

# Device Sampling and Resource Optimization for Federated Learning in Cooperative Edge Networks

Su Wang<sup>1</sup>, Roberto Morabito<sup>2</sup>, *Member, IEEE*, Seyyedali Hosseinalipour<sup>3</sup>, *Member, IEEE*, Mung Chiang, *Fellow, IEEE*, and Christopher G. Brinton<sup>1</sup>, *Senior Member, IEEE*

**Abstract**—The conventional federated learning (FedL) architecture distributes machine learning (ML) across worker devices by having them train local models that are periodically aggregated by a server. FedL ignores two important characteristics of contemporary wireless networks, however: (i) the network may contain heterogeneous communication/computation resources, and (ii) there may be significant overlaps in devices' local data distributions. In this work, we develop a novel optimization methodology that jointly accounts for these factors via intelligent device sampling complemented by device-to-device (D2D) offloading. Our optimization methodology aims to select the best combination of sampled nodes and data offloading configuration to maximize FedL training accuracy while minimizing data processing and D2D communication resource consumption subject to realistic constraints on the network topology and device capabilities. Theoretical analysis of the D2D offloading subproblem leads to new FedL convergence bounds and an efficient sequential convex optimizer. Using these results, we develop a sampling methodology based on graph convolutional networks (GCNs) which learns the relationship between network attributes, sampled nodes, and D2D data offloading to maximize FedL accuracy. Through evaluation on popular datasets and real-world network measurements from our edge testbed, we find that our methodology outperforms popular device sampling methodologies from literature in terms of ML model performance, data processing overhead, and energy consumption.

**Index Terms**—Federated learning (FedL), distributed learning, network optimization.

## I. INTRODUCTION

THE proliferation of smartphones, unmanned aerial vehicles (UAVs), and other devices comprising the Internet of

Manuscript received 16 October 2023; revised 12 May 2024; accepted 19 June 2024; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor R. Pedarsani. The work of Christopher G. Brinton was supported by the National Science Foundation (NSF) under Grant CNS-2146171 and Grant CPS-2313109 and in part by the Office of Naval Research (ONR) under Grant N000142212305 and Grant N0001423C1016. This work is an extension of our prior work titled "Device sampling for heterogeneous federated learning: Theory, algorithms, and implementation" [DOI: 10.1109/INFOCOM42981.2021.9488906] published in the Proceedings of IEEE INFOCOM, 2021. (*Corresponding author: Su Wang.*)

Su Wang was with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA. He is now with the School of Engineering and Applied Science, Princeton University, Princeton, NJ 08540 USA (e-mail: hw5731@princeton.edu).

Roberto Morabito is with the Communication Systems Department, Eurecom, 06410 Biot, France (e-mail: roberto.morabito@eurecom.fr).

Seyyedali Hosseinalipour is with the Department of Electrical Engineering, University at Buffalo (SUNY), Buffalo, NY 14260 USA (e-mail: alipour@buffalo.edu).

Mung Chiang and Christopher G. Brinton are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: chiang@purdue.edu; cgb@purdue.edu).

Digital Object Identifier 10.1109/TNET.2024.3423673

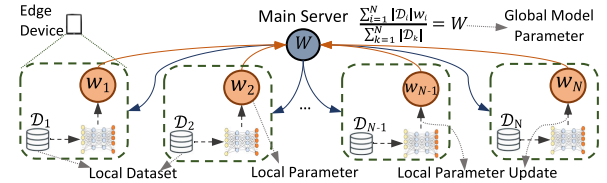


Fig. 1. Architecture of conventional federated learning (FedL).

Things (IoT) is causing an exponential rise in data generation and large demands for machine learning (ML) at the edge [2], [3]. For example, sensor and camera modules on self-driving cars produce up to 1.4 terabytes of data per hour [4] which enable the training of ML models for intelligent navigation and autonomous driving [5]. Unfortunately, traditional ML techniques where model training is conducted at a centralized server are not applicable to such distributed environments as the data often cannot be transferred to a server. There are two challenges in particular: (i) given the large volumes of data at the network edge, data transfer to a server imposes long delays and overloads the network infrastructure [6], and (ii) in certain applications (such as healthcare), some users are unwilling to share their data owing to privacy concerns [7].

Federated learning (FedL) has become a popular distributed ML technique aiming to overcome these challenges [8], [9]. Under FedL, devices train models on their local datasets, typically by means of gradient descent, and a server periodically aggregates the parameters of local models to form a global model. These global model parameters are then transferred back to the devices for the next round of local updates, as depicted in Fig. 1. In standard FedL methodologies, each device processes its own collected data and operates independently within an aggregation period [10], [11]. This will, however, become problematic in terms of upstream device communication and local device processing requirements as implementations scale to networks consisting of millions of heterogeneous wireless devices [12], [13], [14].

At the same time, device-to-device (D2D) communications in cooperative 5G/IoT edge networks can enable *consensual* local offloading of data processing from resource hungry to resource rich devices [15], [16]. Additionally, we can expect that for particular applications, the datasets collected across devices will contain varying degrees of similarity, e.g., images gathered by UAVs surveying the same area [17], [18]. Processing similar data distributions at multiple devices adds unnecessary computational and communication overheads to FedL, thus leading to an opportunity for efficiency improvement.

Motivated by this, we develop a novel methodology for *smart device sampling with data offloading* in FedL. Specifically, we formulate a joint sampling and data offloading optimization problem where devices expected to maximize contribution to model training are sampled for training participation, while devices that are not selected may transfer data to devices that are. This data offloading is motivated by paradigms such as *fog learning* [13], [19], in which data offloading only occurs among trusted devices; devices that have privacy concerns are exempt from data offloading. Furthermore, the data offloading is performed according to estimated data dissimilarities between nodes, which are updated as transfers occur. We show that our methodology yields comparable performance to conventional FedL, and superior performance relative to baselines from literature while significantly reducing network resource utilization.

### A. Related Work

**Resource efficiency in federated learning** Since conventional FedL in large-scale networks can incur heavy communication and computation resource use, many works have investigated avenues to reduce the resource burden of edge devices in FedL. In particular, a few popular and recent approaches have focused on efficient encoding designs to reduce parameter transmission sizes [20], [21], optimizing the frequency of global aggregations [22], [23], [24], and device sampling [25], [26], [27]. Our work falls into the device sampling category, in which methodologies rely on intelligent selection of devices for the global aggregation stage of FedL to improve resource efficiency. In this regard, most works have assumed a static or uniform device selection strategy, e.g., [9], [23], [25], [28], [29], [30], and [31], where the main server chooses a subset of devices either uniformly at random or according to a pre-determined sampling distribution.

While there are many different approaches for device sampling for FedL such as sampling based on wireless channel characteristics in cellular networks (e.g., [32], [33], [34], [35]), we focus on developing a sampling technique that adapts to the heterogeneity of device resources and the similarities/dissimilarities of data distributions across large-scale contemporary wireless edge networks. When investigating literature in this line of device sampling for FedL, popular works have attempted device sampling based on each device's instantaneous contributions to global updates [36], data proportional sampling based on devices' local dataset sizes [26], [37], and sampling devices based on local ML training loss [27]. Specifically, when compared to the literature on device sampling by each device's instantaneous contributions to the global updates [27], [36], our proposed methodology introduces a novel perspective based on device data similarities, and furthermore exploits the proliferation of D2D cooperation at the wireless edge [16], [38], [39], [40] to diversify each selected device's local data via D2D offloading. Our work thus considers the problems of sampling and D2D offloading for FedL jointly, and leads to new analytical convergence bounds and algorithms used by implementations.

**Towards further cooperation in federated learning.** Most works in advancing conventional FedL [9], [29], [41], [42], [43] limit network cooperation to the global aggregation stages. While this is effective in situations requiring absolute data privacy, many practical large-scale and mobile edge networks are not limited to such an extent [16], [44], [45].

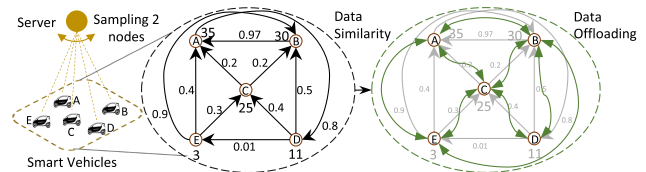


Fig. 2. A motivating example of a wireless network composed of 5 connected vehicles and an edge server. The server can only sample two vehicles to participate in FedL training.

Especially in large-scale and mobile edge networks, devices can be heterogeneous with respect to their data privacy demands [46], [47], [48]. While some devices may require absolute data privacy, many devices may be willing to share non-sensitive data or share all of their data to a device that they trust. Such is the case in recent literature on connected autonomous driving [15], [49], which rely on inter-device cooperation for normal functionality and is now attempting to integrate the benefits of modern ML. With this in mind, works such as [46], [50], and [51] have begun expanding the scope of cooperation in FedL. Specifically, [46] investigates FedL performance in the presence of D2D model offloading among edge devices, [50] proposes device to edge server data offloading so that edge servers can contribute to FedL training, and [51] leverages D2D cooperation in edge networks to develop a modified version of SGD for FedL specifically. Our work contributes to existing literature in cooperative FedL by investigating how D2D data cooperation can lead to optimized device sampling with respect to ML training efficacy and total network resource consumption. Furthermore, our work assumes that devices are heterogeneous with respect to their privacy demands (i.e., they trust some edge devices but not others), resulting in some D2D links enabled and others disabled.

### B. Motivating Toy Example

Consider Fig. 2, wherein five heterogeneous connected vehicles communicate with an edge server to train an object detection model. Due to limited bandwidth, the server can only exploit 2 out of the 5 vehicles to conduct FedL training, but needs to train a model representative of the entire dataset within this network. The computational capability of each vehicle, i.e., the number of processed datapoints in one aggregation period, is shown next to the node representing the vehicle, and the edge weights in the data similarity graph capture the similarity between the local data of the vehicles. Rather than using statistical distance metrics [52], which are hard to compute in this distributed scenario, the data similarities could be estimated by commute routes and geographical proximity [53]. Also, in D2D-enabled environments, nodes can exchange small data samples with trusted neighbors to calculate similarities locally and report them to the server.

In Fig. 2, if the server samples the vehicles with the highest computational capabilities, i.e., A and B, the sampling is inefficient due to the high data similarity between them. Conversely, if it samples those with the lowest similarity, i.e., D and E, the local models will be based on low computational capabilities, which will often result in a low accuracy (and could be catastrophic in this vehicular scenario). Optimal sampling of the vehicles considering both data similarities and computational capabilities is thus critical to the operation of FedL.

We take this one step further to consider how D2D offloading can lead to augmented local distributions of sampled

vehicles. The node sampling must consider the neighborhoods of different vehicles and the capability of data offloading in those neighborhoods. For example, D2D is cheaper in terms of resource utilization among vehicles that are close in proximity. The feasible offloading topology shown in Fig. 2 is represented by the data offloading graph. Given  $C$ 's high processing capability and data dissimilarity with neighboring vehicles  $E$  and  $D$ , sampling  $C$  in a D2D-optimized solution combined with data offloading from  $E$  and  $D$  to  $C$  can yield a composite of all three vehicles' distributions. The purpose of this paper is to model these relationships for general edge/fog networks and optimize the resulting sampling and offloading configurations.

### C. Outline and Summary of Contributions

- We formulate the joint sampling, D2D offloading, and communication and computation resource optimization problem for maximizing FedL model accuracy subject to realistic network constraints (Sec. II).
- We develop an approach to D2D data similarity estimation by comparing the dataset centroids across network devices (Sec. II). This method enables the server to perform the optimization without harvesting actual device data.
- Our theoretical analysis of the offloading subproblem for a fixed sampling strategy yields a new upper bound on the convergence of FedL under an arbitrary data sampling strategy (Sec. III). Using this bound, we derive an efficient sequential convex optimizer for the offloading strategy.
- We propose a novel ML-based methodology that learns the desired combination of sampling and resulting offloading scheme into model features and learn a mapping to the sampling strategy that maximizes expected FedL accuracy.
- We evaluate our methodology through experiments with network parameters obtained from our testbed of wireless IoT devices (Sec. V). Our methodology yields better ML accuracies and less resource consumption than recent device sampling baselines from literature.
- We experimentally characterize our formulation by investigating its behavior with respect to various key optimization variables while varying network and sampling device set sizes (Sec. V).

## II. SYSTEM AND OPTIMIZATION MODEL

In this section, we formulate the joint sampling and offloading optimization (Sec. II-D). We first introduce our edge device (Sec. II-A), network topology and data similarity (Sec. II-B), and ML (Sec. II-C) system models.

### A. Edge Device Model

We consider a set of devices  $\mathcal{N} = \{1, \dots, N\}$  connected to a server, and time span  $t = 0, \dots, T$  for model training. Each device  $i \in \mathcal{N}$  possesses a data processing capacity  $P_i(t) \geq 0$ , which limits the number of datapoints it can process for training at time  $t$ , and a unit data processing cost  $p_i(t) \geq 0$ . Intuitively,  $p_i(t)$ ,  $P_i(t)$  are related to the total CPU cycles, memory (RAM), and power available at device  $i$  [54]. These factors are heterogeneous and time-varying, e.g., as battery power fluctuates and as each device becomes occupied with other tasks. Additionally, for each  $i \in \mathcal{N}$ , we define  $\Psi_i(t) > 0$  as the data transmit budget, and  $\psi_{i,j}(t) > 0$  as the unit data transmission cost from device  $i$  to device

$j$ . Intuitively,  $\psi_{i,j}(t)$ ,  $\Psi_i(t)$  are dependent on factors such as the available bandwidth and distance. For example, devices that are closer in proximity would be expected to have lower  $\psi_{i,j}(t)$ . Our specific derivations of these data processing and communication variables are based on empirical measurements from our IoT testbed, which we describe in detail in Sec. V-A.

Due to resource constraints, the server selects a set  $\mathcal{S} \subseteq \mathcal{N}$  of devices to participate in FedL training. Some devices  $i \in \hat{\mathcal{S}} \triangleq \mathcal{N} \setminus \mathcal{S}$  may be stragglers, i.e., possessing insufficient  $P_i(t)$  to participate in training, but nonetheless gather data. Different from most works, our methodology will seek to leverage the datasets captured by nodes in the unsampled set  $\hat{\mathcal{S}}$  via local D2D communications with nodes in the sampling set  $\mathcal{S}$ .

We denote the dataset at device  $i$  for the specific ML application by  $\mathcal{D}_i(t)$ .  $\mathcal{D}_i(0)$  is the initial data at  $i$ , which evolves as offloading takes place. Henceforth, we use calligraphic font (e.g.,  $\mathcal{D}_i(t)$ ) to denote a set, and non-calligraphic (e.g.,  $D_i(t) = |\mathcal{D}_i(t)|$ ) to denote its cardinality. Each data point  $d \in \mathcal{D}_i(t)$  is represented as  $d = (\mathbf{x}_d, y_d)$ , where  $\mathbf{x}_d \in \mathbb{R}^M$  is a feature vector of  $M$  features, and  $y_d \in \mathbb{R}$  is the target label.

