# Connectivity-Aware Semi-Decentralized Federated Learning over Time-Varying D2D Networks

Rohit Parasnis*
Purdue University
West Lafayette, Indiana, USA
rparasni@purdue.edu

Seyyedali Hosseinalipour
University at Buffalo–SUNY
Buffalo, New York, USA
alipour@buffalo.edu

Yun-Wei Chu
Purdue University
West Lafayette, Indiana, USA
chu198@purdue.edu

Christopher G. Brinton
Purdue University
West Lafayette, USA
cgb@purdue.edu

Mung Chiang
Purdue University
West Lafayette, USA
chiang@purdue.edu

## ABSTRACT

Semi-decentralized federated learning blends the conventional device-to-server (D2S) interaction structure of federated model training with localized device-to-device (D2D) communications. We study this architecture over practical edge networks with multiple D2D clusters modeled as time-varying and directed communication graphs. Our investigation results in an algorithm that controls the fundamental trade-off between (a) the rate of convergence of the model training process towards the global optimizer, and (b) the number of D2S transmissions required for global aggregation. Specifically, in our semi-decentralized methodology, D2D consensus updates are injected into the federated averaging framework based on column-stochastic weight matrices that encapsulate the connectivity within the clusters. To arrive at our algorithm, we show how the expected optimality gap in the current global model depends on the greatest two singular values of the weighted adjacency matrices (and hence on the densities) of the D2D clusters. We then derive tight bounds on these singular values in terms of the node degrees of the D2D clusters, and we use the resulting expressions to design a threshold on the number of clients required to participate in any given global aggregation round so as to ensure a desired convergence rate. Simulations performed on real-world datasets reveal that our connectivity-aware algorithm reduces the total communication cost required to reach a target accuracy significantly compared with baselines depending on the connectivity structure and the learning task.

## CCS CONCEPTS

• **Computer systems organization → Distributed architectures**; **Peer-to-peer architectures**; • **Networks → Topology analysis and generation**.

## KEYWORDS

connectivity, semi-decentralized, federated learning

**ACM Reference Format:**
Rohit Parasnis, Seyyedali Hosseinalipour, Yun-Wei Chu, Christopher G. Brinton, and Mung Chiang. 2023. Connectivity-Aware Semi-Decentralized Federated Learning over Time-Varying D2D Networks. In *Proceedings of Submitted (MobiHoc '23).* ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Federated learning (FL) [14, 22] is a popular paradigm for distributing machine learning (ML) tasks over a network of centrally coordinated devices. By not requiring the devices to share any training data with the central coordinator (server), FL improves privacy and communication efficiency. The first FL technique, known as federated averaging (FedAvg), was proposed in [14, 22] as a distributed optimization algorithm for a "star" topology-based network architecture. In each iteration of the FedAvg algorithm, (i) devices individually performs a number of local stochastic gradient descent (SGD) iterations and transmit their cumulative stochastic gradients to the central server, which then (ii) aggregates a random subset of these gradients to estimate the globally optimal ML model. In recent years, several variants of FedAvg have been proposed to address the challenges encountered by FL at the wireless edge, including different dimensions of heterogeneity in dataset statistics (e.g., varying local data distributions) and in the network system itself (e.g., varying communication and computation capabilities).

An emerging arch of work has been exploring FL under edge networks that diverge from the star learning topology between the devices and the server. This had led to varying degrees of decentralization in FL, reaching fully decentralized, serverless settings that sit at the opposite extreme of the star topology [4, 10, 15, 16, 19, 29, 39]. In between these two extremes is *semi-decentralized FL*, where device-to-device (D2D) communications complement device-to-server (D2S) interactions [6, 9, 20, 38]. These D2D interactions occur locally within *clusters* of devices, with each cluster forming a connected component. In semi-decentralized FL, D2D transmissions are less energy-consuming than D2S interactions and can help reduce the frequency of D2S communications through localized synchronizations of the ML model updates.

Despite these recent investigations, we still do not have a clear understanding of how different D2D topology properties impact

*MobiHoc '23, June 03–05, 2023, Washington, DC*
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

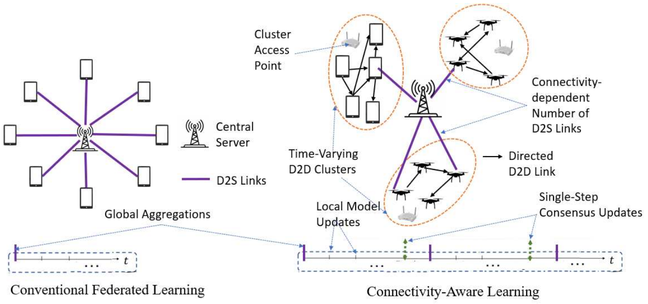
Cluster Access Point
Central Server
D2S Links
Global Aggregations
Time-Varying D2D Clusters
Connectivity-dependent Number of D2S Links
Directed D2D Link
Local Model Updates
Single-Step Consensus Updates
t
t
Conventional Federated Learning
Connectivity-Aware Learning

**Figure 1: Conventional Federated Learning vs. Connectivity-Aware Semi-Decentralized Learning Architecture**

the learning process. For instance, the ratio of the number of D2D interactions to that of D2S interactions will impact the training efficiency differently over different topologies. This becomes especially important in the presence of constraints such as upload/download bandwidths, and stochastic uncertainties such as data heterogeneity, client mobility, and communication link failures. On one hand, edge devices in clustered D2D networks that have little to no cross-cluster interactions are typically in contact with only a small fraction of the rest of the network at any given time instant (e.g., networks of unmanned aerial vehicle (UAV) swarms spread over geo-distributed regions separated by long distances). In such networks, if there is no central coordinator (implying zero D2S interactions) and if the training data are distributed heteregeneously among the edge devices, no practically feasible number of D2D interactions is likely to aggregate a set of local ML models that are diverse enough to approximate the global data distribution [3].

On the other hand, having a high number of D2D interactions is advantageous when D2S interactions take the form of high-latency, high-energy transmissions (e.g., if the UAV swarms in the previous example are miles away from the nearest base station). Moreover, classical star-topology-based FL architectures miss out on an important benefit of D2D cooperation: devices acting as information relays between other devices and the server, effectively sharing with the server more information than it would expect to receive.

We are thus motivated to conduct a formal study of semi-decentralized FL, and reveal the combined impact of D2S and D2D interactions on the training process. After building an understanding of the D2D topologies on which the D2D interactions occur, we propose a novel FL technique that enables us to take into account the degree distributions of the D2D clusters and use this knowledge to tune the number of expensive D2S transmissions while simultaneously ensuring a minimum rate of global training convergence. As shown in Fig. 1, we incorporate two scales of model aggregations: on the first scale, the edge devices perform intra-cluster model aggregations with their one-hop neighbors via distributed averaging, and on the second scale, a central server samples a random set of clients (as in the classical FedAvg architecture [22]) for global aggregation.

Our methodology has several potential use-cases, including the following that we will refer to as examples throughout the paper:

(1) ***UAV Networks for ISR:*** UAVs are being increasingly deployed for intelligence, surveillance, and reconnaissance (ISR) operations in defense settings [28, 33]. With the UAVs partitioned into D2D-enabled swarms deployed across different areas, our connectivity-aware algorithm can facilitate intra-cluster communications and reduce the over-reliance of the model training process on D2S transmissions.
(2) ***Self-driving cars:*** Many learning tasks for self-driving cars call for vehicles to communicate over short distances. In such settings, geographical proximity can be used to partition the traffic network into clusters, which would enable us to design intra-cluster D2S communications that turn out to be more efficient than D2S communications with a far-away server.

## 1.1 Summary of Contributions

Our key contributions are summarized below:

(1) ***Analysis with Time-Varying and Directed Cluster Topologies:*** We consider that each D2D cluster in general is a time-varying directed graph (digraph). We show how the expected optimality gap of the learning process depends on the greatest two singular values of the weighted adjacency matrices used for local aggregations in the clusters. Our analysis is applicable to edge networks with asymmetric D2D communications subjected to link failures.
(2) ***Singular Value Bounds in terms of Node Degrees:*** We derive bounds on the singular values of the cluster-specific weighted adjacency matrices in terms of the degree distribution of every cluster. This introduces new technical challenges as described in Section 1.2, since it is a stark departure from existing analyses of consensus-based FL algorithms that rely heavily on the spectral gaps of symmetric weight matrices (e.g., see [5, 9, 12, 13, 20, 36, 39]).
(3) ***Connectivity-Aware Learning Algorithm:*** We use our singular value bounds to design a time-varying threshold on the number of clients required to be sampled by the central server for global aggregation so as to enforce a desired convergence rate while simultaneously reducing the number of D2S communications. This tradeoff results in a novel connectivity-aware algorithm with significant energy savings, as validated subsequently by our numerical results.
(4) ***Effect of Data Heterogeneity under Mild Gradient Diversity Assumptions:*** We derive a bound on the expected optimality gap that captures the effects of cluster densities as well as the extent of data heterogeneity across the devices. In doing so, we employ a milder definition of gradient diversity [20] than what is typically assumed in literature.

***Notation:*** We denote the set of real numbers by $\mathbb{R}$ and the set of positive integers by $\mathbb{N}$. For any $n \in \mathbb{N}$, we define $[n] := \{1, 2, \ldots, n\}$. For a finite set $S$, we denote its cardinality by $|S|$.

We denote the vector space of $n$-dimensional real-valued column vectors by $\mathbb{R}^n$. We use the superscript notation $^\top$ to denote the transpose of a vector or a matrix. All matrix and vector inequalities are assumed to hold entry-wise. We use $I$ to denote the identity matrix (of the known dimension) and $\mathbf{1}$ to denote the column vector (of the known dimension) that has all entries equal to 1. Similarly, $\mathbf{0}$ denotes the all-zeroes vector. In addition, we use $\|\cdot\|$ to denote the Euclidean norm of a square matrix or a vector, and for any vector $v \in \mathbb{R}^n$ we use $\mathrm{diag}(v)$ to denote the diagonal matrix whose $i$-th diagonal entry is $v_i$.

We say that a vector $v \in \mathbb{R}^n$ is *stochastic* if $v \geq 0$ and $v^\top \mathbf{1} = 1$, and a matrix $A$ is *column-stochastic* if $A$ is non-negative and if each column of $A$ sums to 1, i.e., if $A \geq 0$ and $A^\top \mathbf{1} = \mathbf{1}$.

## 1.2 Related Work

Several different FL approaches with varying degrees of decentralization have been proposed to date. In this section, we focus on those which are most relevant to the present work.

***Semi-decentralized FL:*** [20] proposes a semi-decentralized learning methodology in which the D2D network is partitioned into clusters, as in our paper. The key differences between [20] and the present work are (a) we do not assume the D2D communications to be bidirectional (equivalently, the cluster graphs in our model are not undirected), and (b) our analysis uses column-stochastic consensus matrices that need not satisfy the standard but unrealistic assumption of symmetry (which leads to double stochasticity and may not hold if the cluster graphs are directed). This leads to two significant technical challenges. First, we cannot use standard eigenvalue results in our analysis since we must focus on singular values, which generally differ from eigenvalues for asymmetric matrices. Second, unlike doubly stochastic matrices, column-stochastic aggregation matrices in general do not ensure convergence to consensus in the absence of a central coordinator, which means our analysis must account for the combined effect of global aggregations and column-stochasticity. We address these challenges in this work.

