designed to testify if the tradeoff for a single round exists.

Baselines: We compare DPBalance with three baseline scheduling algorithms, DPK [15], DPF [18] and First-Come-First-Serve (FCFS). DPK allocates resources to the pipelines with lowest weight-to-demand ratio. DPF allocates resources to pipelines with smallest dominant share. First-Come-First-Serve (FCFS) allocates resource to pipelines who arrive first.

Simulation Setup: For the resource side, we create 100 edge devices and each device's global privacy budget ϵ_g follows the Uniform distribution, namely $\epsilon_g \sim U(1.0, 1.5)$. For each device, two new blocks are created every 10 seconds. For the data analysts side, we assume 6 data analysts, each of whom with 25 pipelines, come every 10 seconds following the Poisson process. Depending on the amount of demanding privacy resources, we generate two types of pipelines, 75% mice and 25% elephant pipelines following the setting in DPF [18]. For mice pipeline, it demands data blocks with privacy demand $\epsilon \sim U(0.005, 0.015)$. For elephant pipeline, it demands data blocks with privacy demand $\epsilon \sim U(0.095, 0.105)$. For each pipeline, it demands the latest 10 data blocks with a probability of 0.25 or the latest 1 data block with a probability of 0.75. After being generated, all pipelines are shuffled and put into different data analysts, each of whom consists of 25 pipelines. For each data analyst, it either demands 0.2 of all devices or all the devices with probability 0.5 and 0.5 respectively.

A. Comparison with baselines (Q1)

We compare DPBalance with other baseline scheduling methods. We show the cumulative round efficiency and cumulative round fairness in Fig. 4 and Fig. 5 respectively. We choose three different fairness preference β to represent efficiency-preferred ($\beta = 0.5$), efficiency-fairness-unbiased ($\beta = 2.2$), fairness-preferred ($\beta = 5.0$) situation. We take the last round (round 10) as an example. For efficiency,

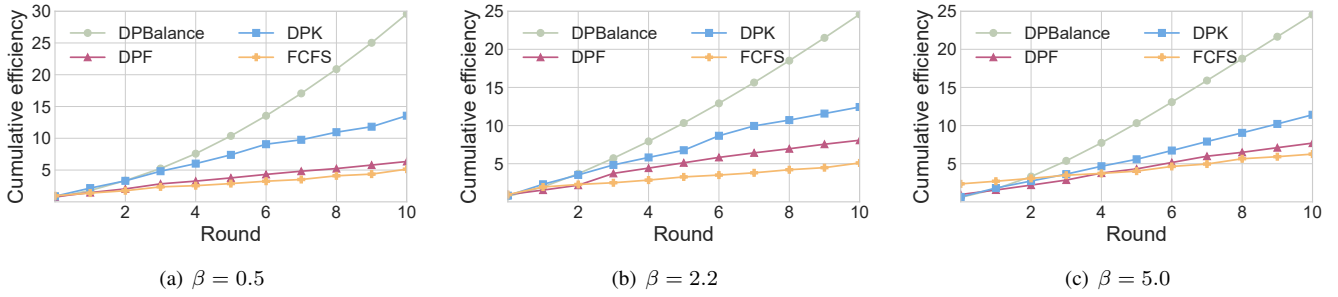


Fig. 4. Comparison of cumulative efficiency under different fairness preference settings