### B. Network Topology and Data Similarity Model

We consider a time-varying network graph  $G = (\mathcal{N}, \mathcal{E}(t))$ , among the set of nodes  $\mathcal{N}$  to represent the available D2D topology. Here,  $\mathcal{E}(t)$  denotes the set of edges or connections between the nodes, where  $(i, j) \in \mathcal{E}(t)$  if node  $i$  is able/willing to transfer data in D2D mode to node  $j$  at time  $t$ . This willingness to transfer data can depend on the trust between the devices as an example, and whether the devices are D2D-enabled. For instance, smart home peripherals can likely transfer data to their owner's smartphone, while different smartphones in an airport may be unwilling to share data. We capture these potential D2D relationships using the adjacency matrix  $\mathbf{A}(t) = [A_{i,j}(t)]_{1 \leq i,j \leq N}$ , where  $A_{i,j}(t) = 1$  if  $(i, j) \in \mathcal{E}(t)$ , and  $A_{i,j}(t) = 0$  otherwise.

We define  $\Phi_{i,j}(t) \in [0, 1]$  as the fraction of node  $i$ 's data offloaded to node  $j$  at time  $t$ . To optimize this, we are interested in the similarity among local datasets. We define the similarity matrix  $\boldsymbol{\lambda}(t) \triangleq [\lambda_{i,j}(t)]_{1 \leq i,j \leq N}$  among the nodes at time  $t$ , where  $0 \leq \lambda_{i,j}(t) \leq 1$ . Lower values of  $\lambda_{i,j}(t)$  imply a higher dataset similarity between nodes  $i$  and  $j$ , and thus less offloading benefit. In practice, neither the server nor the devices have exact knowledge of the local data distributions.

To this end, we consider an interpretation of similarity based on comparing data centroids across network devices. Each centroid is an average of the data contained within a data cluster. In our setting, each device  $i$  has  $\mathcal{C}_i$  clusters, the contents of which are determined based on K-means [55] clustering which is commonly used in existing literature [56], [57] to estimate dataset similarities. Each cluster  $c \in \mathcal{C}_i$  contains the data  $\mathcal{D}_i^c(t)$  so that  $\mathcal{D}_i(t) = \cup_{c \in \mathcal{C}_i} \mathcal{D}_i^c(t)$ . Thus, the  $c$ -th centroid  $\mu_i^c(t)$  at device  $i$  is computed as:

$$\mu_i^c(t) = \frac{1}{|\mathcal{D}_i^c(t)|} \sum_{d \in \mathcal{D}_i^c(t)} \mathbf{x}_d. \quad (1)$$

For each D2D link, we determine the similarity value  $\lambda_{i,j}(t)$  based on the minimum total centroid difference for all clusters between devices  $i$  and  $j$ .<sup>1</sup> Specifically, we first need to match

<sup>1</sup>The goal of the data similarity estimation process is to gauge the data offloading value. If two device pairs have similar minimum centroid difference, then our formulation estimates that these device pairs have similar benefit from data offloading.



each cluster  $b \in \mathcal{C}_j$  to some cluster  $c \in \mathcal{C}_i$  with the smallest centroid difference between them, i.e., we want

$$c = \arg \min_{c \in \mathcal{C}_i} |\sigma_{i,j}^{c,b}(t)|, \quad (2)$$

where  $\sigma_{i,j}^{c,b}(t) = \mu_i^c(t) - \mu_j^b(t)$  and  $\sigma_{i,j}^{c,b}(t) = 0$  implies that clusters  $b$  and  $c$  are essentially identical. However, (2) enables a single cluster  $c$  to be the matching cluster for multiple  $b \in \mathcal{C}_j$ , which fails to fully characterize the D2D similarity between all of the data at devices  $i$  and  $j$  (as its possible for  $\mathcal{D}_i^c(t)$  to be the only subset of data that is characterized at this link). As a result, when a cluster  $c \in \mathcal{C}_i$  has the smallest  $\sigma_{i,j}^{c,b}(t)$  for multiple clusters  $b \in \mathcal{C}_j$ , we choose to link  $c$  to the cluster  $b$  with the smallest  $\sigma_{i,j}^{c,b}(t)$ , and subsequently use the heuristic  $\tilde{\mu}_i^c(t) \rightarrow \infty$  to replace  $\mu_i^c(t)$  in (2) for the remaining similarity computations for the specific link between devices  $i$  and  $j$ . In this way, each cluster  $c \in \mathcal{C}_i$  links to at most one cluster  $b \in \mathcal{C}_j$ . Then the data similarity  $\lambda_{i,j}(t)$  between devices  $i$  and  $j$  can be computed as

$$\lambda_{i,j}(t) = \sum_{b \in \mathcal{C}_j} \min_{c \in \mathcal{C}_i} \frac{|\sigma_{i,j}^{c,b}(t)|}{|\mathcal{C}_i|}. \quad (3)$$

For  $t > 0$ ,  $\lambda_{i,j}(t)$  depends on the changes in  $\sigma_{i,j}^{c,b}(t)$ , which in turn depends on the data offloading ratio  $\Phi_{i,j}^{c,b}(t-1)$  from the previous time step. Using these cluster-specific offloading ratios, we can compute the aggregate data offloading ratio from device  $i$  to  $j$  as

$$\Phi_{i,j}(t) = \sum_{b \in \mathcal{C}_j} \sum_{c \in \mathcal{C}_i} \Phi_{i,j}^{c,b}(t). \quad (4)$$

Because devices lack exact information about each other's datasets, the selection of specific data for offloading from device  $i$  to device  $j$  also depends on the data similarity  $\lambda_{i,j}(t)$  between them, and we explain the exact mechanics during the data offloading process in Sec. II-D. To capture both the node connectivity and data similarity jointly for offloading, we also define the connectivity-similarity matrix  $\Lambda(t) = [\Lambda_{i,j}(t)]$ , and  $\Lambda(t) \triangleq \mathbf{\Lambda}(t) \circ \mathbf{A}(t)$ , where  $\circ$  represents the Hadamard product.

This above process to determine the data similarities involves device-to-server transmissions, as devices will transmit their calculated centroids to the server for further processing such as cluster matching. These computations are performed prior to the start of our optimization process, outlined in Sec. II-D, and involve much smaller communication overhead than ML model transmission during the global aggregations.

### C. Distributed Machine Learning Model

The learning objective of FedL is to train a global ML model, parameterized by a vector  $\mathbf{w} \in \mathbb{R}^p$  (e.g.,  $p$  weights in a neural network), by training at edge devices. Formally, for  $t \in \{0, \dots, T\}$ , each sampled device  $i \in \mathcal{S}$  is concerned with updating its model parameter vector  $\mathbf{w}_i(t)$  on its local dataset  $\mathcal{D}_i(t)$ . The local loss at device  $i \in \mathcal{N}$  is defined as

$$F(\mathbf{w}_i(t)|\mathcal{D}_i(t)) = \frac{\sum_{d \in \mathcal{D}_i(t)} f(\mathbf{w}_i(t), \mathbf{x}_d, y_d)}{D_i(t)}, \quad (5)$$

where  $f(\mathbf{w}_i(t), x_d, y_d)$  denotes the corresponding loss (e.g., squared error for a regression problem, cross-entropy for a classification problem [23]) of each datapoint  $d \in \mathcal{D}_i$ . Each device minimizes its local loss sequentially via gradient descent:

$$\mathbf{w}_i(t) = \mathbf{w}_i(t-1) - \eta \nabla F(\mathbf{w}_i(t-1)|\mathcal{D}_i(t)), \quad (6)$$

where  $\eta > 0$  is the step size and  $\nabla F(\mathbf{w}_i(t-1)|\mathcal{D}_i(t))$  is the average gradient over the local dataset  $\mathcal{D}_i(t)$ . With periodicity  $\tau$ , the server performs a weighted average of  $\mathbf{w}_i(k\tau)$ ,  $i \in \mathcal{S}$ :

$$\mathbf{w}_S(k\tau) = \frac{\sum_{i \in \mathcal{S}} \Delta_i(k\tau) \mathbf{w}_i(k\tau)}{\sum_{i \in \mathcal{S}} \Delta_i(k\tau)}, \quad (7)$$

where  $k$  denotes the  $k$ -th aggregation,  $k \in \{1, \dots, K\}$ ,  $K = \lceil T/\tau \rceil$ , and  $\Delta_i(k\tau) \triangleq \sum_{t=(k-1)\tau+1}^{k\tau} D_i(t)$  denotes the total data located at node  $i$  between  $k-1$  and  $k$ . The server then synchronizes the sampled devices:  $\mathbf{w}_i(k\tau) \leftarrow \mathbf{w}_S(k\tau)$ ,  $\forall i \in \mathcal{S}$ .

Since we are concerned with the performance of the global parameter  $\mathbf{w}_S$ , we define  $\mathbf{w}_S(t)$  as the weighted average of  $\mathbf{w}_i(t)$  as in (7) for each time  $t$ , though it is only computed at the server when  $t = k\tau$ . The global loss that we seek to optimize considers the loss over all the datapoints in the network:

$$F(\mathbf{w}_S(t)|\mathcal{D}_N(t)) = \frac{\sum_{i \in \mathcal{N}} D_i(t) F(\mathbf{w}_S(t)|\mathcal{D}_i(t))}{D_N(t)}, \quad (8)$$

where  $\mathcal{D}_N(t)$  denotes the multiset of the datasets of all the devices at time  $t$ , and  $D_N(t) \triangleq \sum_{i \in \mathcal{N}} D_i(t)$ .

### D. Joint Sampling and Offloading Optimization

The node selection procedure should consider device characteristics mentioned in Sec. II-A and the data similarity among the nodes represented by the connectivity-similarity matrix  $\Lambda(t)$ . Thus, the goal of our optimization is to select (i) the subset of devices  $\mathcal{S}^*$  to sample from a total budget of  $S$  and (ii) the offloading ratios  $\Phi_{i,j}^{c,b}(t)$  among the devices to jointly minimize the loss associated with  $\mathbf{w}_S(t)$  and the associated energy resource consumption (in terms of data processing energy and data offloading energy use). We consider a time average for the objective, as devices may rely on intermediate aggregations for real-time inferences. For the variables, we define the binary vector  $\mathbf{x} \triangleq (x_1, \dots, x_N)$  to represent device sampling status, i.e., if  $i \in \mathcal{S}$  then  $x_i = 1$ , otherwise  $x_i = 0$ , and matrix  $\Phi(t) \triangleq [\Phi_{i,j}(t)]_{1 \leq i,j \leq N}$  to represent the offloading ratios at time  $t$ . The resulting optimization problem  $\mathcal{P}$  is as follows:

$$(\mathcal{P}) : \underset{\mathbf{x}, \{\Phi(t)\}_{t=1}^T}{\text{minimize}} \quad \frac{1}{T} \sum_{t=1}^T \left( \underbrace{\alpha F(\mathbf{w}_S(t)|\mathcal{D}_N(t))}_{(a)} + \underbrace{\beta \sum_{i \in \mathcal{S}} E_i^P(t)}_{(b)} + \underbrace{\gamma \sum_{k \in \mathcal{N}} E_k^{Tx}(t)}_{(c)} \right) \quad (9)$$

subject to

$$D_i^c(t) = D_i^c(t-1) + R_i^c(t), \quad i \in \mathcal{N}, \quad (10)$$

$$R_i^c(t) = \sum_{k \in \mathcal{N}} \sum_{b \in \mathcal{C}_k} D_k^b(t-1) \Phi_{k,i}^{b,c}(t) \Lambda_{k,i}(t-1), \quad i \in \mathcal{N}, \quad (11)$$

$$(1-x_i)(1-x_k) \Phi_{k,i}(t) = 0, \quad i, k \in \mathcal{N}, \quad (12)$$

$$x_k \Phi_{k,i}(t) = 0, \quad i, k \in \mathcal{N}, \quad (13)$$

$$(1-A_{k,i}(t)) \Phi_{k,i}(t) = 0, \quad i, k \in \mathcal{N}, \quad (14)$$

$$\sum_{i \in \mathcal{N}} x_i = S, \quad (15)$$

$$\sum_{c \in \mathcal{C}_i} D_i^c(t) = D_i(t), \quad i \in \mathcal{N}, \quad (16)$$

$$\sum_{c \in \mathcal{C}_i} \sum_{b \in \mathcal{C}_k} \Phi_{k,i}^{b,c}(t) = \Phi_{k,i}(t), \quad i, k \in \mathcal{N}, \quad (17)$$

$$\sum_{b \in \mathcal{C}_k} \Phi_{k,i}^{b,c}(t) \leq C_k, \quad i, k \in \mathcal{N}, \quad (18)$$

$$\sigma_{k,i}^{b,c}(t) = \sigma_{k,i}^{b,c}(t-1)(1 - \Phi_{k,i}^{b,c}(t)), \quad b \in \mathcal{C}_k, c \in \mathcal{C}_i, i, k \in \mathcal{N}, \quad (19)$$

$$\sum_{\substack{b \in \mathcal{C}_k \\ b \neq \arg \min |\sigma_{k,i}^{b,c}(t)|}} \Phi_{k,i}^{b,c}(t) = 0, \quad c \in \mathcal{C}_i, i, k \in \mathcal{N}, \quad (20)$$

$$E_i^P(t) = p_i(t)D_i(t) \leq P_i(t), \quad i \in \mathcal{N}, \quad (21)$$

$$E_k^{Tx}(t) = D_k(t) \sum_{i \in \mathcal{S}} \Phi_{k,i}(t) \psi_{k,i}(t) \leq \Psi_k(t), \quad k \in \mathcal{N}. \quad (22)$$

$$0 \leq \Lambda_{k,i}(t) \leq 1, x_i \in \{0, 1\}, \quad i, k \in \mathcal{N}, \quad (23)$$

$$0 \leq \Phi_{k,i}^{b,c}(t) \leq 1, b \in \mathcal{C}_k, c \in \mathcal{C}_i, \quad i, k \in \mathcal{N}. \quad (24)$$

The objective function (9) captures the balance between three key terms: the FedL loss in (9)(a) scaled by  $\alpha$ , the data processing resource use in (9)(b) scaled by  $\beta$ , and the D2D offloading resource use in (9)(c) scaled by  $\gamma$ . The data at sampled devices, i.e.,  $D_i(t)$  for  $i \in \mathcal{S}$ , changes over time in (10) based on the total received data  $R_i(t)$  for device  $i$ .  $R_i(t)$  is determined in (11) by scaling the data transmissions from each device  $k \in \hat{\mathcal{S}}$  to device  $i$  according to the data similarity, which we explain further in the next paragraph. Through (12)-(14), offloading only occurs between single-hop D2D neighbors from  $k \in \hat{\mathcal{S}}$  to  $i \in \mathcal{S}$ , while (15) ensures that  $(\mathcal{P})$  maintains compliance with the desired sampling size, i.e.,  $|\mathcal{S}^*| = S$ . The two expressions (16)-(17) capture the definitions for  $D_i(t)$  and  $\Phi_{k,i}(t)$ . Next, (18) ensures that the total offload rate at each device cannot exceed the quantity of its local data clusters, (19) is the update rule for centroid differences between two clusters  $b$  and  $c$  at two difference devices  $i$  and  $k$ , and (20) clarifies that data offloading only occurs among the matched data clusters as described in Sec. II-B. The constraints (21)-(22) ensure that our D2D offloading solution adheres to maximum device data processing limits  $P_i(t)$ , and D2D communication limits  $\Psi_i(t)$ . Finally, (23)-(24) express the lower and upper limits of the connectivity-similarity values and the data offloading variables.