Another closely related semi-decentralized learning methodology is [38]. In [38], the goal is to enable edge devices to compute weighted sums of their neighbors' scaled cumulative gradients in order to reduce the dependence of the global training process on unreliable D2S links. [38], however, assumes the D2D communication network to be time-invariant and undirected, thereby disregarding potential communication link failures and client mobility.

***Learning over Clustered D2D Networks:*** Recently, [2] proposed fully decentralized learning over D2D networks in which a small subset of nodes act as bridges between different clusters for cross-cluster model transmission, thereby obviating the role of a server. Their topology design, however, results in a static rather than a dynamic D2D network. Reference [6] also focuses on clustered networks, but it provides a semi-decentralized learning methodology where the basis for clustering is data similarity, whereas our methodology makes no assumptions on the basis for clustering. A complementary approach is proposed in [3], where every cluster is assumed to be a clique and the D2D network is partitioned in such a way that each local dataset is representative of the global data. Network clusters also form the focus of another recent work, [30], which proposes having one edge server per cluster so as to eliminate the need for a central server. Its learning algorithm assumes the edge network topology to be undirected, which gives rise to a symmetric adjacency matrix.

***Other Consensus-based Algorithms:*** Reference [12] provides improved bounds on the convergence rates of certain gradient tracking methods used in decentralized learning by enhancing the analysis of the consensus matrix (referred to as the *mixing matrix* therein) and its spectral gap. However, similar to [20], this work assumes the consensus matrix to be row-stochastic as well as symmetric, and hence, doubly stochastic. In this respect, [7] relaxes the assumptions of symmetry as well as double stochasticity in an online learning setting. However, the matrices therein are row-stochastic, which are not average-preserving and hence, they are not as suitable as column-stochastic matrices for minimizing the average of all the local loss functions. Finally, we remark that there exists abundant literature on distributed optimization over time-varying digraphs characterized by consensus matrices that are not necessarily doubly stochastic (e.g., see [1, 18, 24–26, 32, 35]). However, the effects of both data heterogeneity (or non-i.i.d. data distribution) and graph-theoretic properties (such as the degree distribution of the network in question) on the convergence rate of these algorithms have remained largely unexplored.

# 2 SEMI-DECENTRALIZED FL SETUP

We now introduce the system model, the learning objective, and the network model in semi-decentralized FL.

## 2.1 System Model and Learning Objectives

We consider a collaborative learning environment consisting of $n$ edge devices, or *clients*, and a central parameter server (PS) that is tasked with aggregating all the local model updates generated by the clients. We use $[n]$ to denote the set of clients.

Each client $i \in [n]$ has a local dataset $\mathcal{D}_i$, which is a collection of data samples of the form $\xi = (u, y)$ where $u \in \mathbb{R}^p$ is the *feature vector* of the sample and $y$ is its *label*. On this basis, for any model $x \in \mathbb{R}^p$, we define the *loss function* $L : \mathbb{R}^p \times \cup_{i=1}^n \mathcal{D}_i \to \mathbb{R}$ so that $L(x; \xi)$ denotes the loss incurred by $x$ on a sample $\xi \in \cup_{i=1}^n \mathcal{D}_i$ (where $\cup_{i=1}^n \mathcal{D}_i$ is the global dataset). The average loss incurred by $x$ over the local dataset of client $i$ is given by $f_i(x) := \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} L(x; \xi)$, where $f_i : \mathbb{R}^p \to \mathbb{R}$ denotes the *local loss function* of client $i$.

In collaboration with the PS, the clients seek to minimize the *global loss function* $f : \mathbb{R}^p \to \mathbb{R}$, defined as the unweighted arithmetic mean $f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x)$ of all the local loss functions. The learning objective, therefore, is to determine the *global optimum* $x^* := \arg\min_{x \in \mathbb{R}^p} f(x)$.

## 2.2 D2D and D2S Network Models

We model two types of interactions among the network elements: (i) D2S and (ii) D2D. For D2S interactions, the devices can engage in uplink communications to the PS if prompted by the server, which happens through a sampling procedure explained later.

We model the D2D network as a time-varying directed graph $G(t) = ([n], E(t))$, where $[n]$ denotes the *vertex set* and $E(t)$ the *edge set* of the digraph. The existence of a directed edge from a node $i \in [n]$ to another node $j \in [n]$ in $G(t)$ denotes the existence of a communication link from the $i$-th client to the $j$-th client in the D2D network. In this case, we refer to client $i$ (respectively, client $j$) as the in-neighbor (respectively, out-neighbor) of client $j$ (respectively, client $i$). The set of in-neighbors (respectively, out-neighbors) of a client $i \in [n]$ at time $t$ is denoted by $\mathcal{N}_i^-(t)$ (respectively, $\mathcal{N}_i^+(t)$). The number of in-neighbors (respectively, out-neighbors) is called the in-degree (respectively, out-degree) and is denoted by $d_i^-(t)$ (respectively, $d_i^+(t)$). We let $d_{\max}^-(t)$, $d_{\min}^+(t)$, and $d_{\max}^+(t)$ denote the maximum in-degree, the minimum out-degree, and the maximum out-degree, respectively.

Unlike standard works on distributed learning [1, 24–26], we do not assume the D2D network to be strongly connected or even uniformly strongly connected [24, 25] over time. This gives rise to a number $c > 1$ of strongly connected components of $G(t)$, denoted $\{(V_1(t), E_1(t)), (V_2(t), E_2(t)), \ldots, (V_c(t), E_c(t))\}$ which we refer to as *clusters* of the D2D network. Here, we make the following mild assumptions that apply to many cellular networks:

(1) The number of clusters, $c$, is time-invariant.
(2) There does not exist any communication link between any two clusters. In other words, $E(t) = \cup_{\ell=1}^{c} E_\ell(t)$.
(3) Regardless of any movement of clients from one cluster to another over time, as of time $t$, the server has full knowledge of the vertex sets $\{V_\ell(t)\}_{\ell=1}^{c}$ of all the $c$ clusters.

The third condition is satisfied in practice since the base station (which acts as the PS) is aware of the users in its coverage area.

## 3 PROPOSED METHOD FOR CONNECTIVITY-AWARE LEARNING

We now present our methodology for connectivity-aware learning over the semi-decentralized setup from Sec. 2. Our technique will enable the central server to use limited knowledge of the cluster degree distributions to tune a communication-efficiency trade-off.

### 3.1 Local Model Updates

As in many FL schemes, we assume every client performs multiple rounds of local SGD iterations between any two consecutive rounds of global aggregation. Let $x^{(t)}$ denote the global model that all the clients possess at the end of the $t$-th round of global aggregation. Then, each client $i \in [n]$ performs $T \in \mathbb{N}$ iterations of local SGD. In other words, for each $k \in \{0, 1, \ldots, T-1\}$, we have

$$x_i^{(t,k+1)} = x_i^{(t,k)} - \eta_t \widetilde{\nabla} f_i(x_i^{(t,k)}), \tag{1}$$

where $\eta_t > 0$ is the learning rate or the step-size, and $\widetilde{\nabla} f_i(x) := \frac{1}{|\chi_i|}\sum_{\xi \in \chi_i} L(x; \xi)$ is the stochastic gradient computed by client $i$ by sampling a *mini-batch* or a random subset $\chi_i \subset \mathcal{D}_i$ of its local samples. Note that $x_i^{(t,0)} := x^{(t)}$.

### 3.2 Intra-Cluster Model Aggregations

The next step involves all the clients aggregating their scaled cumulative gradients with their neighbors. This aggregation takes the form of weighted sums. Every client $i \in [n]$ first transmits its *scaled cumulative stochastic gradient* $x_i^{(t,T)} - x^{(t)} = -\eta_t \sum_{k=0}^{T-1} \widetilde{\nabla} f_i(x_i^{(t,k)})$ to each of its out-neighbors $j \in \mathcal{N}_i^+(t)$ before the $t$-th global aggregation round. To facilitate this, we assume that every cluster $\ell \in [c]$ contains an access point to which every client $i \in V_\ell(t)$ sends a list of its in-neighbors (clients whose gradients $i$ has received). The access point then announces the end of the concerned D2D communication round, determines the out-degree sequence $\{d_j^+(t) : j \in V_\ell(t)\}$ of the cluster, and broadcasts this sequence to every client in the cluster.

Subsequently, the client computes the following weighted sum of all the scaled cumulative gradients it receives from its in-neighbors:

$$\Delta_i(t) = \sum_{j \in \mathcal{N}_i^-(t)} \frac{1}{d_j^+(t)} \left( x_j^{(t,T)} - x^{(t)} \right). \tag{2}$$

This rule can be expressed compactly in matrix form as

$$\Delta(t) = A(t) X_{\text{diff}}^\top(t), \tag{3}$$

where $\Delta(t) := [\Delta_1(t) \;\; \Delta_2(t) \;\; \cdots \;\; \Delta_n(t)]^\top$, $X_{\text{diff}}(t) := \left[ x_1^{(t,T)} - x^{(t)} \;\; x_2^{(t,T)} - x^{(t)} \;\; \cdots \;\; x_n^{(t,T)} - x^{(t)} \right]$, and $A(t) \in \mathbb{R}^{n \times n}$ is a matrix whose $(i, j)$-th entry equals $a_{ij}(t) = \frac{1}{d_j^+(t)}$ for all $i, j \in [n]$.

Fact 1. *$A(t)$ is a column-stochastic matrix because the following holds for all $j \in [n]$:*

$$\sum_{i=1}^{n} a_{ij}(t) = \sum_{i \in [n] : j \in \mathcal{N}_i^-(t)} \frac{1}{d_j^+(t)} = \sum_{i \in \mathcal{N}_j^+(t)} \frac{1}{|\mathcal{N}_j^+(t)|} = 1.$$

*It can be verified that $A(t)$ is a block-diagonal matrix with its blocks $\{A_\ell(t)\}_{\ell=1}^{c}$ being the equal-neighbor adjacency matrices of the $c$ clusters in the D2D network.*

Henceforth, we refer to $A(t)$ as the *equal-neighbor adjacency matrix* of $G(t)$ because it represents every client $i \in [n]$ transmitting an equal share (a fraction $\frac{1}{d_i^+(t)}$) of its scaled cumulative gradient to its $d_i^+(t)$ out-neighbors.

### 3.3 Global Aggregation at the PS

For the global aggregation step, the PS samples a random subset of clients $\mathcal{S}(t) \subset [n]$. The cardinality $m(t) \leq n$ of this set is carefully chosen by our algorithm such that the resulting number of D2S interactions is just enough to complement the intra-cluster aggregations without excessively slowing down the training process.

Specifically, this involves three broad steps: (a) The PS first learns the degree distribution of each cluster. (b) It then computes an upper bound on an error quantity $\phi(t)$ that captures the combined effect of random sampling and the cluster degree distributions on the convergence rate. (c) It computes the minimum value of $m(t)$ required to keep $\phi(t)$ below a desired threshold. More specifically:

(1) For the $(t+1)$-th round of global aggregation, the server uses $m(t)$ (computed in the previous iteration) to select $\left\lceil \left( \frac{m(t)}{n} \right) n_\ell(t) \right\rceil$ clients uniformly at random from the $n_\ell(t) := |V_\ell(t)|$ clients that constitute cluster $\ell \in [c]$. This ensures that every cluster has a representation in the global aggregation that is proportionate to its size. The resulting set of randomly sampled clients is denoted by $\mathcal{S}(t)$. The server then updates the global model as follows:

$$x^{(t+1)} = x^{(t)} + \frac{1}{m(t)} \sum_{i \in \mathcal{S}(t)} \Delta_i(t) = x^{(t)} + \frac{1}{m(t)} \sum_{i=1}^{n} \tau_i(t) \Delta_i(t), \tag{4}$$

where $\tau_i(t) := |\{i\} \cap \mathcal{S}(t)|$ is an indicator random variable that takes the value 1 when client $i$ is sampled and the value 0 otherwise. Note that $\sum_{i=1}^{n} \tau_i(t) = |\mathcal{S}(t)| = m(t)$.