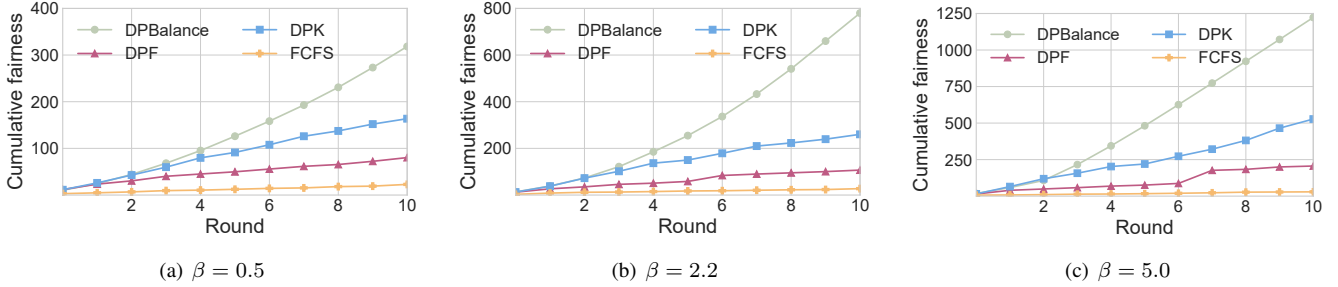


Fig. 5. Comparison of cumulative fairness under different fairness preference settings

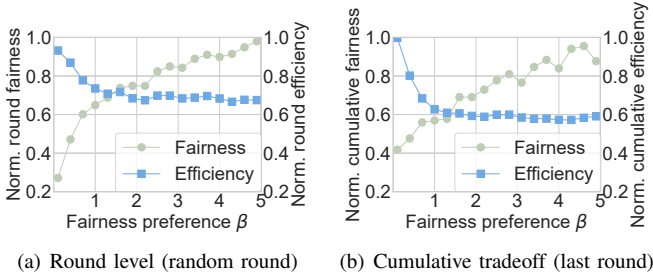


Fig. 6. Validation of efficiency-fairness tradeoff in two rounds

DPBalance outperforms than other three baselines by improving $1.44\times \sim 3.49\times$, $1.41\times \sim 3.04\times$, $1.67\times \sim 2.28\times$ on average under the efficiency-preferred, efficiency-fairness-unbiased and fairness-preferred situation respectively. For fairness, DPBalance outperforms the other three baselines by improving $1.37\times \sim 9.66\times$, $1.77\times \sim 15.58\times$, $1.82\times \sim 24.32\times$ on average under the efficiency-preferred, efficiency-fairness-unbiased and fairness-preferred situation respectively. Therefore, DPBalance outperforms the other three scheduling baselines on cumulative round efficiency and cumulative round fairness under different fairness preferences.

B. Impact of fairness preference β (Q2)

We examine the performance of DPBalance under different fairness preference β increases in Fig. 6. As fairness preference β grows from 0 to 5, which means allocation becomes more fair-preferred, round efficiency and cumulative efficiency decrease 48% and 38% respectively. As fairness preference β grows from 0 to 5, round fairness and cumulative fairness increase 257% and 119% respectively. Therefore, β can serve

as a controlling knob for the system manager to adjust fairness and efficiency.

C. Validation of tradeoff (Q3)

In addition to theoretical proof, we also validate the tradeoff between fairness and efficiency using the result of a randomly selected round. The round efficiency and round fairness are normalized for a better comparison. For a randomly selected round in Fig. 6(a), round efficiency decreases and round fairness increases as fairness preference β increases from 0 to 5. For the last round in Fig. 6(b), similarly, cumulative fairness increases, and cumulative efficiency decreases as fairness preference β increases. Therefore, the tradeoff between dominant efficiency and dominant fairness does exist when conditions in Theorem 5 are satisfied, which verifies the proof of the tradeoff in Section V-C.

VII. CONCLUSION

FLaaS is an emerging cloud-based service that facilitates FL-based model training over crowdsourced data owners. Due to the intrinsic properties of DP, privacy budget set by data owners have to be carefully allocated to different data analysts to improve system efficiency and data analyst fairness. Existing privacy budget scheduling mechanisms consider either efficiency or fairness only. This paper completes the missing piece by proposing a new mechanism, DPBalance, that jointly optimizes efficiency and fairness. We prove theoretically that DPBalance guarantees four key economic properties and there exists an fairness-efficiency tradeoff in practical situations. Extensive experiments further demonstrate the superiority of DPBalance in efficiency and fairness performance.