**Similarity-aware D2D offloading:** The amount of raw data device  $i$  receives from  $k$  is  $\sum_{b \in \mathcal{C}_k} \Phi_{k,i}^{b,c}(t) D_k^b(t-1)$ . Ideally, device  $i$  will receive data from its neighbors that is dissimilar to  $D_i(0)$ . However, neither  $i$  nor  $k$  have full knowledge of each others' datasets in this distributed scenario (nor does the server). Therefore, data offloading in  $\mathcal{P}$  is conducted through a uniformly random selection of  $\Phi_{k,i}^{b,c}(t) D_k^b(t-1)$  data points (for each data cluster  $b \in \mathcal{C}_k$ ) from  $k$  to send to  $i$ . From the viewpoint of cluster  $b$  at device  $k$ , the estimated overlapping data that arrives at  $i$  is  $\Phi_{k,i}^{b,c}(t) D_k^b(t-1)(1 - \Lambda_{k,i}(t-1))$ , and the resulting useful data is  $\Phi_{k,i}^{b,c}(t) D_k^b(t-1) \Lambda_{k,i}(t-1)$ . We use the entirety of  $\Lambda_{k,i}(t)$  for this estimation because it is possible for data in cluster  $b$  to have overlap with multiple clusters in  $\mathcal{C}_i$  at device  $i$ . So, overall data similarity in the form of  $\Lambda_{k,i}(t)$  provides a better holistic estimate that an incoming data from device  $k$  to  $i$  will be similar to *any* existing data at device  $i$ . Subsequently, we update  $\sigma_{k,i}^{b,c}(t)$  per (19) and then use the data similarity calculation in (3) to update  $\Lambda_{k,i}(t)$ . In particular, when  $k$  transfers all of its data to  $i$  (i.e., when

$\Phi_{k,i}^{b,c}(t) = 1 \quad \forall b \in \mathcal{C}_k$ ),  $\sigma_{k,i}^{b,c} = 0 \quad \forall b \in \mathcal{C}_k$  and  $\forall c \in \mathcal{C}_i$ , leading to  $\Lambda_{k,i}(t) \rightarrow 0$ , thus preventing further data offloading according to (11). Imposing these constraints promotes data diversity among the sampled nodes through offloading.

**Solution overview:** Problem  $\mathcal{P}$  faces two major challenges: (i) it requires a concrete model of the loss function in (9)(a) with respect to the datasets, which is, in general, intractable for deep learning models [58], and (ii) even if the loss function is known, the coupled sampling and offloading procedures make this problem an NP-hard mixed integer programming problem. To overcome this, we will first consider the offloading subproblem for a fixed sampling strategy, and develop a sequential convex programming method to solve it in Sec. III. Then, we will integrate this into a graph convolutional network (GCN)-based methodology that learns the relationship between the network properties, sampling strategy (with its corresponding offloading), and the resulting FedL model accuracy in Sec. IV. An overall flowchart of our methodology is given in Fig. 3.

### III. DEVELOPING THE OFFLOADING OPTIMIZER

In this section, we study the offloading optimization subproblem of  $\mathcal{P}$ . Specifically, our theoretical analysis of (9)(a) under common assumptions will yield an efficient approximation of the FedL loss objective in terms of the offloading variables (Sec. III-B). We will then develop our solver for the resulting optimization (Sec. III-C).

#### A. Definitions and Assumptions

To aid our theoretical analysis of FedL, similar to [23], we will consider a hypothetical ML training process that has access to the entire dataset  $\mathcal{D}_{\mathcal{N}}(t)$  at each time instance. The parameter vector  $\mathbf{v}_k(t)$  for this centralized model is trained as follows: (i) at each global aggregation  $t = k\tau$ ,  $\mathbf{v}_k(t)$  is synchronized with  $\mathbf{w}_{\mathcal{S}}(t)$ , i.e.,  $\mathbf{v}_k(t) \leftarrow \mathbf{w}_{\mathcal{S}}(t)$ , and (ii) in-between global aggregation periods,  $\mathbf{v}_k(t)$  is trained based on gradient descent iterations to minimize the global loss  $F(\mathbf{v}_k(t)|\mathcal{D}_{\mathcal{N}}(t))$ .

**Definition 1 (Difference Between Sampled and Unsampled Gradients):** We define the instantaneous difference between  $\nabla F(\mathbf{w}_{\mathcal{S}}(t)|\mathcal{D}_{\mathcal{N}}(t))$ , the gradient with respect to the full dataset across the network, and  $\nabla F(\mathbf{w}_{\mathcal{S}}(t)|\mathcal{D}_{\mathcal{S}}(t))$ , the gradient with respect to the sampled dataset, as:

$$\zeta(\mathbf{w}_{\mathcal{S}}(t)) \triangleq \frac{G_{\mathcal{S}}(t)}{D_{\mathcal{S}}(t)} - \frac{\sum_{i \in \mathcal{N}} D_i(t) \nabla F(\mathbf{w}_{\mathcal{S}}(t)|\mathcal{D}_i(t))}{D_{\mathcal{N}}(t)}, \quad (25)$$

where  $G_{\mathcal{S}}(t) \triangleq \sum_{i \in \mathcal{S}} D_i(t) \nabla F(\mathbf{w}_{\mathcal{S}}(t)|\mathcal{D}_i(t))$  is the scaled sum of gradients on the sampled datasets, and  $D_{\mathcal{S}}(t) = \sum_{i \in \mathcal{S}} D_i(t)$  is the total data across the sampled devices.

**Definition 2 (Difference Between Sampled and Unsampled Gradients):** We define  $\delta_i(t)$  as the upper bound between the gradient computed on  $\mathcal{D}_i(t)$  for  $i \in \mathcal{S}$  and  $\mathcal{D}_{\mathcal{N}}(t)$  at time  $t$ :

$$\|\nabla F(\mathbf{w}_{\mathcal{S}}(t)|\mathcal{D}_i(t)) - \nabla F(\mathbf{w}_{\mathcal{S}}(t)|\mathcal{D}_{\mathcal{N}}(t)) - \zeta(\mathbf{w}_{\mathcal{S}}(t))\| \leq \delta_i(t). \quad (26)$$

We also make the following standard assumptions [23], [39] on the loss function  $F(\mathbf{w})$  for the ML model being trained:

**Assumption 1:** We assume  $F(\mathbf{w})$  is convex with respect to  $\mathbf{w}$ , L-Lipschitz, i.e.,  $\|F(\mathbf{w}) - F(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|$ , and  $\beta$ -smooth, i.e.,  $\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \leq \beta\|\mathbf{w} - \mathbf{w}'\|$ ,  $\forall \mathbf{w}, \mathbf{w}'$ .

Despite these assumptions, we will show in Sec. V that our results still obtain significant improvements in practice for neural networks which do not obey the above assumptions.

### B. Upper Bound on Convergence

For convergence analysis, we assume that devices only offload the same data once, and assume that recipient nodes always keep received data. This must be done to ensure that the optimal global model parameters remain constant throughout time. The following theorem gives an upper bound on the difference between the parameters of sampled FedL and those from the centralized learning, i.e.,  $\|\mathbf{w}_S(t) - \mathbf{v}_k(t)\|$ , over time:

**Theorem 1 (Upper Bound on the Difference Between Sampled FedL and Centralized Learning):** Assuming  $\eta \leq \beta^{-1}$ , the upper-bound on the difference between  $\mathbf{w}_S(t)$  and  $\mathbf{v}_k(t)$  within the local update period before the  $k$ -th global aggregation,  $t \in \{(k-1)\tau + 1, \dots, k\tau\}$ , is given by:

$$\|\mathbf{w}_S(t) - \mathbf{v}_k(t)\| \leq \frac{1}{\beta} \sum_{y=(k-1)\tau+1}^t \left( \Upsilon(y, k) + \|\zeta(\mathbf{w}_S(y-1))\| \right), \quad (27)$$

where  $\Upsilon(y, k) \triangleq \delta_S(y)(2^{y-1-(k-1)\tau} - 1)$ , and

$$\delta_S(t) \triangleq \left( \sum_{i \in \mathcal{S}} D_i(t) \delta_i(t) \right) \left( \sum_{i \in \mathcal{S}} D_i(t) \right)^{-1}. \quad (28)$$

*Proof:* See Appendix A.  $\square$

Through  $\delta_S(t)$ , Theorem 1 establishes a relationship between the difference in model parameters and the datapoints  $D_i(t)$  in the sampled set  $i \in \mathcal{S}$ . Using this, we obtain an upper bound on the difference between our  $\mathbf{w}_S(t)$  and the global minimizer of model loss  $\mathbf{w}^*(t) = \arg \min_{\mathbf{w}} F(\mathbf{w}|\mathcal{D}_N(t))$ :

**Corollary 1 (Upper Bound on the Difference Between Sampled FedL and the Optimal):** The difference of the loss induced by  $\mathbf{w}_S(t)$  compared to the loss induced by  $\mathbf{w}^*(t)$  for  $t \in \{(k-1)\tau, \dots, k\tau - 1\}$ , is given by:

$$\begin{aligned} & F(\mathbf{w}_S(t)|\mathcal{D}_N(t)) - F(\mathbf{w}^*(t)|\mathcal{D}_N(t)) \\ & \leq g(\hat{\Upsilon}(\hat{K})) \triangleq \left( t\xi\eta \left( 1 - \frac{\beta\eta}{2} \right) - \frac{(\hat{K}+1)L}{\beta\epsilon^2} \hat{\Upsilon}(\hat{K}) \right)^{-1}, \end{aligned} \quad (29)$$

where  $\hat{\Upsilon}(\hat{K}) \triangleq \sum_{y=(\hat{K}-1)\tau+1}^{\hat{K}\tau} \left( \Upsilon(\hat{K}, y) + \|\zeta(\mathbf{w}_S(y-1))\| \right)$ ,  $\hat{K} = \lfloor t/\tau \rfloor$ , and  $\xi = \min_k \frac{1}{\|\mathbf{v}_k((k-1)\tau) - \mathbf{w}^*(t)\|^2}$ .

*Proof:* See Appendix B.  $\square$

As our ultimate goal is an expression of (9)(a) in terms of the data  $\mathcal{D}_i(t)$  at each node, we consider the relationship between  $g(\hat{\Upsilon}(\hat{K}))$  and  $\mathcal{D}_i(t)$ , which is clearly non-convex through  $\hat{\Upsilon}(\hat{K})$ . Since  $\hat{\Upsilon}(\hat{K}) \ll 1$  (see Appendix B), (29) can be approximated using the first two terms of its Taylor series:

$$g(\hat{\Upsilon}) \approx \frac{1}{t\xi\eta(1 - \frac{\eta\beta}{2})} + \frac{(\hat{K}+1)L}{\beta\epsilon^2 \left( t\xi\eta(1 - \frac{\eta\beta}{2}) \right)^2} \hat{\Upsilon}. \quad (30)$$

At each time instant, the first term in the right hand side (RHS) of (30) is a constant. Thus, under this approximation, the RHS of (29) becomes proportional to  $\hat{\Upsilon}$ , which is in turn a function of  $\delta_S(t)$ . The final step is to bound the expression for  $\delta_i(t)$ , and thus their weighted sum  $\delta_S(t)$ , in terms of the  $D_i(t)$ ,  $\forall i \in \mathcal{S}$ .

**Proposition 1 (Upper Bound on the Difference Between Local Gradients):** The difference in gradient with respect to a sampled device dataset vs. the full dataset satisfies:

$$\begin{aligned} & \|\nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t)) - \nabla F(\mathbf{w}_S(t)|\mathcal{D}_N(t)) - \zeta(\mathbf{w}_S(t))\| \\ & \leq \left( \frac{D_N(t) - D_S(t)}{D_N(t)} \right) \overline{\nabla F(t)} + \frac{\gamma}{\sqrt{D_i(t)}} + C \equiv \delta_i(t), \end{aligned} \quad (31)$$

where  $C \triangleq (D_N(t))^{-1} \sum_{i \in \hat{\mathcal{S}}} D_i(t) \nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t))$ ,  $\hat{\mathcal{S}} = \mathcal{N} \setminus \mathcal{S}$ ,  $\gamma$  is a constant independent of  $\mathcal{D}_i(t)$ , and

$$\overline{\nabla F(t)} \triangleq (D_S(t))^{-1} \sum_{i \in \mathcal{S}} D_i(t) \nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t)). \quad (32)$$

*Proof:* See Appendix C.  $\square$

The above proposition relates each  $\delta_i(t)$  to the number of instantaneous data points available at device  $i$ .

### C. Offloading as a Sequential Convex Optimization

Using the result of (30) to replace the RHS of (29) implies that the FedL loss term in the objective function (i.e., (9)(a)) is proportional to  $\frac{1}{T} \sum_{t=1}^T \delta_S(t)$ , where  $\delta_S(t)$  is defined in (28) as a sum-of-ratios of  $\delta_i(t)$ . Considering  $\frac{1}{T} \sum_{t=1}^T \delta_S(t)$  as the objective in problem  $\mathcal{P}$  yields the sum-of-ratios problem in fractional programming [59]. The scale of existing solvers for the sum-of-ratios fractional programming problem (e.g. [60]) are on the order of ten ratios, which corresponds to ten devices in our case. Contemporary large-scale networks that may have hundreds of edge devices [2] therefore cannot be solved accurately or in a time-sensitive manner. Motivated by the above fact, we approximate  $\delta_S(t) \approx \frac{1}{S} \sum_{i \in \mathcal{S}} \delta_i(t)$ . Using this with (31), we obtain the following approximation for the loss function term in (9)(a):

$$\tilde{F}(t) = \underbrace{\left[ \left( \frac{D_N(t) - D_S(t)}{D_N(t)} \right) \overline{\nabla F(t)} \right]}_{(a)} + \underbrace{\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{\gamma}{\sqrt{D_i(t)}}}_{(b)}, \quad (33)$$

where term (a) is due to sampling and term (b) is the statistical error from the central limit theorem. Thus, for a known binary vector  $\mathbf{x}$  (i.e., a known  $\mathcal{S}$ ) that satisfies (15), we arrive at the following optimization problem for the D2D data offloading:

$$\begin{aligned} (\mathcal{P}_D) : \min_{\{\Phi(t)\}_{t=1}^T} & \frac{1}{T} \sum_{t=1}^T \left( \alpha \tilde{F}(t) + \beta \sum_{i \in \mathcal{S}} E_i^P(t) + \gamma \sum_{k \in \mathcal{N}} E_k^{Tx}(t) \right) \\ \text{s.t.} & (10) - (14), (16) - (24). \end{aligned}$$

Since the number of datapoints at the unsampled devices is fixed for all time,  $D_N(t)$  can be expressed as  $D_S(t) + D_{\hat{\mathcal{S}}}$ , where  $D_{\hat{\mathcal{S}}} = \sum_{i \in \hat{\mathcal{S}}} D_i$  is a constant. Consequently, both the coefficient of  $\overline{\nabla F(t)}$  in term (a) and the entirety of term (b) in (33) are decreasing functions of the quantity of data  $D_i(t)$  at sampled devices  $i \in \mathcal{S}$ . Furthermore, given  $\overline{\nabla F(t)}$ , both terms (a) and (b) in (33) are convex, via (10) and (11), with respect to the offloading variables in Problem  $\mathcal{P}_D$ . The only remaining challenge is then to obtain  $\overline{\nabla F(t)}$ , which we consider next.