(2) The current round is now $t \longleftarrow t + 1$. All the cluster access points send their respective out-degree sequences to the server. Using this information, the server computes $\alpha_\ell(t) := \frac{1}{n_\ell(t)} \min_{i \in V_\ell(t)} d_i^+(t)$, the *minimum out-degree fraction* of cluster $\ell \in [c]$. The server then uses either of the two sets of singular value bounds that we later derive in Sec.

5 (either (10)-(11) or (15)-(16)) to compute an upper bound $\psi(m(t), \alpha_1(t), \ldots, \alpha_c(t))$ on the *connectivity factor* affecting the convergence rate. This connectivity factor is defined as

$$\phi(t) := \left(\frac{n}{m(t)} - 1\right) \sum_{\ell=1}^{c} \frac{n_\ell(t)}{n} \phi_\ell(t), \tag{5}$$

where $\phi_\ell(t) := \sigma_1^2(A_\ell(t)) + \sigma_2^2(A_\ell(t)) - 1$ depends on the greatest two singular values $\sigma_1(A_\ell(t)) \geq \sigma_2(A_\ell(t))$ of the equal-neighbor adjacency matrix $A_\ell(t)$ of cluster $\ell$. For the upper bound, we will show that

$$\psi(m(t), \alpha_1(t), \ldots, \alpha_c(t)) = \left(\frac{n}{m(t)} - 1\right) \sum_{\ell=1}^{c} \frac{n_\ell(t)}{n} \psi_\ell(t), \tag{6}$$

where either of the following holds (with the indexing $(t)$ on the right hand side omitted for brevity):

$$\begin{aligned} \psi_\ell(t) &= 1 + \varepsilon_\ell + \left(\frac{1}{\alpha_\ell} - 1\right)^2 + 2\varepsilon_\ell \left(1 + \frac{2}{\alpha_\ell} - \frac{1}{\alpha_\ell^2}\right), \\ \psi_\ell(t) &= 2 + 2\varphi_\ell \\ &\quad - \frac{(1-\varepsilon_\ell)^2(1-\alpha_{-\ell}^2)\left((1-\varepsilon_\ell)^2(1-\alpha_{-\ell}^2) - \alpha_{-\ell}\right)}{n_\ell(\varepsilon_{\text{net},\ell}+1)\left(\varepsilon_{\text{net},\ell} - \alpha_{-\ell} + \frac{1}{\alpha_\ell n_\ell}\right)} \end{aligned} \tag{7}$$

with $\varepsilon_\ell(t) := \frac{d^+_{\max}(t) - d^+_{\min}(t)}{d^+_{\min}(t)}$, $\varphi_\ell(t) := \frac{d^-_{\max}(t) - d^+_{\min}(t)}{d^+_{\min}(t)}$, $\alpha_{-\ell}(t) := \frac{1}{\alpha_\ell(t)} - 1$ and $\varepsilon_{\text{net},\ell}(t) = \varphi_\ell(t) + \frac{\varepsilon_\ell(t)}{\alpha_\ell(t)}$.

(3) Finally, the server sets

$$m(t+1) := \min\left\{r \in [n] : \psi(r, \alpha_1(t+1), \ldots, \alpha_c(t+1)) \leq \phi_{\max}\right\}$$

where $\phi_{\max}$ is a threshold given as an input to the algorithm. This step ensures that $\phi(t)$ remains below the threshold $\phi_{\max}$, thereby preserving the convergence rate.

Our algorithm for $t_{\max}$ global rounds is summarized in Alg. 1.

# 4 CONVERGENCE ANALYSIS

We now provide theoretical performance guarantees for Algorithm 1. We also explain how the effect of D2D cluster connectivity on the convergence rate of the algorithm is captured by the singular values of the equal-neighbor adjacency matrices of the clusters. **All the proofs except that of Proposition 5.1 are available in the appendix**.

## 4.1 Assumptions and Preliminaries

*4.1.1 Loss Functions.* We start by making the following standard assumptions on the local loss functions:

Assumption 1 (**Strong Convexity**). *All the local loss functions $\{f_i\}_{i=1}^n$ are $\mu$-strongly convex, i.e., there exists $\mu > 0$ such that $(\nabla f_i(x) - \nabla f_i(y))^\top (x - y) \geq \mu \|x - y\|^2$ for all $x, y \in \mathbb{R}^p$ and all $i \in [n]$.*

Assumption 2 (**Smoothness**). *All the local loss functions $\{f_i\}_{i=1}^n$ are $\beta$-smooth, i.e., there exists a finite $\beta$ such that $\|\nabla f_i(x) - \nabla f_i(y)\| \leq \beta \|x - y\|$ for all $x, y \in \mathbb{R}^p$ and all $i \in [n]$.*

As shown in [20], Assumptions 1 and 2 imply that the global loss function $f$ is both $\mu$-strongly convex and $\beta$-smooth.

**Algorithm 1** Connectivity-Aware Semi-Decentralized Learning

```
Input: $n$, $c$, $T$, $\phi_{\max}$, $t_{\max}$, $m(0)$, $\{n_\ell(t)\}_{t=0}^{t_{\max}}$, $x^{(0)}$
Output: $x^{(t_{\max})}$
for $t \in \{0, 1, \ldots, t_{\max} - 1\}$ do
    Client $i \in [n]$ sets $x_i^{(t,0)} \leftarrow x^{(t)}$
    for $k \in \{0, 1, \ldots, T - 1\}$ do
        Client $i \in [n]$ computes $x_i^{(t,k+1)} \leftarrow x_i^{(t,k)} - \eta_t \widetilde{\nabla} f_i(x_i^{(t,k)})$
    end
    Client $i \in [n]$ transmits its scaled cumulative local gradient
    $-\eta_t \sum_{k=0}^{T-1} \widetilde{\nabla} f_i(x_i^{(t,k)}) = x_i^{(t,T)} - x^{(t)}$ to its out-neighbors $\mathcal{N}_i^+(t)$
    Client $i \in [n]$ computes the following weighted sum of its in-neighbors' cumulative local gradients:
        $\Delta_i(t) \leftarrow \sum_{j \in \mathcal{N}_i^-(t)} \frac{1}{d_j^+(t)} \left(x_j^{(t,T)} - x^{(t)}\right)$
    PS samples $m_\ell(t) = \frac{n_\ell(t)}{n} m(t)$ clients uniformly at random from cluster $\ell \in [c]$
    PS computes $x^{(t+1)} \leftarrow x^{(t)} + \frac{1}{m(t)} \sum_{i=1}^{n} \tau_i(t) \Delta_i(t)$ and broadcasts
    $x^{(t+1)}$ to all clients
    PS computes
$m(t+1) \leftarrow \min\{r \in [n] : \psi(r, \alpha_1(t+1), \ldots, \alpha_c(t+1)) \leq \phi_{\max}\}$
end
return $x^{(t_{\max})}$
```

*4.1.2 SGD Iterations.* Additionally, we make the following standard assumption on the stochastic gradients generated through the SGD procedure for each client:

Assumption 3 (**Unbiasedness and Bounded Variance**). *The SGD noise associated with every client is unbiased, i.e., $\mathbb{E}[\widetilde{\nabla} f_i(x) - \nabla f_i(x) \mid x] = 0$, and it has a bounded variance, i.e., there exists a constant $\varrho > 0$ such that $\mathbb{E}\|\widetilde{\nabla} f_i(x) - \nabla f_i(x)\|^2 \leq \varrho^2$ for all models $x \in \mathbb{R}^p$ and all $i \in [n]$.*

In addition, we assume that the SGD noise is independent across clients, i.e., for all $x \in \mathbb{R}^p$, the random vectors $\left\{\widetilde{\nabla} f_i(x) - \nabla f_i(x)\right\}_{i=1}^n$ are mutually conditionally independent given $x$.

*4.1.3 Gradient Diversity.* Furthermore, we assume that the training data are not distributed uniformly at random among the clients, which gives rise to data heterogeneity among the clients. Unlike the standard assumption on data heterogeneity that imposes a uniform upper bound on $\|\nabla f_i(x) - \nabla f(x)\|$ (see [31] for example), we make a weaker assumption on the diversity of local gradients. In fact, this assumption, which was first proposed in [20], can be derived as a consequence of Assumptions 1 and 2, as shown in [20]. Below, we formally state this observation.

Lemma 4.1 (**Gradient diversity** [20]). *For all $i \in [n]$ and $x \in \mathbb{R}^p$, we have $\|\nabla f_i(x) - \nabla f(x)\| \leq \delta + 2\beta\|x - x^*\|$, where*

$$\delta := \beta \max_{i \in [n]} \|x^* - x_i^*\| = \beta \max_{i \in [n]} \|x^* - \arg\min_{y \in \mathbb{R}^p} f_i(y)\| \tag{8}$$

As argued in [20], the standard assumption (which is a special case of the above inequality with $\beta = 0$) is unrealistic as it does not apply to quadratic and super-quadratic loss functions unless the upper bound $\delta$ is chosen to be unreasonably large.

## 4.2 Results

We now quantify how the singular values of the equal-neighbor matrices and the number of clients sampled by the PS affect the efficiency of our algorithm in terms of its optimality gap.

We first show how the expected optimality gap of our algorithm depends on the expected deviation of the global average $x^{(t+1)} - x^{(t)}$ (i.e., the random vector computed by the PS using the aggregation rule (4)) from the true average of all the scaled cumulative gradients.

Lemma 4.2. *At the end of the $(t+1)$-th round of global aggregation, the expected optimality gap of Algorithm 1 is given by*

$$\mathbb{E}\left\|x^{(t+1)} - x^*\right\|^2 = \mathbb{E}\left\|x^{(t+1)} - \bar{x}^{(t+1)}\right\|^2 + \mathbb{E}\left\|\bar{x}^{(t+1)} - x^*\right\|^2,$$

*where $\bar{x}^{(t+1)} := x^{(t)} + \frac{1}{n}\sum_{i=1}^{n}(x_i^{(t,T)} - x^{(t)})$ is a vector that would equal the global model if the PS were to sample all the n clients.*

Observe that the first term on the RHS depends on $x^{(t+1)} - \bar{x}^{(t+1)}$, which can be easily shown to be the difference between the random average $\frac{1}{m(t)}\sum_{i\in\mathcal{S}(t)}\Delta_i(t)$ and the true average $\frac{1}{n}\sum_{i=1}^{n}\left(x_i^{(t,T)} - x^{(t)}\right)$. Thus, this term captures the error due to random sampling. As the next result shows, this difference depends on the network topology as well as on $m(t)$, the number of clients selected for global aggregation uniformly at random by the PS.

Proposition 4.3. *Let $\delta$ be the constant defined in* (8). *Then Algorithm 1 satisfies the following for every $t \in \mathbb{N} \cup \{0\}$:*

$$\mathbb{E}\left\|x^{(t+1)} - \bar{x}^{(t+1)}\right\|^2 \le \Big(2T\varrho^2\eta_t^2 + 4eT(\varrho^2 + 2\delta^2)\eta_t^2 + 6\delta^2T^2\eta_t^2 + (27 + 4e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2\Big)\phi(t),$$

*where $\phi(t)$ is the connectivity factor defined in* (5).