REFERENCES

- [1] European Commission, “General data protection regulation,” May 2018. [Online]. Available: <https://gdpr-info.eu/>
- [2] M. Chen, R. Mathews, T. Ouyang, and F. Beaufays, “Federated learning of out-of-vocabulary words,” *arXiv*, Apr. 2019.
- [3] S. Ramaswamy, R. Mathews, K. Rao, and F. Beaufays, “Federated learning for emoji prediction in a mobile keyboard,” *arXiv*, Jun. 2019.
- [4] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, “Applied federated learning: Improving google keyboard query suggestions,” *arXiv*, Jan. 2018.
- [5] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, “Federated learning for keyword spotting,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6341–6345.
- [6] W. De Brouwer, “The federated future is ready for shipping,” Mar. 2019. [Online]. Available: https://medium.com/@_doc_ai/the-federated-future-is-ready-for-shipping-d17ff40f43e3
- [7] C. He, S. Li, J. So, M. Zhang, H. Wang, X. Wang, P. Vepakomma, A. Singh, H. Qiu, L. Shen, P. Zhao, Y. Kang, Y. Liu, R. Raskar, Q. Yang, M. Annavaram, and S. Avestimehr, “Fedml: A research library and benchmark for federated machine learning,” *arXiv*, Jun. 2020.
- [8] The FATE Authors, “Federated ai technology enabler,” 2019. [Online]. Available: <https://www.fedai.org/>
- [9] N. Kourtellis, K. Katevas, and D. Perino, “Flaas: Federated learning as a service,” in *Proc. Workshop Distrib. Mach. Learn.*, Dec. 2020, pp. 7–13.
- [10] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proc. ACM SIGSAC conf. comput. commun. secur.*, Oct. 2016, pp. 308–318.
- [11] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor, “Federated learning with differential privacy: Algorithms and performance analysis,” *IEEE Trans. Inf. Forensics Secur.*, vol. 15, no. 1, pp. 3454–3469, Jun. 2020.
- [12] L. Sun, J. Qian, and X. Chen, “LDP-FL: practical private aggregation in federated learning with local differential privacy,” in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 19–27, Aug. 2021, pp. 1571–1578.
- [13] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, “Dominant resource fairness: Fair allocation of multiple resource types,” in *Proc. USENIX Symp. Netw. Syst. Des. Implementation (NSDI)*, Mar. 2011.
- [14] W. Li, L. Xiang, B. Guo, Z. Li, and X. Wang, “Dplanner: A privacy budgeting system for utility,” *IEEE Trans. Inf. Forensics Secur.*, vol. 18, no. 1, pp. 1196–1210, Feb. 2023.
- [15] P. Tholoniati, K. Kostopoulou, M. Chowdhury, A. Cidon, R. Geambasu, M. Lécuyer, and J. Yang, “Packing privacy budget efficiently,” *arXiv*, Jan. 2022.
- [16] J. Yuan, S. Wang, S. Wang, Y. Li, X. Ma, A. Zhou, and M. Xu, “Privacy as a resource in differentially private federated learning,” in *Proc. IEEE Int. Conf. Comp. Commun. (INFOCOM)*, May 2023.
- [17] N. Küchler, E. Opel, H. Lycklama, A. Viand, and A. Hithnawi, “Cohere: Privacy management in large scale systems,” *arXiv*, Jan. 2023.
- [18] T. Luo, M. Pan, P. Tholoniati, A. Cidon, R. Geambasu, and M. Lécuyer, “Privacy budget scheduling,” in *Proc. USENIX Symp. Oper. Syst. Des. Implementation (OSDI)*, Aug. 2021, pp. 55–74.
- [19] C. T. Dinh, N. H. Tran, M. N. H. Nguyen, C. S. Hong, W. Bao, A. Y. Zomaya, and V. Gramoli, “Federated learning over wireless networks: Convergence analysis and resource allocation,” *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 398–409, Oct. 2021.