**Sequential gradient approximation:** Obtaining  $\overline{\nabla F(t)}$  requires the knowledge of real-time gradients,  $\nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t))$ ,  $\forall i \in \mathcal{S}$ , which are unknown a priori. Furthermore, the gradients of the devices are only observed



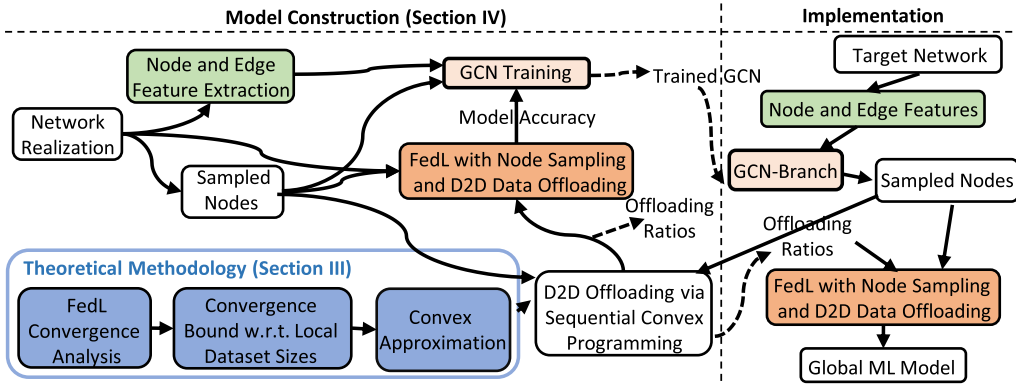


Fig. 3. Overview of the joint sampling and offloading methodology developed in Sec. III&IV. During model construction, our methodology trains a GCN, using various network realizations and sampled sets of nodes with the data offloading optimization from Sec. III. At the implementation stage, the target network uses the GCN-based algorithm developed in Sec. IV to obtain a sampled set of devices, which then undergo the D2D data offloading optimization process. Finally, we apply the results of the sampling and D2D data offloading processes for FedL, yielding a global ML model after training completion.

at the global aggregation time instances  $t = k\tau$ . Motivated by this, we approximate  $\nabla F(t)$  for  $t \in \{k\tau + 1, \dots, (k+1)\tau\}$ ,  $k \in \{1, \dots, K\}$ , using the gradients observed at the most recent global aggregation, i.e.,  $\nabla F(\mathbf{w}_S(k\tau) | \mathcal{D}_i(k\tau))$ ,  $i \in \mathcal{S}$  on which we perform a sequence of corrective approximations. Specifically, since the average loss  $F$  is convex,  $\nabla F(t)$  is expected to decrease over time. We assume that this decrease occurs linearly and approximate the real-time gradient using the previously observed gradient at the server as  $\nabla F(t) \approx \nabla F(k\tau) / \alpha_{k+1}^{t-k\tau}$ ,  $t \in \{k\tau + 1, \dots, (k+1)\tau\}$ ,  $\forall k \in \{1, \dots, K\}$ , where the scaling factor  $\alpha_{k+1} > 1$  is re-adjusted after every global aggregation  $k$ . Through the re-adjustment procedure, the server receives the gradients and computes the scaling factor for the each aggregation period as  $\alpha_{k+1} = \sqrt[t]{\nabla F((k-1)\tau) / \nabla F(k\tau)}$ .

Given the aforementioned characteristics of terms (a) and (b) in (33), our proposed iterative approximation of  $\nabla F(t)$ , and the fact that the constraints of  $\mathcal{P}_D$  are all affine at each time instance, we can solve this problem as a sequence of convex optimization problems over time. For this, we employ the CVXPY convex optimization software [61].

#### IV. SMART DEVICE SAMPLING WITH D2D OFFLOADING

We now turn to the sampling decisions in problem  $\mathcal{P}$ , which must be coupled with the offloading solution to  $\mathcal{P}_D$ . After explaining the rationale for our GCN-based approach (Sec. IV-A), we will detail our training procedure encoding the network characteristics (Sec. IV-B). Finally, we will develop an iterative procedure for selecting the sample set (Sec. IV-C).

##### A. Rationale and Overview of GCN Sampling Approach

Sampling the optimal subset of nodes from a resource-constrained network to maximize a utility function (in our case, minimizing the ML loss) has some similarity to 0-1 knapsack problem [62]. In this combinatorial optimization problem, a set of weights and values for  $n$  items are given, where each item can be either added or left out to maximize the value of the items within the knapsack subject to a weight capacity. Analogously, our sampling problem aims to maximize FedL accuracy while adhering to a sampling

budget  $S = \sum_{i \in \mathcal{N}} x_i$ . Strategies for the knapsack problem become unsuitable here because the value that each device provides to FedL is difficult to quantify: it depends on the ML loss function, the gradient descent procedure, and the D2D relationships from Sec. III.

To address these complexities, we propose a (separate) ML technique to model the relationship between network characteristics, the sampling set, and the resulting FedL model quality. Specifically, we develop a sampling technique based on active filtering of a Graph Convolutional Network (GCN) [63]. In a GCN, the learning procedure consists of sequentially feeding an input (the network graph) through a series of graph convolution [64] layers, which generalize the traditional convolution operation into non-Euclidean spaces and thus captures connections among nodes in different neighborhoods.

Our methodology is depicted in Fig. 4. GCNs excel at graph-based classification tasks, as they learn over the intrinsic graph structure. However, GCNs by themselves have performance issues when there are multiple good candidates for the classification problem [65]. This holds for our large-scale network scenario, as many high performing sets of sampling candidates can be expected. The data offloading scheme adds another important dimension: a sampled node  $i$  may perform poorly when considered in isolation, but it may have high processing capacities  $P_i(t)$  and be connected to unsampled nodes  $j \in \hat{\mathcal{S}}$  with large quantities of local data and high transfer limits  $\Psi_{j,i}(t)$ . We address these issues by (i) incorporating the solution from Sec. III into the GCN training procedure, and (ii) proposing *sampling GCN-branch*, a network-based post-processing technique that maps the GCN output to a sampling set by considering the underlying connectivity-similarity matrix.

##### B. GCN Architecture and Training Procedure

We consider a GCN function  $H(\pi, \tilde{\mathbf{A}})$  with two inputs: (i)  $\pi \in \mathbb{R}^{N \times U}$ , a matrix of  $U$  node features, and (ii)  $\tilde{\mathbf{A}} \in \mathbb{R}^{N \times N}$ , the augmented connectivity-similarity matrix. The feature vector for each node  $i$  is defined as  $\pi_i \triangleq [D_i(0), P_i(0), p_i(0), \theta_i(0)]$ , forming the rows of  $\pi$ , and the augmented connectivity-similarity matrix is defined as  $\tilde{\mathbf{A}} \triangleq$

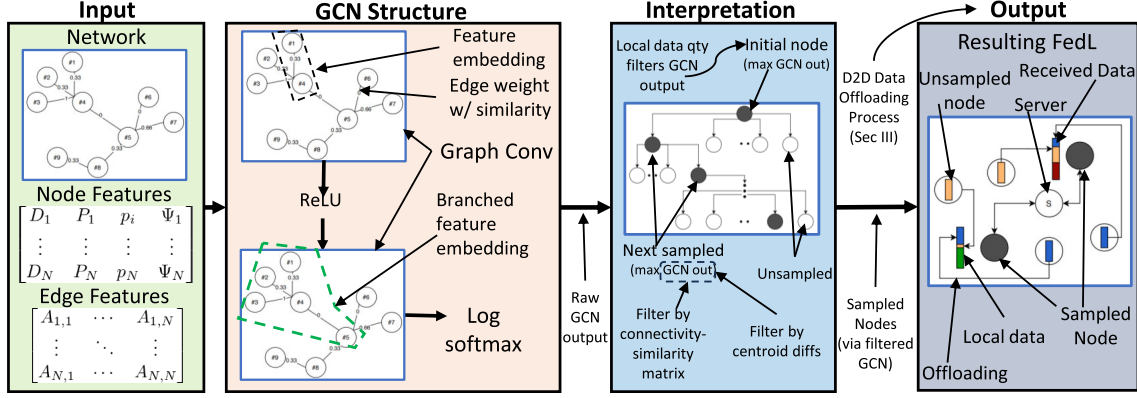


Fig. 4. Architecture of our GCN-branch sampling algorithm. Given a target network, GCN-branch extracts node and edge features, passing them through two GCN layers, each of which convolves features in local neighborhoods. This process returns raw output probabilities that we filter, based on data quantity, connectivity-similarity matrix, and centroid differences, to obtain a sampled set of nodes in the interpretation stage. With the resulting sampled set, we perform the D2D data offloading process from Sec. III, yielding an output ML model training process for FedL.

$\Lambda(0) + \mathbf{I}_N$ , where  $\mathbf{I}_N$  denotes the identity matrix [64].  $H$  consists of two graph convolutional layers separated by a rectified linear unit (ReLU) activation layer [63], as depicted in Fig. 4. The outputs of each layer are defined as:

$$\mathbf{H}^{(l)} \triangleq \sigma \left( \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(l-1)} \mathbf{Q}^{(l)} \right), \quad l \in \{1, 2\}, \quad (34)$$

where  $\tilde{\mathbf{D}}$  is the degree matrix of  $\tilde{\mathbf{A}}$ ,  $\mathbf{Q}^{(l)}$  denotes the trainable weights for the  $l$ -th layer, and  $\sigma$  represents ReLU activation. Note that  $\mathbf{H}^{(0)} = \boldsymbol{\pi}$ ,  $\mathbf{Q}^{(1)} \in \mathbb{R}^{U \times O}$ , and  $\mathbf{Q}^{(2)} \in \mathbb{R}^{O \times 1}$ , where  $O$  is the dimension of the second layer. Finally, log-softmax activation is applied to  $\mathbf{H}^{(2)} \in \mathbb{R}^N$  to convert the results into a vector of probabilities, i.e.,  $\boldsymbol{\Gamma} \in [0, 1]^N$ , representing the likelihood of each node belonging to the sampled set.

**GCN training procedure:** To train the GCN weights, we generate a set of sample network and node data realizations  $e = 1, \dots, E$  with the properties from Sec. II-A&II-B. For each realization, we calculate the matrices  $\boldsymbol{\pi}_e$  and  $\tilde{\mathbf{A}}_e$  corresponding to the inputs of the GCN. Then, for each candidate sampling allocation  $\mathbf{x}_e^s = [(x_e^s)_i]_{1 \leq i \leq N}$  (with  $\sum_i (x_e^s)_i = S$ ), we solve  $\mathcal{P}_D$  from Sec. III to obtain the offloading scheme, and then determine the loss of FedL resulting from model training and D2D offloading. Among these, we choose the  $\mathbf{x}_e^*$  that yields the smallest objective to be the target GCN output. The collection of  $[(\boldsymbol{\pi}_e, \tilde{\mathbf{A}}_e, \mathbf{x}_e^*)]_{e=1}^E$  form the training samples for the GCN.

As the number of devices  $N$  increases, the number of choices that will be considered for the sampled set increases combinatorially as  $\binom{N}{S}$ . An advantage of this GCN procedure is that it can be adapted to networks of varying size: once trained on a set of realizations for tolerable-sized values of  $N$ , the graph convolutional layer weights  $\mathbf{Q}^{(l)}$ ,  $l \in \{1, 2\}$ , can be applied to the desired network of varying size by shifting through the graph. Our obtained performance results in Sec. V verify this experimentally.

### C. Offloading-Aware Smart Device Sampling

Given any network graph, our procedure must solve the sampling problem at the point of FedL initialization, i.e.,  $t = 0$ . With the trained GCN in hand, we obtain  $\boldsymbol{\pi}$  and  $\tilde{\mathbf{A}}$  for the target network and calculate  $\boldsymbol{\Gamma} = H(\boldsymbol{\pi}, \tilde{\mathbf{A}})$ ,  $\boldsymbol{\Gamma} = [\Gamma_i]_{1 \leq i \leq N}$ .

Given this output, our sampling GCN-branch algorithm populates the set  $\mathcal{S}$  as follows. Let  $\mathcal{N}_p \subset \mathcal{N}$  be the subset of nodes in the 95th percentile of initial data quantity. Starting with  $\mathcal{S} = \emptyset$ , the first node is added according to  $\mathcal{S} = \mathcal{S} \cup \{s_1\}$ , where  $s_1 = \arg \max_{i \in \mathcal{N}_p} \Gamma_i$ . In this way, the first node sampled is the device with the highest GCN probability among nodes with large local data generation. To choose subsequent sampled nodes, the algorithm performs a recursive branch-based search on the initial connectivity-similarity matrix  $\Lambda(0)$  for nodes with the highest sampling probabilities and the smallest aggregate data similarity, as measured by maximin centroid difference, to the previously sampled nodes. In doing so, our algorithm filters the search space for the GCN to yield subsequent nodes with the most contribution to the existing sampled set of nodes. We develop this procedure for two main reasons. Firstly, the GCN scores devices individually, so two devices with high GCN probabilities may not necessarily yield a cohesive fit. Secondly, existing work [64], [65] has shown that a GCN individually may struggle to return a single optimal set in scenarios where multiple optimal sets are feasible, which is the case in large-scale networks.

Formally, we choose the  $n$ -th node addition as  $\mathcal{S} = \mathcal{S} \cup \{s_n\}$ , where  $s_n = \arg \max_{i \in \mathcal{R}_{s_{n-1}}} \Gamma_i$  with  $\mathcal{R}_{s_{n-1}}$  denoting the unsampled nodes above the 95th percentile of link dissimilarity to  $s_{n-1}$  (i.e., based on  $\Lambda(0)$ ) and the 80th percentile of maximin centroid difference relative to all the nodes in  $\mathcal{S}$  thus far. In this way, our branch algorithm relies on the GCN to decide which branch the sampling scheme will follow given its current sampled nodes (visualization in Fig. 4), so that subsequent selections are more likely to contain nodes with (i) different data distributions while (ii) leading to new neighborhoods that can contribute to the current set. Once the sampled set  $\mathcal{S}$  is determined, the offloading is scheduled for  $t = 0, \dots, T$  per the solver for  $\mathcal{P}_D$  from Sec. III.