In other words, $\mathbb{E}\left\|x^{(t+1)} - \bar{x}^{(t+1)}\right\|$ depends on the previous optimality gap $\mathbb{E}\left\|x^{(t)} - x^*\right\|^2$ via $\phi(t)$, i.e., the connectivity factor that captures the combined effect of global aggregation (via $m(t)$) and the D2D network topology within each cluster (via $\phi_\ell(\alpha_\ell(t))$).

Moreover, Lemma 4.2 and Proposition 4.3 together show that the singular values of the equal-neighbor adjacency matrices can be used to derive an upper bound on the expected optimality gap (and ultimately establish theoretical performance guarantees) for our connectivity-aware algorithm. Doing so yields the following.

Proposition 4.4. *Let $\delta$ be as defined in* (5), *let $\phi(t)$ be the connectivity factor defined in* (5), *let $\Gamma := f(x^*) - \frac{1}{n}\sum_{i=1}^{n}\min_{x\in\mathbb{R}^p} f_i(x)$, and let e denote the exponential constant. Then the expected optimality gap of Algorithm 1 satisfies the following for all $t \in \mathbb{N}_0$:*

$$\begin{aligned}
&\mathbb{E}\left\|x^{(t+1)} - x^*\right\|^2\\
&\le \Big((1-\mu\eta_t)^T + (27+4e)T^2\beta^2\eta_t^2(2T+\phi(t))\Big)\mathbb{E}\left\|x^{(t)} - x^*\right\|^2\\
&\quad + T\left(\frac{\varrho^2}{n} + 6\beta\Gamma + 4T\varrho^2 + 8eT(\varrho^2 + 2\delta^2) + 12\delta^2T^2\right)\eta_t^2\\
&\quad + \left(2T\varrho^2 + 4eT(\varrho^2 + 2\delta^2) + 6\delta^2T^2\right)\phi(t)\eta_t^2.
\end{aligned}$$

A recursive expansion on the inequality stated by Proposition 4.4 results in our main theoretical result, which we state below.

Theorem 4.5. *Consider a connectivity factor threshold $\phi_{\max} \ge 0$, and suppose that $\phi(t) \le \phi_{\max}$ for all times $t \ge 0$. In addition, suppose $\eta_t = \frac{4}{T\mu(t+t_1)}$, where*

$$t_1 := \left\lfloor 4\left(1 - \frac{1}{T}\right) + (16T + 8\phi_{\max})\left(\frac{\beta}{\mu}\right)^2 + 1\right\rfloor.$$

*Then the expected optimality gap of Algorithm 1 satisfies the following for all $t \ge 0$:*

$$\begin{aligned}
&\mathbb{E}\left\|x^{(t)} - x^*\right\|^2\\
&\le \left(\frac{t_1}{t+t_1}\right)^2\mathbb{E}\left\|x^{(0)} - x^*\right\|^2 + \frac{16\left(\frac{1}{nT}\left(\frac{\varrho}{\mu}\right)^2 + 6\frac{\beta\Gamma}{T\mu^2}\right)}{t+t_1}\\
&+ \frac{(32T + 16\phi_{\max})\left(\frac{2}{T}\left(\frac{\varrho}{\mu}\right)^2 + \frac{4e}{T}\left(\left(\frac{\varrho}{\mu}\right)^2 + 2\left(\frac{\delta}{\mu}\right)^2\right) + 6\left(\frac{\delta}{\mu}\right)^2\right)}{t+t_1}. \quad (9)
\end{aligned}$$

Theorem 4.5 reveals that the convergence rate of our algorithm is $\mathcal{O}(1/t)$, which coincides with that of FedAvg and its semi-decentralized variants such as [20]. In fact, $\mathcal{O}(1/t)$ resembles the convergence rate of vanilla centralized SGD. It also shows that suitably tuning the connectivity factor (by choosing an appropriate value of $\phi_{\max}$) is critical to the efficiency of the algorithm: as $\phi_{\max}$ increases the bound gets worse/larger; however, $\phi_{\max}$, by its definition, is non-negative, which means it can at best be made equal to 0, which forces $m = n$, in which case the inequality boils down to an upper bound on the convergence rate of FedAvg with full device sampling. At the other extreme, setting $\phi_{\max}$ to $\infty$ results in $m = 1$, which happens when our semi-decentralized FL architecture collapses to full decentralization.

Moreover, Theorem 4.5 jointly captures the effect of the following factors on the expected instantaneous optimality gap and hence on the convergence rate: (i) the initial optimality gap $\mathbb{E}\|x^{(0)} - x^*\|^2$ (via the first term), (ii) The SGD noise variance $\varrho^2$ and the strong convexity $\mu$ and smoothness parameters $\beta$ (via the second term), and finally, (iii) the combined effect of cluster connectivity levels and random sampling-based global aggregations (via the third term, which depends on $\phi_{\max}$, which in turn prevents the connectivity factor $\phi(t)$ from becoming too large). It can be seen that higher values of the SGD noise variance $\varrho^2$ and the data heterogeneity measure $\Gamma$ lead to a larger value of the bound, implying that our algorithm is sensitive to the size of the mini-batches used for computing the stochastic gradients as well as to the non-i.i.d.-ness of the local datasets.

## 5 SINGULAR VALUE BOUNDS

Having established the role of the connectivity factor $\phi(t)$ in the performance of Algorithm 1, we now analyze two important quantities associated with $\phi(t)$: the top two singular values of the equal-neighbor adjacency matrices of the clusters. Since a precise estimation of these singular values requires full knowledge of the cluster topologies, which is challenging to obtain in practice, we are motivated to derive a set of novel upper bounds on these values in terms of the node degrees of the cluster digraphs, which are easy to obtain/measure in practice. To the best of our knowledge, this is one of the first attempts at connecting the singular values of adjacency matrices with minimal topological information such as node degrees of the digraphs.

To conduct our analysis, for any digraph $G = ([s], E)$, we first define $\varepsilon = \varepsilon_G := \frac{d^+_{\max}(G) - d^+_{\min}(G)}{d^+_{\min}(G)}$, which quantifies the heterogeneity of out-degree of the nodes across the digraph. We also let $\alpha(G) := \frac{d^+_{\min}(G)}{s}$ capture the minimum fraction of the node population that any node is out-connected to. In addition, we let $W(G) = (w_{ij})$ and $D^+(G) := \mathrm{diag}([d^+_1 \; d^+_2 \; \cdots \; d^+_s]^\top)$ denote the binary adjacency matrix and the out-degree matrix of $G$, respectively. In the sequel, we drop the indexing $(G)$ for brevity.

We are now equipped to state our first set of bounds on the greatest two singular values of $G$ under certain regularity assumptions on the digraph.

Proposition 5.1. *Suppose $G = ([s], E)$ is a directed graph in which every node has its in-degree equal to its out-degree, i.e., $d^+_i = d^-_i$ for all $i \in [n]$. Then the greatest two singular values $\sigma_1$ and $\sigma_2$ of the equal-neighbor adjacency matrix $A$ of $G$ satisfy the following inequalities for $\alpha > \frac{1}{2}$ and $\varepsilon \ll 1$:*

$$\sigma_1^2 \le 1 + \varepsilon + O(\varepsilon^2), \tag{10}$$

$$\sigma_2^2 \le \left(\frac{1}{\alpha} - 1\right)^2 + 2\varepsilon\left(1 + \frac{2}{\alpha} - \frac{1}{\alpha^2}\right) + O(\varepsilon^2), \tag{11}$$

*where $O(\cdot)$ is the big-O notation used in the context of $\varepsilon \longrightarrow 0$.*

Proof. To simplify our notation, we define $D := D^+$ for the remainder of this proof. Observe that

$$A^\top = D^{-1}W = D^{-\frac{1}{2}}(D^{-\frac{1}{2}}WD^{-\frac{1}{2}})D^{\frac{1}{2}},$$

which means $A^\top$ is similar to the normalized adjacency matrix defined as $A_N := D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.

On the other hand, we have $d^+_{\min} = d^+_{\max}(1-\varepsilon) \le d^+_i \le d^+_{\max}$ for all $i \in [s]$, which implies the existence of a diagonal matrix $E_3$ such that $O \le E_3 \le I$ and $D = d^+_{\max}\left((1-\varepsilon)I + \varepsilon E_3\right)$. Using similar arguments, it can be easily shown that there exist diagonal matrices $E_1$ and $E_2$ such that $O \le E_1, E_2 \le I$, $D^{\frac{1}{2}} = \sqrt{d^+_{\max}}\left((1-\frac{\varepsilon}{2})I + \frac{\varepsilon}{2}E_1\right) + O(\varepsilon^2)$, and $D^{-\frac{1}{2}} = \frac{1}{\sqrt{d^+_{\max}}}\left(I + \frac{\varepsilon}{2}E_2\right) + O(\varepsilon^2)$. As a result, the following holds up to an additive error of $O(\varepsilon^2)$:

$$\begin{aligned} A_N = D^{\frac{1}{2}}A^\top D^{-\frac{1}{2}} &= \sqrt{d^+_{\max}}\left((1-\frac{\varepsilon}{2})I + \frac{\varepsilon}{2}E_1\right)\frac{A}{\sqrt{d^+_{\max}}}\left(I + \frac{\varepsilon}{2}E_2\right) \\ &= A^\top + \frac{\varepsilon}{2}\left((E_1 - I)A + AE_2\right), \end{aligned}$$

i.e., $A^\top - A_N = -\frac{\varepsilon}{2}\left((E_1 - I)A^\top + A^\top E_2\right) + O(\varepsilon^2)$. In conjunction with standard bounds on singular value perturbations (e.g., see [23]), this implies the following up to an additive error of $O(\varepsilon^2)$:

$$\begin{aligned} \sigma_j(A) &= \sigma_j(A^\top) \\ &\le \sigma_j(A_N) + \frac{\varepsilon}{2}\left\|(E_1 - I)A^\top + A^\top E_2\right\| \\ &\le \sqrt{\lambda_j(A_N A_N^\top)} + \frac{\varepsilon}{2}\left(\|E_1 - I\|\left\|A^\top\right\| + \left\|A^\top\right\|\|E_2\|\right) \\ &\overset{(a)}{\le} \sqrt{\lambda_j(D^{-\frac{1}{2}}WD^{-1}W^\top D^{-\frac{1}{2}})} + \frac{\varepsilon}{2}(1\cdot\left\|A^\top\right\| + \left\|A^\top\right\|\cdot 1) \\ &\overset{(b)}{=} \sqrt{\lambda_j(D^{-1}WD^{-1}W^\top)} + \varepsilon\sigma_1(A^\top). \end{aligned} \tag{12}$$

Here, $(a)$ holds because $I - E_1$ being a diagonal matrix along with $O \le I - E_1 \le I$ implies that $\|I - E_1\| = \max_{i\in[s]}|(I - E_1)_{ii}| \le 1$. $(b)$ holds because $\sigma_1(A^\top) = \left\|A^\top\right\|$ and because $D^{-1}WD^{-1}W^\top$ and $D^{-\frac{1}{2}}WD^{-1}W^\top D^{-\frac{1}{2}}$, being similar, have the same eigenvalues.