- [20] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, “Joint device scheduling and resource allocation for latency constrained wireless federated learning,” *IEEE Trans. Wirel. Commun.*, vol. 20, no. 1, pp. 453–467, Dec. 2021.
- [21] T. Li, M. Sanjabi, A. Beirami, and V. Smith, “Fair resource allocation in federated learning,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, Apr. 2020, pp. 1–27.
- [22] W. Y. B. Lim, J. S. Ng, Z. Xiong, J. Jin, Y. Zhang, D. Niyato, C. Leung, and C. Miao, “Decentralized edge intelligence: A dynamic resource allocation framework for hierarchical federated learning,” *IEEE Trans. Parallel Distributed Syst.*, vol. 33, no. 3, pp. 536–550, Sep. 2022.
- [23] A. M. Girgis, D. Data, and S. N. Diggavi, “Renyi differential privacy of the subsampled shuffle model in distributed learning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021, pp. 29 181–29 192.
- [24] I. Mironov, “Rényi differential privacy,” in *Proc. IEEE Comput. Secur. Found. Symp. (CSF)*, Aug. 2017, pp. 263–275.
- [25] M. Uchida and J. Kurose, “An information-theoretic characterization of weighted alpha-proportional fairness,” in *Proc. IEEE Int. Conf. Comp. Commun. (INFOCOM)*, Apr. 2009, pp. 1053–1061.
- [26] T. Lan, D. T. H. Kao, M. Chiang, and A. Sabharwal, “An axiomatic theory of fairness in network resource allocation,” in *Proc. IEEE Int. Conf. Comp. Commun. (INFOCOM)*, Mar. 2010, pp. 1343–1351.
- [27] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang, “Multiresource allocation: Fairness–efficiency tradeoffs in a unifying framework,” *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1785–1798, May 2013.
- [28] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in *Proc. IEEE Symp. Secur. Privacy (S&P)*, Jun. 2017, pp. 3–18.
- [29] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, “The secret sharer: Evaluating and testing unintended memorization in neural networks,” in *Proc. USENIX Secur. Symp. (USENIX Security)*, Aug. 2019, pp. 267–284.
- [30] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury, “Oort: Efficient federated learning via guided participant selection,” in *Proc. USENIX Symp. Oper. Syst. Des. Imple., (OSDI)*, Jul. 2021, pp. 19–35.
- [31] E. Altman, K. Avrachenkov, and A. Garnaev, “Generalized a-fair resource allocation in wireless networks,” in *Proc. Conf. Decis. Control (CDC)*, Dec. 2008, pp. 2414–2419.
- [32] Gurobi Optimization, LLC, “Gurobi optimizer reference manual,” 2022. [Online]. Available: <https://www.gurobi.com/documentation/current/refman/index.html>
- [33] Y. Liu, Z. Wang, Y. Zhu, and C. Chen, “Dpbalance: Efficient and fair privacy budget scheduling for federated learning as a service (technical report),” Dec. 2023. [Online]. Available: <https://sites.ji.sjtu.edu.cn/yifei-zhu/wp-content/uploads/sites/19/2023/12/INFOCOM24DPBalanceTechReport.pdf>
- [34] A. Tang, J. Wang, and S. H. Low, “Counter-intuitive throughput behaviors in networks under end-to-end control,” *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 355–368, Jul. 2006.
- [35] A. Rényi, “On measures of entropy and information,” in *Proc. Berkeley Symp. Math. Statist. Probability, Volume 1: Contributions to the Theory of Statistics*, vol. 4, 1961, pp. 547–562.
- [36] C. F. Menezes and D. L. Hanson, “On the theory of risk aversion,” *Int. Econ. Rev.*, vol. 11, no. 3, pp. 481–487, 1970.