**Summary of methodology:** Fig. 3 summarizes our methodology developed in Sec. III&IV for solving  $\mathcal{P}$ . Algorithm 1 also provides an overview of the key steps in the process, focusing specifically on the steps (i) to build the GCN model, (ii) for implementation on a target network, and (iii) for FedL operation on the target network. The sequential convex optimization for offloading (Sec. III) is embedded within the



**Algorithm 1** Summary of Key Steps in Overall Methodology

- 1: // GCN model construction - **Sec. IV-B**
- 2: *Step 1*: Generate randomized network realizations, each having graph  $G_z$  with  $z$  corresponding to a unique realization, for FedL.
- 3: *Step 2*: For each realization, find the best sets of sampled devices, e.g., via brute force search, while factoring in optimized data offloading via  $(\mathcal{P}_D)$  in **Sec. III-C**.
- 4: *Step 3*: Extract node features (i.e.,  $P_i, p_i, D_i, \Psi_i \forall i \in \mathcal{N}$ ) and edge features (i.e.,  $\psi_{i,j}, A_{i,j}, \lambda_{i,j} \forall i, j \in \mathcal{N}, i \neq j$ ) for the network realizations.
- 5: *Step 4*: Use the results of Step 2 and 3 as the ground truth to train the GCN's parameters,  $\mathbf{Q}^{(l)}$ ,  $l \in \{1, 2\}$  per (34) in **Sec. IV-B**.
- 6:
- 7: // Implementation and operation for a given network - follows **Sec. IV-C**
- 8: *Step 1*: Extract node features (i.e.,  $P_i, p_i, D_i, \Psi_i \forall i \in \mathcal{N}$ ) and edge features (i.e.,  $\psi_{i,j}, A_{i,j}, \lambda_{i,j} \forall i, j \in \mathcal{N}, i \neq j$ ) for a target network,  $G = \{\mathcal{N}, \mathcal{E}(t)\}$ .
- 9: *Step 2*: Pass the target network  $G = \{\mathcal{N}, \mathcal{E}(t)\}$  and its features (i.e.,  $P_i, p_i, D_i, \Psi_i, \psi_{i,j}, A_{i,j}, \lambda_{i,j} \forall i, j \in \mathcal{N}, i \neq j$ ) through the trained GCN, which has trained weights,  $\mathbf{Q}^{(l)}$ ,  $l \in \{1, 2\}$ , in the form of (34) as a result of the procedure in **Sec. IV-B**.
- 10: *Step 3*: Perform GCN-branch to obtain the set of sampled devices  $\mathcal{S}$ , described in **Sec. IV-C**.
- 11: *Step 4*: Perform the D2D data offloading and resource optimization process of  $(\mathcal{P}_D)$  from **Sec. III-C** to determine the offloading ratios  $\Phi(t)$  for the sampled set  $\mathcal{S}$ , given some optimization regime  $\alpha, \beta$ , and  $\gamma$ .
- 12:
- 13: // Operation with device sampling and D2D data offloading - **Sec. V**
- 14: *Step 1*: Network uses GCN-branch to select devices  $i \in \mathcal{S}$  as the sampled nodes and  $(\mathcal{P}_D)$  to obtain the offloading ratios  $\Phi(t)$ .
- 15: *Step 2*: Sampled devices  $i \in \mathcal{S}$  are initialized with ML model parameters  $\mathbf{w}_i^0$ , and locally update their ML model parameters,  $\mathbf{w}_i(t)$ , via gradient descent in (6).
- 16: *Step 3*: At each training iteration  $t$ , sampled devices  $i \in \mathcal{S}$  receive data  $\Phi_{j,i}(t)D_j(t)$  from nearby devices  $j \in \mathcal{N}, j \notin \mathcal{S}$ .
- 17: *Step 4*: At each  $k$ -th aggregation, the server aggregates the ML model parameters  $\mathbf{w}_S(k\tau) = \frac{\sum_{i \in \mathcal{S}} \Delta_i(k\tau) \mathbf{w}_i(k\tau)}{\sum_{i \in \mathcal{S}} \Delta_i(k\tau)}$ , and then synchronizes the sampled devices' ML model parameters to  $\mathbf{w}_S(k\tau)$ .
- 18: *Step 5*: Training process continues until  $t = T$ .

GCN-based sampling procedure (Sec. IV). Once the model is trained on sample network realizations, it is applied to the target network to generate the  $\mathcal{S}$  and  $\Phi(t)$  for FedL.

## V. EXPERIMENTAL EVALUATION

### A. Setup and Experimental Procedure

1) *Network Characteristics via Wireless Testbed*: We employed our IoT testbed in Fig. 5 to obtain device and

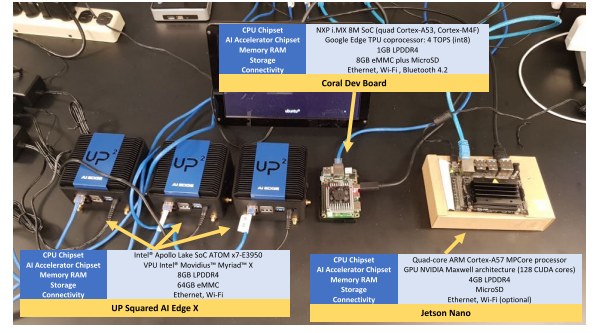


Fig. 5. IoT testbed used to generate device and link characteristics.

TABLE I

GLOBAL AGGREGATIONS REQUIRED BY EACH SCHEME TO REACH A CERTAIN PERCENTAGE OF FEDL MODEL ACCURACY ON MNIST (69%) AND F-MNIST (41%) WITH  $S = 6$  AND VARYING  $N$ . OUR SAMPLING METHODOLOGY WITH OFFLOADING CONSISTENTLY OBTAINS THE FASTEST TRAINING TIME. DASHES INDICATE THAT A METHOD WAS UNABLE TO REACH THE ACCURACY THRESHOLD

Sampling Method	Devices (MNIST)			Devices (F-MNIST)		
	600	700	800	600	700	800
Ours w/ offload	4	5	3	3	3	2
DPP w/ offload	12	12	9	15	15	16
PoC w/ offload	—	14	—	13	17	13
PYR w/ offload	—	17	17	—	—	—
Ours	10	12	14	10	12	13
DPP	13	16	10	16	18	20
PoC	—	—	—	18	—	—
PYR	—	19	17	—	—	—

communication characteristics. It consists of Jetson Nano, Coral Dev Board TPU, and UP Squared AI Edge boards configured to operate in D2D mode. We used Dstat [66] to collect the device resources and power consumption. We map the measured computing resources (in CPU cycles and RAM) and the corresponding power consumption (in mW) at devices to the costs and capacities in our model by calculating the Gateway Performance Efficiency Factor (GPEF) [54]. Specifically, to determine the processing costs  $p_i(t)$ , we measured the GPEF of the devices running gradient iterations on the MNIST dataset [67]. For the processing capacities  $P_i(t)$ , we pushed the devices to 100% load and measured the GPEF. We initialized the devices at 25%-75% loads, and treated the available remaining capacity as the receive buffer parameter  $\theta_i(t)$ .

For the transmission costs, we measured GPEF spent on D2D offloading over WiFi. Our WiFi links, when only devoted to D2D offloading, consistently saturated at 12 Mbps. To simulate the effect of external tasks, we limit available bandwidth for D2D to 1, 6, and 9 Mbps. We then calculated the transmission resource budget for devices as transfer limits  $\Psi_i(t)$ , and modelled unit transfer costs  $\psi_{i,j}(t)$  as normalized D2D latency.

2) *Datasets and Large-Scale Network Generation*: For FedL training, we use MNIST and Fashion-MNIST (F-MNIST) [68] image classification datasets. We consider a CNN predictor composed of two convolutional layers with ReLU activation and dropout. The devices perform  $\tau = 5$  rounds of gradient descent with a learning rate  $\eta = 0.01$ . Following [39], we generate network topologies with  $N = 100$  to

800 devices using Erdős–Rényi graphs with link formation probability (i.e.,  $A_{i,j} = 1$ ) of 0.1. To produce local datasets across the nodes that are both overlapping and non-i.i.d., the datapoints at each node are chosen uniformly at random with replacement from datapoints among three labels (i.e., image classes). Differentiating the labels between devices captures dataset heterogeneity (i.e., from different devices collecting data from different labels). The number of initial datapoints  $D_i(0)$  at each device follows a normal distribution with mean  $\mu = (D_{\mathcal{N}}(0))N^{-1}$  and variance  $\sigma^2 = 0.2\mu$ . We further estimate the initial similarity weights  $\lambda_{i,j}(0)$  based on the procedure discussed in Sec. II-B.

For the GCN-based sampling procedure, we train the model on small network realizations of ten devices. We consider sampling budgets of  $S = 3$  to 6, with corresponding training samples  $E$  for each case. We save the resulting graph convolutional layer weights  $\mathbf{Q}^{(1)}$  and  $\mathbf{Q}^{(2)}$  for each choice of  $S$  and reapply them on the larger target target networks.

## B. Results and Discussion

In the following experiments, we compare our methodology to several baseline sampling and offloading schemes. The four sampling strategies considered are *Ours*, data proportional probabilities (*DPP*) [26], [37], power-of-choice (*PoC*) [27], and PyramidFL (*PYR*) [69]. *DPP* is a random device sampling technique based on the quantity of local data at network devices, with larger datasets incurring a higher probability of sampling for FedL aggregations. On the other hand, *PoC* is a sampling technique based on the instantaneous training losses across network devices, with larger losses equating to a higher probability of being part of the sampled device set. *PYR* is a time-varying sampling technique based on a combination of exploration and exploitation across network devices, with exploitation referring to the best performing nodes and exploration representing careful selection of other nodes. For *PYR*, we use an even ratio of 0.5 for both exploration and exploitation. Each of the four sampling schemes is investigated both *with* and *without* data offloading. For our sampling methodology, we employ our corresponding data offloading methodology. For *DPP*, *PoC*, and *PYR* sampling methodologies, we perform a greedy offloading that minimizes the communication resource costs while maintaining the same nominal offloaded data quantity as our methodology. A baseline of FedL with no sampling (i.e., all nodes active) and no offloading is also included.

In the following, we will first investigate classification performances, convergence speeds, and resource utilization of the various sampling methodologies. Then, we will qualitatively examine characteristics of our optimization formulation. Finally, we further examine different optimization regimes (i.e., combinations of  $\alpha$ ,  $\beta$ , and  $\gamma$  on the performance of our methodology), the impact of time-varying links, and integration with FedDrop [70] in the Appendices.

1) *Model Accuracy*: Figs. 6 and 7 show FedL accuracy for different sampling baselines on two popular ML datasets in six different combinations ( $S = 3, 4, 5$  and  $N = 100, 200$ ). For these experiments, our proposed methodology solves ( $\mathcal{P}_D$ )

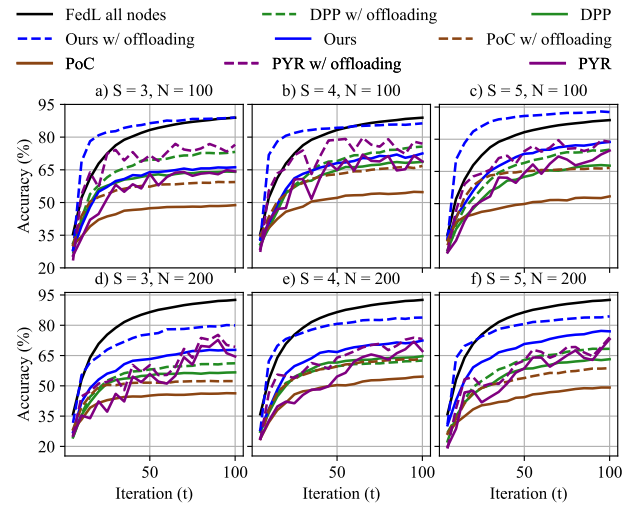


Fig. 6. Testing classification accuracies over training iterations on MNIST obtained by the sampling schemes with and without offloading, and by FedL using all nodes, for different sampled sizes ( $S$ ) and nodes ( $N$ ). For  $S > 3$ , our proposed sampling methodology with offloading consistently obtains a wide margin of improvement over all schemes.

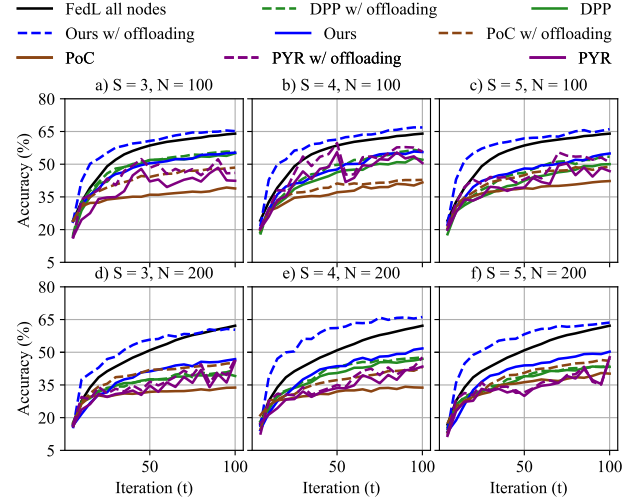


Fig. 7. Testing classification accuracies on F-MNIST for the same setup as in Fig. 6. The results are consistent with the MNIST dataset. The wide margin of improvement obtained by our proposed sampling scheme with offloading vs. without emphasizes the benefit of considering these two aspects jointly for FedL optimization.

for the data offloading and resource optimization component using  $\alpha = 100$ ,  $\beta = 0.001$ , and  $\gamma = 0.01$  for MNIST or  $\gamma = 0.006$  for FMNIST.<sup>2</sup> We further compare the above  $\alpha$ ,  $\beta$ , and  $\gamma$  combination, which we term as the balanced regime, versus the high energy costs regime for  $\alpha$ ,  $\beta$ , and  $\gamma$  in Appendix D, which increases  $\beta$  and  $\gamma$  by an order of magnitude for both datasets.

For the experiments involving data offloading, *Ours w/ offloading* consistently demonstrates at least a 5% accuracy improvement over the baseline methodologies for MNIST (Fig. 6) with a maximum of 13% accuracy improvement in Fig. 6c). Similarly on F-MNIST (Fig. 7), our methodology is able to continue outperforming the baseline sampling

<sup>2</sup>These values of  $\alpha$ ,  $\beta$ , and  $\gamma$  enable the three objective function terms of ( $\mathcal{P}_D$ ) to all have the same order of magnitude, i.e.,  $\mathcal{O}(10^2)$ .

methodologies for various  $S$  and  $N$  combinations. The final accuracy attained by *Ours w/ offloading* is even able to match or outperform *FedL all nodes* in some cases. Our joint sampling and data offloading methodology is able to achieve better performance due to two main reasons: (i) it minimizes the data skew resulting from unbalanced label frequencies, and (ii) it ensures higher quality of local datasets at sampled nodes, which reduces bias caused by multiple local gradient descents. Without offloading, our sampling methodology is also able to either outperform or match the accuracies obtained by the *DPP*, *PoC*, and *PYR* sampling strategies. Additionally, our method with offloading is able to offer consistent improvements over sampling without offloading in most cases, whereas the effects of data offloading are smaller or even detrimental for the baseline methodologies. This emphasizes the importance of designing the sampling and offloading schemes for FedL jointly.