We now bound $\sigma_1(A)$ and $\sigma_2(A)$ individually. As for $\sigma_1(A)$, the derivation (12) and the fact that $\sigma_1(A) = \sigma_1(A^\top)$ imply that

$$\begin{aligned} \sigma_1(A) \le \frac{\sqrt{\lambda_1(D^{-1}WD^{-1}W^T)}}{1-\varepsilon} &= \sqrt{\lambda_1(D^{-1}WD^{-1}W^T)}(1+\varepsilon) \\ &\quad + O(\varepsilon^2). \end{aligned} \tag{13}$$

So, it is enough to bound $\lambda_1(D^{-1}WD^{-1}W^T)$. For this purpose, note that $A$ being column-stochastic implies that $D^{-1}W\mathbf{1} = \mathbf{1}$ and hence also that $W\mathbf{1} = D\mathbf{1}$. Besides, our assumption on in-degrees and out-degrees can be expressed as $\sum_{j=1}^s w_{ij} = \sum_{j=1}^s w_{ji}$ for each $i \in [s]$, or equivalently, $W^\top\mathbf{1} = W\mathbf{1} = D\mathbf{1}$. As a result, we have $D^{-1}W^\top\mathbf{1} = \mathbf{1}$. Thus, $D^{-1}WD^{-1}W^\top = A^\top D^{-1}W^\top$ is a product of row-stochastic matrices and hence, it is row-stochastic in itself. Consequently, $\lambda_1(D^{-1}WD^{-1}W^T) = 1$. In light of this, (13) implies (10).

It remains to prove (11). We do this by using Theorem 2.2 of [21], which helps derive a bound in terms of $\sigma_1$ and the minimum positive entry $\delta$ of the matrix $D^{-1}WD^{-1}W^T$. We first note that

$$\begin{aligned} (D^{-1}WD^{-1}W^T)_{ij} &\overset{(a)}{\ge} \frac{1}{(d^+_{\max})^2}\sum_{k=1}^{s}(W)_{ik}(W^\top)_{kj} \\ &= \frac{1}{(d^+_{\max})^2}\left|\{k \in [s] : w_{ik} = w_{jk} = 1\}\right| \overset{(b)}{\le} \frac{(2\alpha - 1)s}{(d^+_{\max})^2}, \end{aligned}$$

where $(a)$ follows from the fact that $D^{-1} \ge \frac{1}{d^+_{\max}}I$ and $(b)$ holds because the number of common out-neighbors of any two nodes $i, j \in [s]$ is at least $(2\alpha - 1)s$. We can now apply Theorem 2.2 of [21] by setting $x = \frac{1}{\sqrt{s}}$ in the theorem (because $\frac{1}{\sqrt{s}}\mathbf{1}$, as explained above, is the unit-norm principal eigenvector of $D^{-1}WD^{-1}W^\top$). Thus,

$$\begin{aligned} \lambda_2(D^{-1}WD^{-1}W^\top) &\le \lambda_1(D^{-1}WD^{-1}W^\top) - \frac{(2\alpha - 1)s^2}{(d^+_{\max})^2} \\ &= 1 - \left(\frac{2}{\alpha} - \frac{1}{\alpha^2}\right)(1 - 2\varepsilon) + O(\varepsilon^2), \end{aligned} \tag{14}$$

where the last step follows from the observation that $d^+_{\max} = \frac{\alpha s}{1-\varepsilon}$.

Combining (10), (12) and (14) now gives

$$\sigma_2(A) \le \sqrt{1 - \left(\frac{2}{\alpha} - \frac{1}{\alpha^2}\right)(1 - 2\varepsilon) + O(\varepsilon^2)} + \varepsilon\left(1 + \varepsilon + O(\varepsilon)^2\right).$$

Squaring both sides and rearranging the terms results in (11). □

Remark 1. *Observing the bounds in Proposition 5.1, we can see that setting $\alpha = 1$, which corresponds to $G$ being a clique, in the bounds yields $\sigma_1 \leq 1 + O(\varepsilon)$ and $\sigma_2 = O(\varepsilon)$. These inequalities, for $\varepsilon \ll 1$, are tight with respect to the well-known lower bounds $\sigma_1 \geq 1$ and $\sigma_2 \geq 0$. This implies that the bounds* (10) *and* (11) *can be expected to be reasonably tight for high edge density (i.e., whenever $\alpha \approx 1$). Another implication of the bounds is decreasing $\varepsilon$, which inherently measures how irregular the digraph is, leads to* (11) *becoming sharper.*

The above singular value bounds are especially tight for digraphs that are approximately regular (or digraphs that do not exhibit significant variations in their in-degrees and out-degrees) since such digraphs satisfy $\varepsilon \ll 1$. This happens in practice, when the D2D clusters are dense (e.g., in the wireless setting, when the nodes are closer to each other or when they can move and communicate over time). Furthermore, the same holds for the condition on $\alpha$ in Proposition 5.1 (i.e., $\alpha > \frac{1}{2}$), which is always met when the clusters are dense.

However, the bounds (10) and (11) are obtained under the assumption that every node has its in-degree equal to its out-degree, which can be restrictive in practical settings. This observation further motivates us to find a new set of singular value bounds that work well under milder assumptions. We thus provide the following bounds, which not only relax the said restrictive assumption, but also apply to digraphs with more general out-degree distributions (and hence subsume digraphs with wider out-degree variations).

Proposition 5.2. *Let $\varphi = \frac{d^{in}_{\max} - d_{\min}}{d_{\min}}$, where $d^{in}_{\max}$ denotes the maximum in-degree of the digraph $G$. If $\alpha \geq \frac{1}{2}$, we have the following bounds:*

$$\sigma_1^2 \leq 1 + \varphi, \tag{15}$$

$$\sigma_2^2 \leq 1 + \varphi - \frac{(1-\varepsilon)^2(1-\alpha_{-1}^2)\left((1-\varepsilon)^2(1-\alpha_{-1}^2) - \alpha_{-1}\right)}{s(\varepsilon_{net}+1)\left(\varepsilon_{net} - \alpha_{-1} + \frac{1}{\alpha s}\right)}, \tag{16}$$

*where $\varepsilon_{net} := \varphi + \frac{\varepsilon}{\alpha}$ and $\alpha_{-1} := \frac{1}{\alpha} - 1$.*

The bounds obtained in Proposition 5.2 (**proved in the appendix**) are particularly effective when the D2D cluster digraphs are dense but irregular. This is often the case in practical systems, when there is communication heterogeneity (e.g., in wireless sensor networks consisting of sensors with different radii).

In conjunction with Theorem 4.5, the bounds derived in Propositions 5.1 and 5.2 capture the inherent dependence of the expected optimality gap, and hence that of the convergence rate, on the degree distributions of the D2D clusters. In particular, upon having approximately regular D2D clusters, the bounds in Propositions 5.1 along with the result of Theorem 4.5 determine the convergence rate of Algorithm 1. The same holds when using the result of Proposition 5.2 with Theorem 4.5, which will characterize the convergence rate of Algorithm 1 upon having irregular D2D clusters.

# 6 NUMERICAL VALIDATION

We now conduct numerical experiments to validate our methodology. Overall, our simulations show that compared with baselines, Algorithm 1 obtains significant reductions in total communication cost for the same or similar levels of testing accuracy.

## 6.1 Implementation

*6.1.1 Network Architecture.* We simulate a network consisting of $n = 70$ edge devices partitioned into $c = 7$ clusters with $n_\ell = 10$ nodes per cluster. In every global aggregation round, the digraph for each cluster $\ell \in [c]$ is constructed as follows: (i) we generate a $k$-regular directed graph (a digraph in which every node has its in-degree and out-degree equal to $k$) with the value of $k$ being chosen uniformly at random from the set $\{6, \ldots, 9\}$; (ii) we delete a fraction $p \in (0,1)$ of the directed edges uniformly at random so as to incorporate D2D link failures due to client mobility and bandwidth issues. The result is an approximately regular digraph whose degree distribution may deviate significantly from that of regular digraphs, while satisfying $\alpha_\ell(t) > \frac{1}{2}$.

*6.1.2 Datasets.* All our simulations are performed on MNIST [37] and Fashion-MNIST (F-MNIST) [34] datasets. The MNIST dataset consists of 70K images (60K for training and 10K for testing), and each image is a hand-written digit between 0 to 9 (i.e., the dataset has 10 labels). The same applies to the FMNIST dataset, the only difference being that it consists of images of fashion products.

*6.1.3 ML Models and Implementation.* We use the neural network model from [22] in our simulations. In particular, we use a convolutional neural network (CNN) with two $5 \times 5$ convolution layers, the first of which has 32 channels and the second 64 channels, where each of these layers precedes a $2 \times 2$ max pooling, resulting in a total model dimension of $1,663,370$. We use the PyTorch implementation of this setup provided in [11] with cross-entropy loss. Each dataset is distributed among the clients in a non-i.i.d. manner: the samples (from either of the two datasets) are first sorted by their labels, partitioned into chunks of equal size, and each of the 70 clients is assigned only two chunks (i.e., each client will end up having only two labels). This results in extreme data heterogeneity, which leads to strong empirical guarantees for our approach.

All of our simulations are performed using the following hyperparameter values/ranges: $T = 5$, $t_{\max} \in \{15, 30\}$, $p \in \{0.1, 0.2\}$, and $\eta_t = 0.02(0.1)^t$ where $t$ is the global aggregation index.

## 6.2 Results

We compare the energy vs. accuracy trade-offs associated with Algorithm 1 with those associated with two baselines, FedAvg [22] and collaborative relaying (COLREL) [38]. The second baseline is a recently proposed semi-decentralized FL algorithm that incorporates single-step consensus updates. Under the D2D and D2S connectivity constraints introduced in Section 2, COLREL is a variant of FedAvg that incorporates one round of column-stochastic D2D aggregations before every global aggregation round but does not provide any criterion to control the sampling size $m$, which we assume to be fixed throughout its implementation. The fundamental difference between our method and COLREL is that our method takes into account the change in the connectivity of D2D clusters, optimally tuning the value of $m$ according to the set of novel upper bounds on the singular values we obtained in Section 5.

We consider these tradeoffs under different D2S connectivity levels. Intuitively speaking, on one hand, as the D2S connectivity improves, we expect to see that our algorithm leads to a lower energy and cost savings as compared to FedAvg. This is because

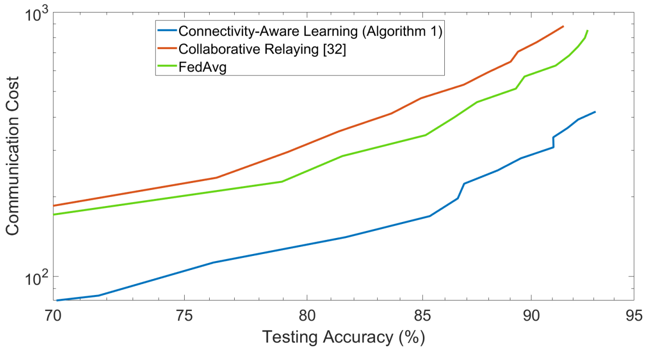

**Figure 2: Communication cost vs. testing accuracy under high D2S connectivity (Dataset: MNIST).**

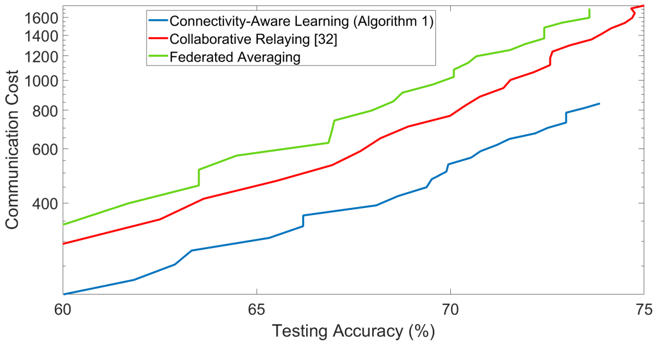

**Figure 3: Communication cost vs. testing accuracy under high D2S connectivity (Dataset: F-MNIST).**

our algorithm will naturally collapse to FedAvg and D2D communications will become less useful since more devices would engage in uplink communications, which by itself degrades the benefit of D2D local aggregations. On the other hand, as D2S connectivity improves, we expect to see that our algorithm achieves significant energy savings as compared to COLREL. This is because the impact of tuning $m$ becomes more prominent when there is a possibility of D2S communications.