2) *Model Convergence Speed*: We next compare the convergence speeds of our methodology to the other schemes in terms of the number of global aggregations needed to reach specific accuracy thresholds. Table I compares the convergence speeds on MNIST (to reach 69%) and F-MNIST (to reach 41%), respectively, for  $N = 600, 700, 800$  and  $S = 6$ , and uses a horizontal line to indicate when a methodology is unable to reach the specific accuracy thresholds. We investigate sampling methodologies with and without data offloading.

For methodologies involving data offloading, our method obtains the fastest convergences rates. Specifically, our method is at least 58% and 80% faster, in terms of global aggregations needed to reach the accuracy threshold, than the other baselines on MNIST and F-MNIST respectively. Furthermore, our method is only one to consistently be able to reach the accuracy thresholds. For example, while the baseline *PoC w/ offload* is able to reach the threshold for MNIST with  $N = 700$ , it fails to do for  $N = 600$  and  $N = 800$  on the same dataset.

Even without data offloading, our sampling methodology is still able to consistently reach the accuracy thresholds for both MNIST and F-MNIST. By contrast, both *DPP* and *PoC* baselines are unable to consistently obtain the accuracy thresholds for F-MNIST. Overall, we see that our joint sampling and offloading methodology obtains faster training speeds than the other methods. Enabling offloading is also seen to improve the convergence speeds of each sampling scheme; in fact, without offloading, several baseline cases fail to reach the given percentage of the FedL baseline.

3) *Resource Utilization*: Finally, we compare the resource utilization for the different schemes in terms of the total data processed and energy used (measured by GPEF consumption) across a sampled set  $S$  of 6 nodes in a network of  $N = 700$  edge devices. Fig. 8 depicts the total data processed while Fig. 9 shows the corresponding energy costs in GPEF. Both figures compare the resource consumption (data processed or energy used) that sampling methodologies with and without offloading incur in order to reach specific accuracy thresholds on both MNIST and F-MNIST. We see that our sampling methodology (both with and without offloading) requires fewer

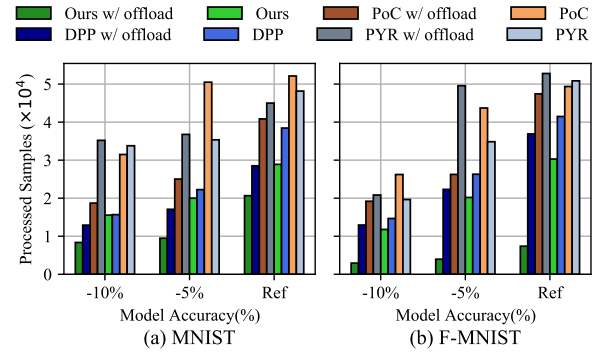


Fig. 8. Number of samples processed by the schemes with and without offloading to reach within a certain percentage of a reference testing accuracy (Ref, 69% for MNIST and 41% for F-MNIST). Our smart sampling with offloading methodology scales the best in terms of data needed for ML model training.

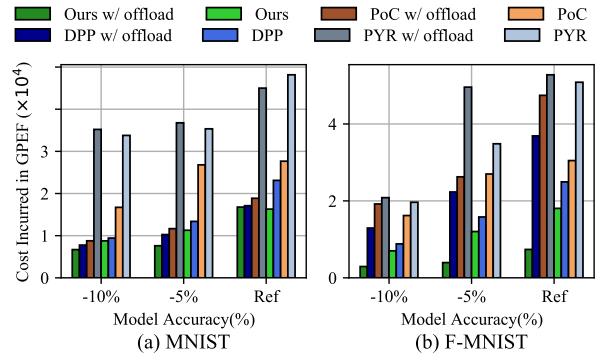


Fig. 9. The total associated data processing and communication costs measured in GPEF linked to Fig. 8. Our proposed sampling methodology with and without offloading obtain some of the lowest costs among all the baseline comparisons regardless of dataset.

data to be processed and thus results in less energy use relative to the baselines for various accuracy thresholds.

Specifically, on the processed data side (Fig. 8), *Ours w/ offload* processes the least amount of training data across all cases. As the reference accuracy level increases, our methodology consistently requires fewer datapoints compared to the other methods (over 60% fewer on average), emphasizing its ability to get the most value from each processed datum. This emphasizes the benefits of well-designed data offloading in reducing the computational burden across the network as a whole, which is especially valuable in IoT and mobile edge applications. Furthermore, *Ours* without offloading requires fewer processed data than all other methods under evaluation aside from *Ours w/ offload*, which highlights the resource efficiency of our sampling methodology on its own.

Our methodologies are able to maintain their advantages on the energy costs side (Fig. 9). Similar to Fig. 8, our methodology with and without data offloading incurs the least amount of energy use as measured in terms of GPEF. On average, our method uses less energy than the baseline methods on MNIST and significantly fewer energy than the baseline methods on FMNIST. While *Ours w/ offload* processed the least amount of training data in Fig. 8, we see that, for some cases, the energy costs incurred by it are fairly similar to those total costs incurred by *Ours* without data offloading in Fig. 9. So even though *Ours w/ offload* may process fewer data and thus have a lower data processing cost, it seems that the



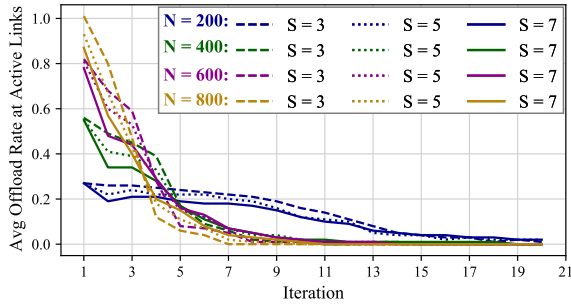


Fig. 10. The average data offloading rates at active D2D links while varying network sizes and sampling quantities. Larger networks exhibit higher initial average offloading rates and correspondingly have data offloading rates that decay faster in the FedL training process.

data offloading costs incurred over D2D links can, in specific scenarios, balance out the energy savings from fewer processed data for our methodology.

These experiments demonstrate that our joint optimization method exceeds baseline performances in terms of model accuracy, convergence speed, and resource utilization. Next, we experimentally characterize the behavior of the optimization formulation ( $\mathcal{P}$ ) by investigating its responses over training iterations, and the two energy scaling variables,  $\beta$  and  $\gamma$ .

4) *Optimization Characterization:* We characterize the behavior of our optimization formulation ( $\mathcal{P}$ ) in three key aspects: (i) its behavior over time as the FedL process converges, (ii) its sensitivity to the data processing energy use (scaled by  $\beta$ ), and (iii) its responses to the data offloading energy use (scaled by  $\gamma$ ). For all of these experiments, we investigate on networks of  $N \in \{200, 400, 600, 800\}$  nodes while sampling  $S \in \{3, 5, 7\}$  devices. In addition, these experiments all use  $\alpha = 100$ ,  $\beta = 0.001$ , and  $\gamma = 0.001$  unless stated otherwise.

First, we characterize the behavior of ( $\mathcal{P}$ ) with respect to training iterations in Fig. 10. Here, Fig. 10 highlights the average offloading rate at active links (i.e., links that have a non-zero data offloading rate). One of the major effects shown in Fig. 10 is device/link saturation, in which D2D links gradually cease to be used for data offloading. The main reason for this is that, over time, sampled devices gather enough data from nearby willing unsampled devices that they no longer have any remaining data processing capabilities (as described in (21)). This effect becomes more noticeable as the edge network grows in size, which leads to more D2D connections from unsampled to sampled devices and thus faster device/link saturation. For instance in Fig. 10, the average data offloading rate at active D2D links rapidly decreases (in less than 10 iterations) from more than 0.8 to 0 for  $N = 800$  regardless of sampled device set size, while for  $N = 200$ , the average data offloading rate requires roughly 20 iterations to approach 0 for all  $S \in \{3, 5, 7\}$ .

More D2D connections among devices also leads to higher initial average data offloading rates among active D2D links in Fig. 10. The two reasons for this are: (i) more links leads to more valuable data offloading opportunities and thereby more data offloading, and (ii) simulation setup in Sec. V-A.2 in

which the datasets for larger networks are initialized at smaller sizes in order to form a partition of the full ML dataset.

Next, we investigate the sensitivity of ( $\mathcal{P}$ ) to  $\beta$  in Fig. 11. As we vary  $\beta$  from 0.001 to 10, the cost of processing data becomes more expensive relative to the FedL loss term in ( $\mathcal{P}$ ) and (9). Initially, for small  $\beta$  in  $N = 200$ , the cost of processing data is essentially negligible and the average data processed is at a stable point until the  $\beta$  scaling reaches a point that is comparable in value to the FedL loss term in (9). Further increases in  $\beta$  then lead to reductions in the average data processed per device until nearly no data is processed, and, correspondingly, the estimated and normalized FedL loss increases.

Fig. 11 also highlights some of the effects of large-scale networks and sampled device set size. Here, as the networks grow in size from 200 to 800 devices, we see that the average data processed per sampled device is increasing, specifically from under 200 data per device when  $N = 200$  to over 400 data per device when  $N = 800$ . This is because larger networks have more links, and therefore more valuable data offloading, which leads to more D2D offloading and more data processed. On the other hand, as the sampled device set size increases from  $S = 3$  to  $S = 7$ , the opposite effect occurs as the average data processed per sampled device decreases. More sampled devices not only leads to more D2D data offloading opportunities, but also enables the unsampled devices to spread out their data to more sampled devices. As a result, the burden of data processing can be shared among more sampled devices, thus leading to the decrease in average data processed per sampled device.

Finally, we observe the responsiveness of ( $\mathcal{P}$ ) to changes in the value of D2D offloading costs in Fig. 12 by measuring the change in average data offload quantity relative to  $\gamma$ , the scaling of D2D offloading. As  $\gamma$  increases from 0.00001 to 0.01, the cost of D2D data offloading becomes more expensive relative to the data processing and FedL cost terms from (9). Small  $\gamma$  initially has little impact on average data offloading in small networks of  $N = 200$  until the  $\gamma$  scaling term makes the D2D communication cost comparable to that of the FedL performance term in (9). Then average data offloading quantity rapidly decreases to 0 and correspondingly the normalized FedL term increases.

Fig. 12 provides two additional insights. First, as the network grows in size from  $N = 200$  to  $N = 800$ , the average data offload quantity per D2D link remains quite stable, with all Fig. 12a)-Fig. 12d) demonstrating average data offload rates between 10 to 20 for  $\gamma = 0.00001$ . That being said, larger networks do demonstrate heightened sensitivity to  $\gamma$ . For example, for all  $S \in \{3, 5, 7\}$ , while the average data offloaded in a network with  $N = 200$  remains stable for  $\gamma \in [0.00001, 0.001]$ , the average data offloaded decreases more and more rapidly as the network size grows to 400, 600, and then 800 nodes for the same  $\gamma$  range. The second is that as the sampled device set size increases from  $S = 3$  to  $S = 7$ , the average data offloaded per D2D link increases. This is due to the existence of more D2D connections between unsampled and sampled devices, enabling more links to be active.

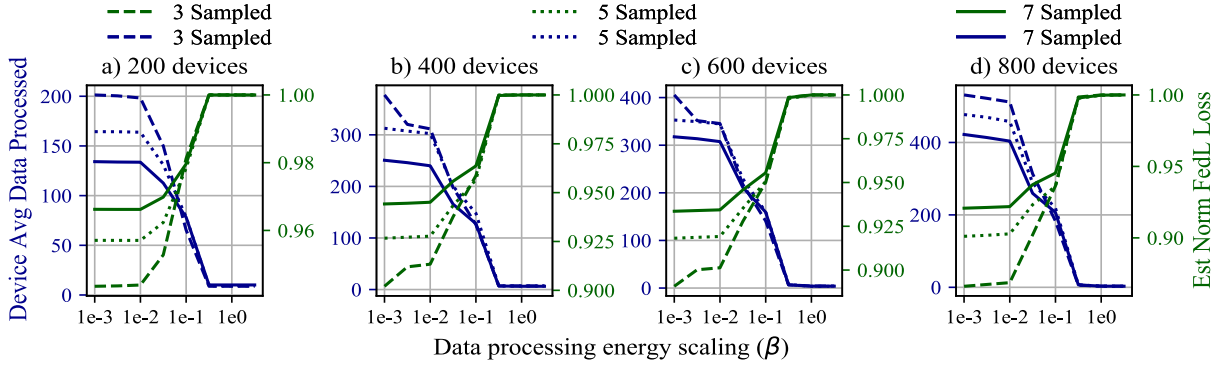


Fig. 11. Average data processed per sampled device when varying the data processing energy scaling,  $\beta$  from (9). As  $\beta$  increases, networks are disincentivized to perform costly ML training, which leads to lower average data processed per sampled device and correspondingly increases in the FedL loss. In addition, as networks grow in size, the average data processed per sampled device increases as more cost-effective D2D links become available.

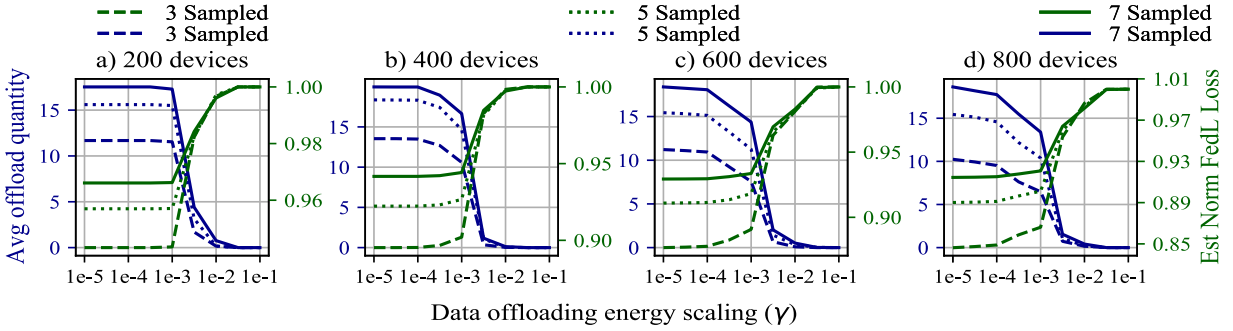


Fig. 12. Average data offloaded from unsampled to sampled device when varying the data offloading energy scaling,  $\gamma$  from (9). As  $\gamma$  increases in value, D2D links become more expensive to use, leading to an expected decrease in average offloaded data quantity and a corresponding growth in the FedL loss term. The effect of sampling size can also be seen to influence the average data offloading quantity, with smaller sampling quantities yielding lower average data offloading rates as there are physically less links from unsampled to sampled devices.

## VI. CONCLUSION AND FUTURE WORK

In this paper, we developed a novel methodology to solve the joint sampling and D2D offloading optimization problem for FedL. Our theoretical analysis of the offloading subproblem produced new convergence bounds for FedL, and led to a sequential convex optimization solver. We then developed a GCN-based algorithm that determines the sampling strategy by learning the relationships between the network properties, the offloading topology, the sampling set, and the FedL accuracy. Our implementations using popular datasets and real-world IoT measurements from our testbed demonstrated that our methodology obtains significant improvements in terms of datapoints processed, training speed, and resulting model accuracy compared to several other algorithms, including FedL using all devices. Future work may consider the integration of further realistic network characteristics such as unlabeled data [71], [72] on FedL, or investigate the integration of realistic wireless conditions, such as fading, shadowing, and interference among others [73], for D2D cooperation within large-scale networks for FedL.