All of the following plots and discussion are based on the assumption that the ratio of the energy required for D2D communication to that of up-link (D2S) transmission, denoted by $\frac{E_{\text{D2D}}}{E_{\text{Glob}}}$, equals 0.1. This is a pessimistic estimate in favor of D2S considering that most ratios reported in the literature [8, 20, 40] take values less than 0.1. Thus, the communication costs reported are (#D2S transmissions) + 0.1 × (#D2D transmissions).

*6.2.1 Case 1: Cost savings under high D2S connectivity and a low link failure probability.* When the PS has a high downlink bandwidth and the connectivity between the devices and the PS is reliable, implementing FedAvg or COLREL has the effect of setting $m$ to a value close to $n$. As an example, we implement FedAvg and COLREL with $m = 57$ and $m = 52$, respectively (note that COLREL requires fewer up-link transmissions because it uses D2D consensus updates in addition to global aggregations). The results for MNIST are shown in Fig. 2: choosing $\phi_{\max} = 0.06$ and a low D2D link failure probability $p = 0.1$ results in Algorithm 1 achieving a testing accuracy of 90% while consuming about 46% less energy than FedAvg (thereby incurring proportionately lower communication costs). With respect to COLREL, the energy saving is even higher because COLREL also expends energy on D2D aggregations with relatively little gain in testing accuracy.

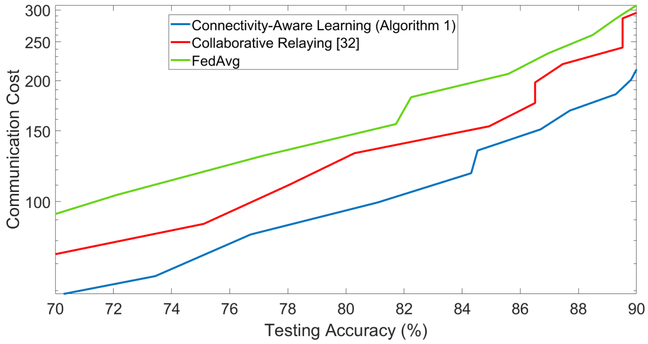

**Figure 4: Communication cost vs. testing accuracy under low D2S connectivity (Dataset: MNIST).**

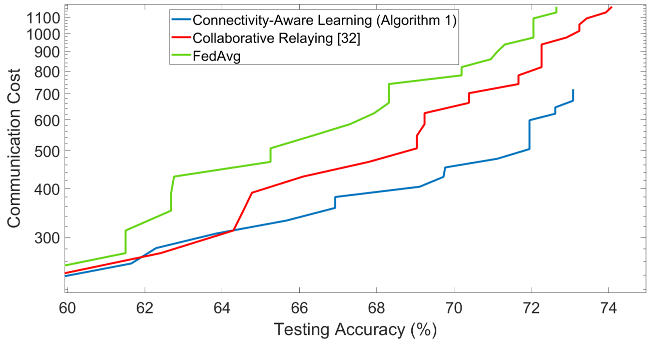

**Figure 5: Communication cost vs. testing accuracy under low D2S connectivity (Dataset: F-MNIST).**

Repeating this experiment on FMNIST results in a similar performance, depicted in Fig. 3. We see that Algorithm 1 (with $\phi_{\max} = 0.06$) consumes about 30% less energy than COLREL for achieving a testing accuracy of 70%.

*6.2.2 Case 2: Cost savings under low D2S connectivity and a high link failure probability.* When the connectivity between the devices and the PS is poor, implementing FedAvg or COLREL has the effect of setting $m$ to a value significantly smaller than $n$. As an example, we implement FedAvg and COLREL with $m = 26$ and $m = 15$, respectively. Choosing $\phi_{\max} = 0.2$ and a high D2D link failure probability $p = 0.2$ results in Algorithm 1 consuming about 30% less energy than FedAvg for achieving a testing accuracy of 90% on MNIST, as shown in Fig. 4. The cost saving is lower than in Case 1 as we expect because the singular value bounds incorporated by our algorithm into its choice of $m(t)$ are looser for higher values of the link failure probability $p$. Repeating this experiment on FMNIST results in a similar performance, depicted in Fig. 5.

## 7 CONCLUSION

We have investigated consensus-based semi-decentralized learning over clustered D2D networks modeled as time-varying digraphs. We first revealed the connection between the singular values of the column-stochastic matrices used for D2D model aggregations and the convergence rate of the learning process. We then derived a set of novel upper bounds on these singular values in terms of the degree distributions of the cluster digraphs, and we used the resulting bounds to design a novel connectivity-aware FL algorithm that enables the central parameter server to tune the number of up-link transmissions by using its knowledge of the time-varying degree distributions of clusters. Our algorithm maintains a continuous balance between the number of model aggregations occurring at the server and those occurring over the edge network, thereby enhancing the resource efficiency of the learning process without compromising convergence.

Future works include obtaining upper bounds on singular values under more general assumptions, and obtaining optimal device sampling schemes for irregular clusters.

## REFERENCES

[1] Mohammad Akbari, Bahman Gharesifard, and Tamás Linder. 2015. Distributed online convex optimization on time-varying directed graphs. *IEEE Transactions on Control of Network Systems* 4, 3 (2015), 417–428.

[2] Mohammed S Al-Abiad, Mohanad Obeed, Md Hossain, Anas Chaaban, et al. 2022. Decentralized aggregation for energy-efficient federated learning via overlapped clustering and D2D communications. *arXiv preprint arXiv:2206.02981* (2022).

[3] Aurélien Bellet, Anne-Marie Kermarrec, and Erick Lavoie. 2022. D-cliques: Compensating for data heterogeneity with topology in decentralized federated learning. In *2022 41st International Symposium on Reliable Distributed Systems (SRDS)*. IEEE, 1–11.

[4] Enrique Tomás Martínez Beltrán, Mario Quiles Pérez, Pedro Miguel Sánchez Sánchez, Sergio López Bernal, Gérôme Bovet, Manuel Gil Pérez, Gregorio Martínez Pérez, and Alberto Huertas Celdrán. 2022. Decentralized Federated Learning: Fundamentals, State-of-the-art, Frameworks, Trends, and Challenges. *arXiv preprint arXiv:2211.08413* (2022).

[5] Aleksandr Beznosikov, Pavel Dvurechensky, Anastasia Koloskova, Valentin Samokhin, Sebastian U Stich, and Alexander Gasnikov. 2021. Decentralized local stochastic extra-gradient for variational inequalities. *arXiv preprint arXiv:2106.08315* (2021).

[6] Christopher Briggs, Zhong Fan, and Peter Andras. 2020. Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–9.

[7] Chaoyang He, Conghui Tan, Hanlin Tang, Shuang Qiu, and Ji Liu. 2019. Central server free federated learning over single-sided trust social networks. *arXiv preprint arXiv:1910.04956* (2019).

[8] Mariem Hmila, Manuel Fernández-Veiga, Miguel Rodriguez-Perez, and Sergio Herrería-Alonso. 2019. Energy efficient power and channel allocation in underlay device to multi device communications. *IEEE transactions on communications* 67, 8 (2019), 5817–5832.

[9] Seyyedali Hosseinalipour, Sheikh Shams Azam, Christopher G Brinton, Nicolo Michelusi, Vaneet Aggarwal, David J Love, and Huaiyu Dai. 2022. Multi-stage hybrid federated learning over large-scale D2D-enabled fog networks. *IEEE/ACM Transactions on Networking* 30, 4 (2022), 1569–1584.

[10] Yifan Hua, Kevin Miller, Andrea L Bertozzi, Chen Qian, and Bao Wang. 2022. Efficient and reliable overlay networks for decentralized federated learning. *SIAM J. Appl. Math.* 82, 4 (2022), 1558–1586.

[11] Shaoxiong Ji. 2018. A PyTorch Implementation of Federated Learning.

[12] Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. 2021. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems* 34 (2021), 11422–11435.

[13] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. 2020. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*. PMLR, 5381–5393.

[14] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[15] Anusha Lalitha, Shubhanshu Shekhar, Tara Javidi, and Farinaz Koushanfar. 2018. Fully decentralized federated learning. In *Third workshop on bayesian deep learning (NeurIPS)*, Vol. 2.

[16] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine* 37, 3 (2020), 50–60.

[17] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189* (2019).

[18] Shu Liang, George Yin, et al. 2019. Dual averaging push for distributed convex optimization over time-varying directed graph. *IEEE Trans. Automat. Control* 65, 4 (2019), 1785–1791.

[19] Frank Po-Chen Lin, Seyyedali Hosseinalipour, Sheikh Shams Azam, Christopher G Brinton, and Nicolò Michelusi. 2021. Federated learning beyond the star: Local D2D model consensus with global cluster sampling. In *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.

[20] Frank Po-Chen Lin, Seyyedali Hosseinalipour, Sheikh Shams Azam, Christopher G Brinton, and Nicolo Michelusi. 2021. Semi-decentralized federated learning with cooperative D2D local model aggregations. *IEEE Journal on Selected Areas in Communications* 39, 12 (2021), 3851–3869.

[21] M Stuart Lynn and William P Timlake. 1969. Bounds for Perron eigenvectors and subdominant eigenvalues of positive matrices. *Linear Algebra Appl.* 2, 2 (1969), 143–152.

[22] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[23] Carl D Meyer. 2000. *Matrix analysis and applied linear algebra*. Vol. 71. Siam.

[24] Angelia Nedić and Alex Olshevsky. 2014. Distributed optimization over time-varying directed graphs. *IEEE Trans. Automat. Control* 60, 3 (2014), 601–615.

[25] Angelia Nedić and Alex Olshevsky. 2016. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Trans. Automat. Control* 61, 12 (2016), 3936–3947.

[26] Angelia Nedic, Alex Olshevsky, and Wei Shi. 2017. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization* 27, 4 (2017), 2597–2633.

[27] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. 2020. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021–2031.

[28] Dan Shen, Genshe Chen, Jose B Cruz, and Erik Blasch. 2008. A game theoretic data fusion aided path planning approach for cooperative UAV ISR. In *2008 IEEE Aerospace Conference*. IEEE, 1–9.

[29] Tao Sun, Dongsheng Li, and Bao Wang. 2022. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[30] Bin Wang, Jun Fang, Hongbin Li, Xiaojun Yuan, and Qing Ling. 2022. Confederated Learning: Federated Learning with Decentralized Edge Servers. *arXiv preprint arXiv:2205.14905* (2022).

[31] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. 2019. Adaptive federated learning in resource constrained edge computing systems. *IEEE journal on selected areas in communications* 37, 6 (2019), 1205–1221.

[32] Zheng Wang and Huaqing Li. 2019. Edge-based stochastic gradient algorithm for distributed optimization. *IEEE Transactions on Network Science and Engineering* 7, 3 (2019), 1421–1430.