### APPENDIX A PROOF OF THEOREM 1

Since  $\mathbf{v}_k(t) = \mathbf{v}_k(t-1) - \eta \nabla F(\mathbf{v}_k(t-1) | \mathcal{D}_{\mathcal{N}}(t))$ ,  $\mathbf{w}_S(t) = \mathbf{w}_S(t-1) - \eta (\nabla F(\mathbf{w}_S(t-1) | \mathcal{D}_{\mathcal{N}}(t)) + \zeta(\mathbf{w}_S(t-1)))$ , we get:

$$\begin{aligned} & \|\mathbf{w}_S(t) - \mathbf{v}_k(t)\| \\ &= \|\mathbf{w}_S(t-1) - \mathbf{v}_k(t-1) - \eta \zeta(\mathbf{w}_S(t-1))\| \end{aligned}$$

$$- \eta \nabla F(\mathbf{w}_S(t-1) | \mathcal{D}_{\mathcal{N}}(t)) + \eta \nabla F(\mathbf{v}_k(t-1) | \mathcal{D}_{\mathcal{N}}(t))\| \quad (35)$$

We simplify (35) through the following steps:

$$\begin{aligned} & \|\mathbf{w}_S(t) - \mathbf{v}_k(t)\| \stackrel{(a)}{\leq} \|\mathbf{w}_S(t-1) - \mathbf{v}_k(t-1)\| \\ & + \eta \sum_{i \in \mathcal{N}} \frac{D_i(t-1)}{D_{\mathcal{N}}(t)} \|\nabla F(\mathbf{w}_S(t-1) | \mathcal{D}_i(t-1)) \\ & - \nabla F(\mathbf{v}_k(t-1) | \mathcal{D}_i(t-1))\| + \eta \|\zeta(\mathbf{w}_S(t-1))\| \\ & \stackrel{(b)}{\leq} \|\mathbf{w}_S(t-1) - \mathbf{v}_k(t-1)\| + \eta \|\zeta(\mathbf{w}_S(t-1))\| \\ & + \eta \beta \sum_{j \in \mathcal{N}} \frac{D_j(t-1)}{D_{\mathcal{N}}(t) D_{\mathcal{S}}(t)} \sum_{i \in \mathcal{S}} D_i(t-1) \|\mathbf{w}_S(t-1) - \mathbf{v}_k(t-1)\| \\ & \stackrel{(c)}{\leq} \|\mathbf{w}_S(t-1) - \mathbf{v}_k(t-1)\| + \eta \|\zeta(\mathbf{w}_S(t-1))\| \\ & + \sum_{j \in \mathcal{N}} \frac{D_j(t-1)}{D_{\mathcal{N}}(t) D_{\mathcal{S}}(t)} \sum_{i \in \mathcal{S}} D_i(t-1) \frac{\delta_i(t)}{\beta} (2^{t-1-(k-1)\tau} - 1) \\ & \stackrel{(d)}{\leq} \|\mathbf{w}_S(t-1) - \mathbf{v}_k(t-1)\| + \frac{1}{\beta} \|\zeta(\mathbf{w}_S(t-1))\| \\ & + \sum_{j \in \mathcal{N}} \frac{D_j(t-1)}{D_{\mathcal{N}}(t)} \frac{\delta_{\mathcal{S}}(t)}{\beta} (2^{t-1-(k-1)\tau} - 1), \quad (36) \end{aligned}$$

where (a) results from expanding  $\nabla F(\mathbf{v}_k(t-1) | \mathcal{D}_{\mathcal{N}}(t))$  and applying the triangle inequality repeatedly, (b) follows from using the  $\beta$ -smoothness of the loss function and the triangle inequality, (c) applies Lemma 3 from [23], and (d) uses the

expanded form of  $\delta_S(t)$  in (28). We then rearrange (36):

$$\begin{aligned} & \|\mathbf{w}_S(t) - \mathbf{v}_k(t)\| - \|\mathbf{w}_S(t-1) - \mathbf{v}_k(t-1)\| \\ & \leq \frac{\Upsilon(t, k)}{\beta} + \frac{1}{\beta} \|\zeta(\mathbf{w}_S(t-1))\|. \end{aligned} \quad (37)$$

Since  $\|\mathbf{w}_S(t) - \mathbf{v}_k(t)\| = 0$  when re-synchronization occurs at  $t = k\tau$ ,  $\forall k \in \{1, \dots, K\}$ , we express  $\|\mathbf{w}_S(t) - \mathbf{v}_k(t)\|$  as:

$$\begin{aligned} & \|\mathbf{w}_S(t) - \mathbf{v}_k(t)\| \\ & = \sum_{y=(k-1)\tau+1}^t \|\mathbf{w}_S(y) - \mathbf{v}_k(y)\| - \|\mathbf{w}_S(y-1) - \mathbf{v}_k(y-1)\| \\ & \leq \frac{1}{\beta} \sum_{y=(k-1)\tau+1}^t (\Upsilon(y, k) + \|\zeta(\mathbf{w}_S(y-1))\|). \end{aligned} \quad (38)$$

## APPENDIX B

### PROOF OF COROLLARY 1

We first define  $\theta_k(t) \triangleq F(\mathbf{v}_k(t)) - F(\mathbf{w}^*(t)) \geq \epsilon$ . Since  $F$  is  $L$ -Lipschitz, we apply the result of Theorem 1, and Lemmas 2 and 6 from [23] to obtain:

$$\begin{aligned} & \theta_{\hat{K}+1}(t)^{-1} - \theta_1(0)^{-1} \\ & = \left( \theta_{\hat{K}+1}(t)^{-1} - \theta_{\hat{K}+1}(\hat{K}\tau)^{-1} \right) \\ & \quad + \left( \theta_{\hat{K}+1}(\hat{K}\tau)^{-1} - \theta_{\hat{K}}(\hat{K}\tau)^{-1} \right) + \left( \theta_{\hat{K}}(\hat{K}\tau)^{-1} - \theta_1(0)^{-1} \right) \\ & \geq \left( (t - \hat{K}\tau)\xi\eta \left( 1 - \frac{\beta\eta}{2} \right) \right) + \hat{K}\tau\xi\eta \left( 1 - \frac{\beta\eta}{2} \right) - \frac{L}{\beta\epsilon^2} \hat{\Upsilon}(\hat{K}) \\ & \quad - (\hat{K} - 1) \frac{L}{\beta\epsilon^2} \hat{\Upsilon}(\hat{K}) = t\xi\eta \left( 1 - \frac{\beta\eta}{2} \right) - \hat{K} \frac{L}{\beta\epsilon^2} \hat{\Upsilon}(\hat{K}), \end{aligned} \quad (39)$$

where  $\hat{\Upsilon}(\hat{K}) \triangleq \sum_{y=(\hat{K}-1)\tau+1}^{\hat{K}\tau} \Upsilon(\hat{K}, y) + \|\zeta(\mathbf{w}_S(y-1))\|$ . Using the assumptions from Lemma 2 of [23] and the fact that  $t \geq \hat{K}\tau$ , the RHS of (39) is strictly positive, implying that  $\hat{\Upsilon}(\hat{K}) \ll 1$  as  $t\xi\eta(1 - \frac{\beta\eta}{2})$  is very small. Next, defining  $\varrho(t) \triangleq F(\mathbf{w}_S(t)|\mathcal{D}_N(t)) - F(\mathbf{w}^*(t)|\mathcal{D}_N(t)) \geq \epsilon$ , we get:

$$\begin{aligned} & \varrho(t)^{-1} - \frac{1}{\theta_{\hat{K}+1}(t)} \\ & = \frac{F(\mathbf{v}_{\hat{K}}(t)|\mathcal{D}_N(t)) - F(\mathbf{w}_S(t)|\mathcal{D}_N(t))}{\theta_{\hat{K}+1}(t)\varrho(t)} \\ & \geq \frac{-L}{\beta\epsilon^2} \sum_{y=(\hat{K}-1)\tau+1}^{\hat{K}\tau} \Upsilon(\hat{K}, y) + \|\zeta(\mathbf{w}_S(y-1))\|. \end{aligned} \quad (40)$$

Combining (39) and (40), and since  $\theta_1(0) > 0$ , we get  $\varrho(t)^{-1} \geq t\xi\eta(1 - \frac{\beta\eta}{2}) - \frac{(\hat{K}+1)L}{\beta\epsilon^2} \hat{\Upsilon}(\hat{K})$ , taking the reciprocal of which leads to (29).

## APPENDIX C

### PROOF OF PROPOSITION 1

Since  $\nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t))$  is the average of  $\nabla F(\mathbf{w}_S(t), x_d, y_d)$ ,  $\forall (x_d, y_d) \in \mathcal{D}_i(t)$ , we apply the central limit theorem to view  $\nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t))$  as  $\mathcal{D}_i(t)$  samples of  $\nabla F(\mathbf{w}_S(t), x_d, y_d)$  from a distribution with mean  $\nabla F(\mathbf{w}_S(t)|\mathcal{D}_N(t))$ . Then, (31) can be upper bounded using the definition in (1):

$$\begin{aligned} & \|\nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t)) - \nabla F(\mathbf{w}_S(t)|\mathcal{D}_N(t)) - \zeta(\mathbf{w}_S(t))\| \\ & \leq \|\zeta(\mathbf{w}_S(t))\| + \frac{\gamma}{\sqrt{D_i(t)}} \end{aligned}$$

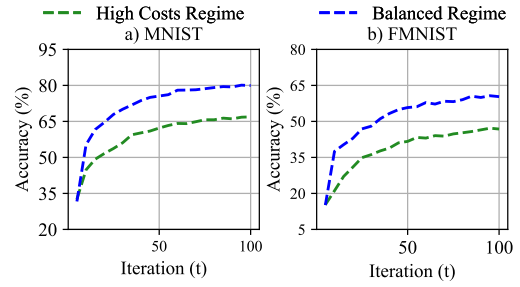


Fig. 13. Comparing the performance of our methodology under the high energy cost regime versus the standard balanced cost regime for MNIST and FMNIST for a network with  $N = 200$  and  $S = 3$ .

TABLE II

OPTIMIZATION VALUES OF  $(\mathcal{P}_D)$  FOR A NETWORK WITH  $N = 200$  AND  $S = 3$  ON MNIST AND FMNIST. WE EXAMINE THE VALUES OF THE UN-SCALED TERMS OF  $(\mathcal{P}_D)$ , I.E., THE AVERAGED ESTIMATED FL LOSS, THE DATA PROCESSING ENERGY, AND THE DATA TRANSMISSION ENERGY, FOR BOTH THE BALANCED REGIME AND THE HIGH ENERGY COSTS REGIME

	MNIST		FMNIST	
	Balanced	High Costs	Balanced	High Costs
Estimated Loss	155.54	156.05	155.76	156.05
Data Processing Energy	120.79	99.87	148.36	142.03
Data Transmission Energy	188.84	0	128.47	0

$$\begin{aligned} & \leq \left\| \frac{-1}{D_N(t)} \sum_{i \in \mathcal{S}} D_i(t) \nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t)) \right. \\ & \quad \left. + \frac{D_N(t) - D_S(t)}{D_N(t)D_S(t)} \sum_{i \in \mathcal{S}} D_i(t) \nabla F(\mathbf{w}_S(t)|\mathcal{D}_i(t)) \right\| + \frac{\gamma}{\sqrt{D_i(t)}}. \end{aligned} \quad (41)$$

Applying the triangle inequality on the above gives the result.

## APPENDIX D

### HIGH ENERGY COSTS REGIME ON PERFORMANCE

We investigate the behavior of our methodology in a scenario with energy restrictions, and compare it relative to the balanced regime. In the balanced regime, the three objective function terms in  $(\mathcal{P}_D)$  are expected to have a similar degree of importance, which we can emulate by having similar order of magnitude via careful choice of  $\alpha$ ,  $\beta$ , and  $\gamma$ . By contrast, if devices were disconnected from stable power, they would have energy limitations, and therefore the energy cost for data processing and transmissions is likely to be more restrictive. In response, a network may increase  $\beta$  and  $\gamma$  by an order of magnitude, thus entering the high energy cost regime.

We examine these two regimes in Fig. 13 and Table II. For the balanced regime, we employ  $\alpha = 100$ ,  $\beta = 0.001$ , and  $\gamma = 0.01$  for MNIST or  $\gamma = 0.006$  for FMNIST as in Sec. V-B, while, for the high energy costs regime, we will use  $\alpha = 100$ ,  $\beta = 0.01$ , and  $\gamma = 0.1$  for MNIST or  $\gamma = 0.06$  for FMNIST.

As the high energy cost regime places greater emphasis on reducing the data processing and D2D data offloading, we see a reduction in the data processing and transmission energies in Table II for the high energy cost regime relative to the balanced regime. This reduced data processing and transmission leads



TABLE III  
EXAMINING THE IMPACT OF TIME-VARYING INACTIVE LINKS  
(I.E., ADJACENCY MATRIX CHANGES OVER TIME) ON THE  
PERFORMANCE OF THE PROPOSED SAMPLING WITH D2D  
DATA OFFLOADING METHODOLOGY FOR A  
NETWORK WITH  $N = 200$  AND  $S = 3$

Link Failure Rate	MNIST		FMNIST	
	Final Acc (%)	Links Avail- able	Final Acc (%)	Links Avail- able
0%	79.93	7694	60.32	7699
25%	77.96	5706	59.19	5714
50%	73.53	3788	57.97	3792
75%	73.26	1891	53.16	1893
100%	66.67	0	47.94	0

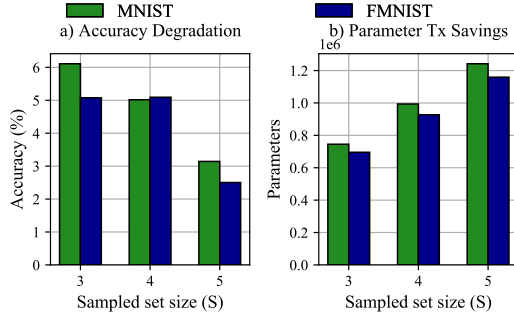


Fig. 14. Examining the effectiveness of FedDrop [70] integrated with our proposed device sampling methodology on both MNIST and FMNIST datasets for a network with  $N = 100$ . FedDrop has yields a small decrease in the final classification accuracies as well as a significant decrease in the number of parameters that are transmitted from the sampled devices to the server.

to a corresponding increase in the estimated loss in Table II, which is reflected by the decreased classification accuracies shown in Fig. 13 for the high energy cost regime relative to the balanced regime.