[33] Zutong Wang, Mingfa Zheng, Jiansheng Guo, and Hanqiao Huang. 2017. Uncertain UAV ISR mission planning problem with multiple correlated objectives. *Journal of Intelligent & Fuzzy Systems* 32, 1 (2017), 321–335.

[34] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[35] Ran Xin and Usman A Khan. 2018. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters* 2, 3 (2018), 315–320.

[36] Hong Xing, Osvaldo Simeone, and Suzhi Bi. 2021. Federated learning over wireless device-to-device networks: Algorithms and convergence analysis. *IEEE Journal on Selected Areas in Communications* 39, 12 (2021), 3723–3741.

[37] L Yan, C Corinna, and CJ Burges. 1998. The MNIST dataset of handwritten digits.

[38] Michal Yemini, Rajarshi Saha, Emre Ozfatura, Deniz Gündüz, and Andrea J Goldsmith. 2022. Semi-decentralized federated learning with collaborative relaying. In *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 1471–1476.

[39] Shahryar Zehtabi, Seyyedali Hosseinalipour, and Christopher G Brinton. 2022. Event-Triggered Decentralized Federated Learning over Resource-Constrained Edge Devices. *arXiv preprint arXiv:2211.12640* (2022).

[40] Aiqing Zhang and Xiaodong Lin. 2017. Security-aware and privacy-preserving D2D communications in 5G. *IEEE Network* 31, 4 (2017), 70–77.

## Appendix: Proofs and Auxiliary Results

### Proof of Lemma 4.2

*Proof.* The key steps are to note that

$$\begin{aligned}\mathbb{E}\left\|x^{(t+1)}-x^{*}\right\|^{2}=&\ \mathbb{E}\left\|x^{(t+1)}-\bar{x}^{(t+1)}\right\|^{2}+\mathbb{E}\left\|\bar{x}^{(t+1)}-x^{*}\right\|^{2}\\&+2\mathbb{E}\left[\left(x^{(t+1)}-\bar{x}^{(t+1)}\right)^{T}\left(\bar{x}^{(t+1)}-x^{*}\right)\right]\end{aligned}\tag{17}$$

and to show that the cross-term above vanishes. To this end, let $v(t)\in\mathbb{R}^{n}$ denote the vector such that $v_i(t)=\frac{1}{m}$ if $i\in[n]$ is sampled by the PS at time $t$ and $v_i(t)=0$ otherwise. We then observe that $x^{(t+1)}-\bar{x}^{(t+1)}=X_{\text{diff}}\left(v^{T}(t)A(t)-\frac{1}{n}\mathbf{1}^{T}\right)$. As a result, we have the following:

$$\begin{aligned}&\mathbb{E}\left[\left(x^{(t+1)}-\bar{x}^{(t+1)}\right)^{T}\left(\bar{x}^{(t+1)}-x^{*}\right)\right]\\&=\mathbb{E}\left[\left(A^{T}(t)v(t)-\frac{1}{n}\mathbf{1}\right)^{T}X_{\text{diff}}^{T}(t)\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i^{(t,T)}-x^{(t)}\right)\right)\right]\\&=\mathbb{E}\left[\left(A^{T}(t)v(t)-\frac{1}{n}\mathbf{1}\right)^{T}X_{\text{diff}}^{T}(t)X_{\text{diff}}(t)\frac{\mathbf{1}}{n}\right]\\&\overset{(a)}{=}\frac{1}{n}\left(\mathbb{E}\left[A^{T}(t)v(t)-\frac{1}{n}\mathbf{1}\right]\right)^{T}\mathbb{E}\left[X_{\text{diff}}^{T}(t)X_{\text{diff}}(t)\right]\mathbf{1}\\&\overset{(b)}{=}\frac{1}{n}\left(\frac{1}{n}A^{T}(t)\mathbf{1}-\frac{1}{n}\mathbf{1}\right)^{T}\mathbb{E}\left[X_{\text{diff}}^{T}(t)X_{\text{diff}}(t)\right]\mathbf{1}\\&\overset{(c)}{=}0,\end{aligned}$$

where $(a)$ holds because $X_{\text{diff}}(t)$, which is determined by $\{x^{(i,T)}-x^{(t)}\}_{i=1}^{n}$, is independent of the random subset of nodes sampled by the PS and hence also independent of $v(t)$, $(b)$ follows from the observation that $\mathbb{E}[v(t)]=\frac{1}{n}\mathbf{1}$, which holds because every node is sampled with the same probability, and $(c)$ holds because $A(t)$ is column-stochastic. □

### Auxiliary Lemmas

The proof of Proposition 4.3 is founded on the following auxiliary lemmas.

**Lemma 3.** *For any set of $k\in\mathbb{N}$ random vectors $\{Y^{(i)}\}_{i=1}^{k}$ that take values in $\mathbb{R}^{q}$ (where $q\in\mathbb{N}$), we have $\mathbb{E}\left\|\sum_{i=1}^{k}Y^{(i)}\right\|^{2}\le k\sum_{i=1}^{k}\mathbb{E}\left\|Y^{(i)}\right\|^{2}$.*

*Proof.* Let $Y_j^{(i)}$ denote the $j$-th entry of $Y^{(i)}$ for each $j\in[q]$ and $i\in[k]$. Additionally, let $\widetilde{Y}^{(j)}:=$

$[Y_j^{(1)}\ Y_j^{(2)}\ \ \ldots,\ Y_j^{(k)}]^T$ for each $j \in [q]$. We then have

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{k} Y^{(i)}\right\|^2 = \sum_{j=1}^{m}\mathbb{E}\left(\sum_{i=1}^{k} Y_j^{(i)}\right)^2 &= \sum_{j=1}^{m}\mathbb{E}\left(\mathbf{1}^T\widetilde{Y}^{(j)}\right)^2 \\
&\leq \sum_{j=1}^{m}\mathbb{E}\left[\|\mathbf{1}\|^2\left\|\widetilde{Y}^{(j)}\right\|^2\right] \\
&= k\sum_{j=1}^{m}\mathbb{E}\left\|\widetilde{Y}^{(j)}\right\|^2 \\
&= k\mathbb{E}\sum_{j=1}^{m}\sum_{i=1}^{k}\left(Y_j^{(i)}\right)^2 = k\mathbb{E}\sum_{i=1}^{k}\mathbb{E}\left\|Y^{(i)}\right\|^2,
\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality. This proves the lemma. □

**Lemma 4.** *Every $n \times n$ matrix $\mathbf{W}$ satisfies $\|\mathbf{W}\|^2 \leq \sum_{i=1}^{n}\|W_i\|^2$, where $\{W_i\}_{i=1}^{n}$ are the columns of $\mathbf{W}$.*

*Proof.* We have

$$\sum_{i=1}^{n}\|W_i\|^2 \overset{(a)}{=} \operatorname{trace}(\mathbf{W}^T\mathbf{W}) \overset{(b)}{=} \sum_{i=1}^{n}\sigma_i^2(\mathbf{W}) \geq \sigma_1^2(\mathbf{W}) \overset{(c)}{=} \|\mathbf{W}\|^2, \tag{18}$$

where $(a)$ follows from the property [23, (5.2.1)] of Frobenius norms, $(b)$ holds because trace$(\mathbf{W}^T\mathbf{W})$ equals the sum of the eigenvalues of $\mathbf{W}^T\mathbf{W}$, which are the squares of the singular values of $\mathbf{W}$, and $(c)$ holds because $\|\mathbf{W}\|^2 = \min_{\|z\|=1}\|\mathbf{W}z\|^2 = \min_{\|z\|=1} z^T\mathbf{W}^T\mathbf{W}z = \sigma_1^2(\mathbf{W})$, where the last equality follows from the Courant-Fischer theorem (see [23]). □

**Lemma 5.** *For $s \in \mathbb{N}$, let $A \in \mathbb{R}^{s\times s}$ be an irreducible non-negative matrix with positive diagonal entries. Then the matrix $AA^T$ is irreducible.*

*Proof.* Note that for any two indices $i, j \in [s]$, we have $(AA^T)_{ij} = \sum_{k=1}^{n} a_{ik}a_{jk} \geq a_{ij}a_{jj}$, which in conjunction with the positive diagonal assumption implies that $(AA^T)_{ij} > 0$ whenever $a_{ij} > 0$. Thus, $\mathcal{G}(AA^T)$ is a super-digraph of $\mathcal{G}(A)$, which, in turn, is a strongly connected digraph because $A$ is irreducible [23]. Hence, $\mathcal{G}(AA^T)$ is strongly connected. Equivalently, $AA^T$ is irreducible. □

**Lemma 6.** *For $s \in \mathbb{N}$, let $v \in \mathbb{R}^s$ be a stochastic vector, and let $A \in \mathbb{R}^{s\times s}$ be an irreducible column-stochastic matrix with positive diagonal entries. Then*

$$\left\|A^Tv - \frac{1}{n}\mathbf{1}\right\|^2 \leq (\sigma_1^2 + \sigma_2^2 - 1)\|v_\perp\|^2,$$

*where $v_\perp := v - \frac{1}{s}\mathbf{1}$ is the component of $v$ that is orthogonal to $\mathbf{1}$, and $\sigma_1$ and $\sigma_2$ are the largest and the second-largest singular values of $A$, respectively.*

*Proof.* We first show that the quantity in question equals $v_\perp^T AA^T v_\perp$, we derive an inequality connecting the principal eigenvector of $AA^T$ with $\sigma_1$ and $\sigma_2$, and we then use the results of each of these steps to obtain the desired upper bound.

For the first step, we expand the square of the norm in question as follows:

$$\begin{aligned}
&\left\|v^TA - \frac{1}{s}\mathbf{1}^T\right\|^2 \\
&= v^TAA^Tv - \frac{2}{s}\mathbf{1}^TA^Tv + \frac{1}{s^2}\mathbf{1}^T\mathbf{1} \\
&= \left(v_\perp^T + \frac{1}{s}\mathbf{1}^T\right)AA^T\left(v_\perp + \frac{1}{s}\mathbf{1}\right) - \frac{2}{s}\mathbf{1}^TA^T\left(v_\perp + \frac{1}{s}\mathbf{1}\right) + \frac{1}{s^2}\mathbf{1}^T\mathbf{1} \\
&= v_\perp^TAA^Tv_\perp + \frac{1}{s^2}\left(A^T\mathbf{1}\right)^TA^T\mathbf{1} + \frac{2}{s}\left(A^T\mathbf{1}\right)^TA^Tv_\perp - \frac{2}{s}\mathbf{1}^TA^Tv_\perp - \frac{2}{s^2}\mathbf{1}^TA^T\mathbf{1} + \frac{1}{s^2}\mathbf{1}^T\mathbf{1} \\
&\overset{(a)}{=} v_\perp^TAA^Tv_\perp, \qquad (19)
\end{aligned}$$

where $(a)$ is obtained on simplifying the concerned expressions using the relation $A^T\mathbf{1} = \mathbf{1}$.