#### APPENDIX E IMPACT OF TIME-VARYING LINKS

We examine the ability of our methodology to adapt to networks with time varying links (i.e., changing adjacency matrices) in Table III. This experiment randomly deactivates links within the adjacency matrix  $\mathbf{A}$  such that at any given iteration  $t$ , each  $i, j$ -th link will be deactivated, i.e.,  $\mathbf{A}_{i,j}(t) = 0$ , with probability 0%, 25%, 50%, 75%, or 100%. From Table III, higher link failure rate does reduce the classification performance for the proposed methodology, which makes sense as there are fewer links available and thus fewer D2D data offloading opportunities. Nonetheless, the resulting accuracies in the scenarios involving time-varying links are quite close to the case with fixed links. For both MNIST and FMNIST, the gap between 0% link failure rate and 75% link failure rate is within 7%, approximately. That being said, the case with total link failures (i.e., 100% link failure rate) has over 6% reduced final classification accuracy from the case with 75% link failure rate. Thus, we can see that D2D data offloading has substantial impact on the resulting system performance.

#### APPENDIX F EXAMINING INTEGRATION OF FEDDROP [70]

In the following, we examine the impact of federated dropout (FedDrop from [70]), a methodology designed to

dropout model parameters from the fully connected layers of ML models in federated settings, on our device sampling methodology. We use a dropout rate of 75% for MNIST and 70% for FMNIST on the fully connected layers of our CNNs. As our experiments rely on CNNs with 21840 total parameters of which 16560 parameters are for the fully connected layers, these dropout settings yield device-to-server ML model transmission savings of roughly 43% and 46% for MNIST and FMNIST, respectively. The classification performance results in Fig. 14 show that FedDrop integrated with our methodology yields a substantial decrease in total ML model parameters transmitted across all global aggregations. Moreover, these substantial resource savings only induce a minor decrease, under roughly 6%, in the final classification accuracies. Thus, in federated scenarios with device-to-server communication limitations, our methodology combined with FedDrop offers significant communication overhead resource savings.

#### REFERENCES

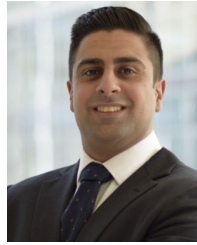
- [1] S. Wang, M. Lee, S. Hosseinalipour, R. Morabito, M. Chiang, and C. G. Brinton, "Device sampling for heterogeneous federated learning: Theory, algorithms, and implementation," in *Proc. INFOCOM IEEE Conf. Comput. Commun.*, May 2021, pp. 1–10.
- [2] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2923–2960, 4th Quart., 2018.
- [3] M. A. Alsheikh, "Five common misconceptions about privacy-preserving Internet of Things," *IEEE Commun. Mag.*, vol. 6, no. 5, pp. 151–157, May 2023.
- [4] S. Heinrich. (2017). *Flash Memory in the Emerging Age of Autonomy*. Accessed: Aug. 8, 2020. [Online]. Available: [https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808\\_FT12\\_Heinrich.pdf](https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808_FT12_Heinrich.pdf)
- [5] M. Wu, F. R. Yu, and P. X. Liu, "Intelligence networking for autonomous driving in beyond 5G networks with multi-access edge computing," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 5853–5866, Jun. 2022.
- [6] X. Li, G. Feng, Y. Sun, S. Qin, and Y. Liu, "A unified framework for joint sensing and communication in resource constrained mobile edge networks," *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 5643–5656, Oct. 2023.
- [7] A. Gupta, S. Misra, N. Pathak, and D. Das, "FedCare: Federated learning for resource-constrained healthcare devices in IoMT system," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1587–1596, Aug. 2023.
- [8] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2017, pp. 1273–1282.
- [9] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.
- [10] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "A novel framework for the analysis and design of heterogeneous federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 5234–5249, 2021.
- [11] S. Lu, Z. Gao, Q. Xu, C. Jiang, A. Zhang, and X. Wang, "Class-imbalance privacy-preserving federated learning for decentralized fault diagnosis with biometric authentication," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9101–9111, Dec. 2022.
- [12] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 46–51, Jun. 2020.
- [13] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated to fog learning: Distributed machine learning over heterogeneous wireless networks," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 41–47, Dec. 2020.
- [14] S. Zehtabi, S. Hosseinalipour, and C. G. Brinton, "Decentralized event-triggered federated learning with heterogeneous communication thresholds," in *Proc. IEEE 61st Conf. Decis. Control (CDC)*, Sep. 2022, pp. 4680–4687.
- [15] C. Chang et al., "BEV-V2X: Cooperative birds-eye-view fusion and grid occupancy prediction via V2X-based data sharing," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 11, pp. 4498–4514, Nov. 2023.

- [16] S. Wang et al., "Towards cooperative federated learning over heterogeneous Edge/Fog networks," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 54–60, Dec. 2023.
- [17] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.
- [18] S. Wang, S. Hosseinalipour, M. Gorlatova, C. G. Brinton, and M. Chiang, "UAV-assisted online machine learning over multi-tiered networks: A hierarchical nested personalized federated learning approach," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 2, pp. 1847–1865, Jun. 2023.
- [19] S. Hosseinalipour et al., "Parallel successive learning for dynamic distributed model training over heterogeneous wireless networks," *IEEE/ACM Trans. Netw.*, vol. 32, no. 1, pp. 222–237, Feb. 2024.
- [20] T. Huang, B. Ye, Z. Qu, B. Tang, L. Xie, and S. Lu, "Physical-layer arithmetic for federated learning in uplink MU-MIMO enabled wireless networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Jul. 2020, pp. 1221–1230.
- [21] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-IID data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Nov. 2019.
- [22] W. Sun, S. Lei, L. Wang, Z. Liu, and Y. Zhang, "Adaptive federated learning and digital twin for industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5605–5614, Aug. 2020.
- [23] S. Wang et al., "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.
- [24] T. Wu et al., "Joint edge aggregation and association for cost-efficient multi-cell federated learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Sep. 2023, pp. 1–10.
- [25] S. Ji, W. Jiang, A. Walid, and X. Li, "Dynamic sampling and selective masking for communication-efficient federated learning," *IEEE Intell. Syst.*, vol. 37, no. 2, pp. 27–34, Mar./Apr. 2021.
- [26] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5132–5143.
- [27] Y. J. Cho, J. Wang, and G. Joshi, "Towards understanding biased client selection in federated learning," in *Proc. 25th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Mar. 2022, pp. 10351–10375.
- [28] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2020.
- [29] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2019, pp. 1387–1395.
- [30] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," 2018, *arXiv:1812.06127*.
- [31] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2021–2031.
- [32] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, Jan. 2020.
- [33] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.
- [34] W. Xia, T. Q. S. Quek, K. Guo, W. Wen, H. H. Yang, and H. Zhu, "Multi-armed bandit based client scheduling for federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7108–7123, Nov. 2020.
- [35] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, Nov. 2020.
- [36] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-IID data with reinforcement learning," in *Proc. IEEE Conf. Comput. Commun.*, Aug. 2020, pp. 1698–1707.
- [37] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–26.
- [38] Z. Su et al., "Secure and efficient federated learning for smart grid with edge-cloud collaboration," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 1333–1344, Feb. 2022.
- [39] S. Wang, Y. Ruan, Y. Tu, S. Wagle, C. G. Brinton, and C. Joe-Wong, "Network-aware optimization of distributed learning for fog computing," *IEEE/ACM Trans. Netw.*, vol. 29, no. 5, pp. 2019–2032, Oct. 2021.
- [40] X. Lin, J. Wu, J. Li, X. Zheng, and G. Li, "Friend-as-learner: Socially-driven trustworthy and efficient wireless federated edge learning," *IEEE Trans. Mobile Comput.*, vol. 22, no. 1, pp. 269–283, Jan. 2023.
- [41] F. Dou, J. Lu, T. Zhu, and J. Bi, "On-device indoor positioning: A federated reinforcement learning approach with heterogeneous devices," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 3909–3926, Feb. 2024.
- [42] J. Zhang, W. Liu, Y. He, Z. He, and M. Guizani, "Semi-asynchronous model design for federated learning in mobile edge networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 16280–16292, Dec. 2023.
- [43] Y. Qu, C. Dong, T. Wu, Y. Zhuang, H. Dai, and F. Wu, "Efficient edge intelligence under clustering for UAV swarm networks," in *Proc. Int. Conf. Space-Air-Ground Comput. (SAGC)*, Oct. 2021, pp. 112–117.
- [44] J. Xie, D. Guo, X. Shi, H. Cai, C. Qian, and H. Chen, "A fast hybrid data sharing framework for hierarchical mobile edge computing," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Jul. 2020, pp. 2609–2618.
- [45] X. Xia et al., "OL-MEDC: An online approach for cost-effective data caching in mobile edge computing systems," *IEEE Trans. Mobile Comput.*, vol. 22, no. 3, pp. 1646–1658, Mar. 2023.
- [46] Y. Li et al., "Energy-aware, device-to-device assisted federated learning in edge computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 34, no. 7, pp. 2138–2154, Jul. 2023.
- [47] S. Savazzi, M. Nicoli, M. Bennis, S. Kianoush, and L. Barbieri, "Opportunities of federated learning in connected, cooperative, and automated industrial systems," *IEEE Commun. Mag.*, vol. 59, no. 2, pp. 16–21, Feb. 2021.
- [48] Z. Lin, H. Liu, and Y.-J. A. Zhang, "Relay-assisted cooperative federated learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7148–7164, Sep. 2022.
- [49] L. Barbieri, S. Savazzi, M. Brambilla, and M. Nicoli, "Decentralized federated learning for extended sensing in 6G connected vehicles," *Veh. Commun.*, vol. 33, Jan. 2022, Art. no. 100396.
- [50] B. Ganguly et al., "Multi-edge server-assisted dynamic federated learning with an optimized floating aggregation point," *IEEE/ACM Trans. Netw.*, vol. 31, no. 6, pp. 2682–2697, Dec. 2023.
- [51] Y. Guo, Y. Sun, R. Hu, and Y. Gong, "Hybrid local SGD for federated learning with heterogeneous communications," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–42.
- [52] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.
- [53] T. Jeske, "Floating car data from smartphones: What Google and Waze know about you and how hackers can control traffic," in *Proc. BlackHat Euro.*, Mar. 2013, pp. 1–12.
- [54] R. Morabito, R. Petrolo, V. Loscri, and N. Mitton, "LEGIOT: A lightweight edge gateway for the Internet of Things," *Future Gener. Comput. Syst.*, vol. 81, pp. 1–15, Apr. 2018.
- [55] K. Krishna and M. N. Murty, "Genetic K-means algorithm," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 29, no. 3, pp. 433–439, Jun. 1999.
- [56] K. Jacksi, R. Kh. Ibrahim, S. R. M. Zeebaree, R. R. Zebari, and M. A. M. Sadeeq, "Clustering documents based on semantic similarity using HAC and K-mean algorithms," in *Proc. Int. Conf. Adv. Sci. Eng. (ICOASE)*, Dec. 2020, pp. 205–210.
- [57] A. D. Ribas, J. G. Colonna, C. M. S. Figueiredo, and E. F. Nakamura, "Similarity clustering for data fusion in wireless sensor networks using k-means," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–7.
- [58] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [59] S. Schaible and J. Shi, "Fractional programming: The sum-of-ratios case," *Optim. Methods Softw.*, vol. 18, no. 2, pp. 219–229, Apr. 2003.
- [60] T. Kuno, "A branch-and-bound algorithm for maximizing the sum of several linear ratios," *J. Glob. Optim.*, vol. 22, nos. 1–4, pp. 155–174, 2002.
- [61] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, 2016.
- [62] S. Martello, D. Pisinger, and P. Toth, "New trends in exact algorithms for the 0–1 knapsack problem," *Eur. J. Oper. Res.*, vol. 123, no. 2, pp. 325–332, Jun. 2000.
- [63] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

- [64] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
- [65] Z. Li, Q. Chen, and V. Koltun, "Combinatorial optimization with graph convolutional networks and guided tree search," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 539–548.
- [66] *Dstat: Versatile Resource Statistics Tool*. Accessed: Aug. 8, 2020. [Online]. Available: <http://dag.wiee.rs/home-made/dstat/>
- [67] *The MNIST Database of Handwritten Digits*. Accessed: Aug. 8, 2020. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [68] *Fashion-MNIST*. Accessed: Aug. 8, 2020. [Online]. Available: <https://research.zalando.com/welcome/mission/research-projects/fashion-mnist/>
- [69] C. Li, X. Zeng, M. Zhang, and Z. Cao, "PyramidFL: A fine-grained client selection framework for efficient federated learning," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2022, pp. 158–171.
- [70] D. Wen, K.-J. Jeon, and K. Huang, "Federated dropout—A simple approach for enabling federated learning on resource constrained devices," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 923–927, May 2022.
- [71] S. Wang, S. Hosseinalipour, and C. G. Brinton, "Multi-source to multi-target decentralized federated domain adaptation," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 3, pp. 1011–1025, Jun. 2024.
- [72] S. Wagle, S. Hosseinalipour, N. Khosravan, M. Chiang, and C. G. Brinton, "Embedding alignment for unsupervised federated learning via smart data exchange," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2022, pp. 492–497.
- [73] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.



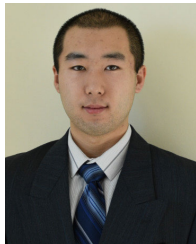
**Roberto Morabito** (Member, IEEE) received the Ph.D. degree in networking technology from Aalto University, Finland, in May 2019. He is an Assistant Professor with the Communication Systems Department, Eurecom, France. From June 2019 to March 2021, he was a Post-Doctoral Researcher with the Edge Laboratory, School of Electrical and Computer Engineering, Princeton University, USA. His work intersects the IoT, edge computing, and distributed AI, focusing on trade-offs in AI service provisioning and orchestration under computing and networking resource constraints.



**Seyyedali Hosseinalipour** (Member, IEEE) received the B.S. degree in electrical engineering from the Amirkabir University of Technology, Tehran, Iran, in 2015, and the M.S. and Ph.D. degrees in electrical engineering from North Carolina State University, NC, USA, in 2017 and 2020, respectively. He is currently an Assistant Professor with the Department of Electrical Engineering, University at Buffalo (SUNY). His research interests include the analysis of modern wireless networks, synergies between machine learning methods and fog computing systems, distributed machine learning, and network optimization.



**Mung Chiang** (Fellow, IEEE) is the President and a Roscoe H. George Distinguished Professor of electrical and computer engineering with Purdue University. Previously, he was an Arthur LeGrand Doty Professor with Princeton University and has founded Princeton Edge Laboratory. He is a member of the American Academy of Arts and Sciences, the National Academy of Inventors, and the Royal Swedish Academy of Engineering Sciences. He received the 2013 Alan T. Waterman Award.



**(Henry) Su Wang** received the B.S. (with distinction) and Ph.D. degrees in electrical and computer engineering from Purdue University, West Lafayette, in 2018 and 2023, respectively. He is currently a Post-Doctoral Research Associate with the Department of Electrical and Computer Engineering, Princeton University. His research interests lie at the intersection of machine intelligence and networking.



**Christopher G. Brinton** (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Princeton in 2013 and 2016, respectively, both in electrical engineering. He is the Elmore Rising Star Assistant Professor of electrical and computer engineering (ECE) with Purdue University. His research interest is at the intersection of networking, communications, and machine learning, specifically in fog/edge network intelligence, distributed machine learning, and AI/ML-inspired wireless network optimization.