Next, we let $\hat{p}$ denote the unit-norm principal eigenvector of $AA^T$, and we relate $\sigma_1$ and $\sigma_2$ to $\hat{p}$. To do so, we first note that $AA^T$ is irreducible by Lemma 5, and hence, the Perron-Frobenius theorem implies that $\hat{p}$ is unique up to scaling by a complex scalar of unit magnitude. We now let $\{\hat{v}_j\}_{j=2}^s$ denote the unit-norm eigenvectors of $AA^T$ corresponding to its eigenvalues $\{\sigma_j^2\}_{j=2}^s$, where $\sigma_j$ denotes the $j$-th largest singular value of $A$. Observe that

$$\begin{aligned}
s = \mathbf{1}^T\mathbf{1} = \mathbf{1}^TAA^T\mathbf{1} &\overset{(a)}{=} \mathbf{1}^T\left(\sigma_1^2\hat{p}\hat{p}^T + \sum_{j=2}^s \sigma_j^2\hat{v}_j\hat{v}_j^T\right)\mathbf{1} \\
&= \sigma_1^2(\mathbf{1}^T\hat{p})^2 + \sum_{j=2}^s \sigma_j^2(\mathbf{1}^T\hat{v}_j) \\
&\le \sigma_1^2(\mathbf{1}^T\hat{p})^2 + \sigma_2^2\sum_{j=2}^s(\mathbf{1}^T\hat{v}_j)^2 \\
&\overset{(b)}{=} \sigma_1^2(\mathbf{1}^T\hat{p})^2 + \sigma_2^2(s - (\mathbf{1}^T\hat{p})^2),
\end{aligned}$$

where $(a)$ follows from the spectral decomposition theorem for symmetric matrices, and $(b)$ holds because of the following reasons: the symmetry of $AA^T$ implies that $\mathcal{E} := \cup_{j=2}^s\{\hat{v}_j\} \cup \{\hat{p}\}$ is an orthonormal eigenvector basis for $\mathbb{R}^s$, which further implies that $\mathbf{1}$ has a representation $[\mathbf{1}^T\hat{p} \;\; \mathbf{1}^T\hat{v}_2 \;\; \cdots \;\; \mathbf{1}^T\hat{v}_s]^T$ in the basis $\mathcal{E}$. Therefore, we have $(\mathbf{1}^T\hat{p})^2 + \sum_{j=2}^s(\mathbf{1}^T\hat{v}_j)^2 = \|\mathbf{1}\|^2 = s$, which proves $(b)$. We have thus shown that

$$\frac{1}{s}(\mathbf{1}^T\hat{p})^2 \ge \frac{1-\sigma_2^2}{\sigma_1^2 - \sigma_2^2}. \qquad (20)$$

The final step is to upper bound $v_\perp^TAA^Tv_\perp$. To this end, let $\hat{p}$ denote the unit-norm principal eigenvector of $AA^T$, let $\hat{p}_\perp := \hat{p} - \frac{1}{s}(\hat{p}^T\mathbf{1})\mathbf{1}$ denote the component of $\hat{p}$ that is orthogonal to $\mathbf{1}$, and let $v_{\perp\perp} := v_\perp - (v_\perp^T\hat{p})\hat{p}$

denote the component of $v_\perp$ that is orthogonal to $\hat{p}$. We then have

$$\begin{aligned}
v_\perp^T A A^T v_\perp &= \left(v_{\perp\!\!\!\perp} + (v_\perp^T \hat{p})\hat{p}\right) A A^T \left(v_{\perp\!\!\!\perp} + (v_\perp^T \hat{p})\hat{p}\right) \\
&\overset{(a)}{=} v_{\perp\!\!\!\perp}^T A A^T v_{\perp\!\!\!\perp} + (v_\perp^T \hat{p})^2 \hat{p}^T A A^T \hat{p} \\
&\overset{(b)}{\leq} \sigma_2^2 \left\| v_{\perp\!\!\!\perp} \right\|^2 + (v_\perp^T \hat{p})^2 \sigma_1^2 \\
&\overset{(c)}{=} \sigma_2^2 (\|v_\perp\|^2 - (v_\perp^T \hat{p})^2) + (v_\perp^T \hat{p})^2 \sigma_1^2 \\
&\overset{(d)}{=} (\sigma_1^2 - \sigma_2^2)(v_\perp^T \hat{p}_\perp)^2 + \sigma_2^2 \left\| v_\perp \right\|^2 \\
&\overset{(e)}{\leq} \left( (\sigma_1^2 - \sigma_2^2) \left\| \hat{p}_\perp \right\|^2 + \sigma_2^2 \right) \left\| v_\perp \right\|^2 \\
&\overset{(f)}{=} \left( (\sigma_1^2 - \sigma_2^2) \left( 1 - \frac{1}{s} (\mathbf{1}^T \hat{p})^2 \right) + \sigma_2^2 \right) \left\| v_\perp \right\|^2 \\
&\overset{(g)}{\leq} (\sigma_1^2 + \sigma_2^2 - 1) \left\| v_\perp \right\|^2 ,
\end{aligned} \tag{21}$$

where $(a)$ holds because $v_{\perp\!\!\!\perp}^T \hat{p} = 0$, $(b)$ follows from the Courant-Fischer theorem and the fact that $v_{\perp\!\!\!\perp}^T$ is orthogonal to the principal eigenspace $\{\beta \hat{p} : \beta \in \mathbb{R}\}$ of $AA^T$, $(c)$ follows from the Pythagoras theorem, $(d)$ holds because $v_\perp^T \mathbf{1}$ being 0 implies that $v_\perp^T \hat{p} = v_\perp^T \left( \hat{p}_\perp + \frac{1}{s} (\hat{p}^T \mathbf{1}) \mathbf{1} \right) = v_\perp^T \hat{p}_\perp$, $(e)$ follows from the Cauchy-Schwarz inequality, $(f)$ follows from the Pythagoras theorem, the orthogonality of $\hat{p}_\perp$ with $\frac{1}{\sqrt{s}} \mathbf{1}$ and the fact that $\|\hat{p}\| = 1$, and $(g)$ follows from (20).

Combining (21) with (19) now yields the required upper bound. □

**Lemma 7.** *Let $\{Z_i\}_{i=1}^4$ be random vectors of the same dimension such that $\mathbb{E}[Z_1^T Z_2] = \mathbb{E}[Z_1^T Z_4] = 0$. Then $\mathbb{E}\|Z_1 + Z_2 + Z_3 + Z_4\|^2 \leq 2\mathbb{E}\left\|Z_1\right\|^2 + 4\mathbb{E}\left\|Z_2\right\|^2 + 3\left(\mathbb{E}\left\|Z_3\right\|^2 + \mathbb{E}\left\|Z\right\|^4\right)$.*

*Proof.* Observe that

$$\begin{aligned}
&\mathbb{E}\|Z_1 + Z_2 + Z_3 + Z_4\|^2 \\
&= \sum_{i=1}^4 \mathbb{E}\left\|Z_i\right\|^2 + 2\mathbb{E}[Z_1^T Z_2] + 2\mathbb{E}[Z_2^T Z_3 + Z_3^T Z_4 + Z_2^T Z_4] \\
&\overset{(a)}{\leq} \sum_{i=1}^4 \mathbb{E}\left\|Z_i\right\|^2 + \mathbb{E}\left\|Z_1\right\|^2 + \mathbb{E}\left\|Z_2\right\|^2 + \mathbb{E}\left\|Z_2 + Z_3 + Z_4\right\|^2 - \sum_{j=2}^4 \mathbb{E}\left\|Z_j\right\|^2 \\
&= 2\mathbb{E}\left\|Z_1\right\|^2 + \mathbb{E}\left\|Z_2\right\|^2 + \mathbb{E}\left\|Z_2 + Z_3 + Z_4\right\|^2 \\
&\overset{(b)}{\leq} 2\mathbb{E}\left\|Z_1\right\|^2 + \mathbb{E}\left\|Z_2\right\|^2 + 3\sum_{j=2}^4 \mathbb{E}\left\|Z_j\right\|^2 ,
\end{aligned}$$

where $(a)$ holds because $2Z_1^T Z_2 - \left\|Z_1\right\|^2 - \left\|Z_2\right\|^2 = -\left\|Z_1 - Z_2\right\|^2 \leq 0$ and $(b)$ holds because of Lemma 3. This completes the proof. □

**Lemma 8.** *Suppose $k \in [T]$ and let $\{h_q : \mathbb{R}^p \to \mathbb{R}^p\}_{q=0}^{k-1}$ be a collection of non-random vector-valued functions. Then $\mathbb{E}\left[\left(\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)})\right)^T \left(\widetilde{\nabla} f_i(x_i^{(t,r)}) - \nabla f_i(x_i^{(t,r)})\right)\right] = 0$ for all $i \in [n]$ and all $q, r \in \{0, 1, \ldots, k-1\}$ with $q \neq r$.*

*Proof.* Without loss of generality, suppose $q > r$. Then note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)})\right)^T \left(\widetilde{\nabla} f_i(x_i^{(t,r)}) - \nabla f_i(x_i^{(t,r)})\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)})\right)^T \left(\widetilde{\nabla} f_i(x_i^{(t,r)}) - \nabla f_i(x_i^{(t,r)})\right) \,\middle|\, x_i^{(t,q)}, x_i^{(t,r)}, \widetilde{\nabla} f_i(x_i^{(t,r)})\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)}) \,\middle|\, x_i^{(t,q)}, x_i^{(t,r)}, \widetilde{\nabla} f_i(x_i^{(t,r)})\right]^T \left(\widetilde{\nabla} f_i(x_i^{(t,r)}) - \nabla f_i(x_i^{(t,r)})\right)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\mathbf{0}^T \left(\widetilde{\nabla} f_i(x_i^{(t,r)}) - \nabla f_i(x_i^{(t,r)})\right)\right] \\
&= 0,
\end{aligned}$$

where $(a)$ follows from Assumption 3. □

**Lemma 9.** *Suppose the learning rate satisfies $\eta_t \leq \frac{1}{\sqrt{2}T\beta}$ for all $t \geq 0$. Then, for all $i \in [n]$, we have $\sum_{k=0}^{T-1} \mathbb{E}\left\|x_i^{(t,k)} - \beta^{(t,k)}\right\|^2 \leq 2eT^2\eta_t^2 \left(\varrho^2 + 2\delta^2 + 8\beta^2 \mathbb{E}\left\|x^{(t)} - x^*\right\|^2\right)$, where $e$ is the exponential constant.*

*Proof.* Let $a_i^{(t,k)} := \mathbb{E}\left\|x_i^{(t,k)} - \beta^{(t,k)}\right\|^2$. Then, by the definitions of $x_i^{(t,k)}$ and $\beta^{(t,k)}$, we have the following for all $k \leq T$:

$$\begin{aligned}
a_i^{(t,k)} &= \mathbb{E}\left\|\eta_t \sum_{q=0}^{k-1} \left(\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(\beta^{(t,q)})\right)\right\|^2 \\
&= \eta_t^2 \mathbb{E}\left\|\sum_{q=0}^{k-1} \left(\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)})\right) + \sum_{q=0}^{k-1} \left(\nabla f_i(x_i^{(t,q)}) - \nabla f_i(\beta^{(t,q)})\right)\right\| \\
&\overset{(a)}{=} 2\eta_t^2 \left(\sum_{q=0}^{k-1} \mathbb{E}\left\|\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)})\right\|^2 + k\sum_{q=1}^{k-1} \mathbb{E}\left\|\nabla f_i(x_i^{(t,q)}) - \nabla f_i(\beta^{(t,q)})\right\|^2\right) \\
&\overset{(b)}{\leq} 2\eta_t^2 \left(k\varrho^2 + k\beta^2 \sum_{q=1}^{k-1} \mathbb{E}\left\|x_i^{(t,q)} - \beta^{(t,q)}\right\|^2\right) \\
&\leq 2T\eta_t^2\varrho^2 + 2T\eta_t^2\beta^2 \left(a_i^{(t,1)} + \cdots + a_i^{(t,k-1)}\right) \qquad (22)
\end{aligned}$$

where $(a)$ follows from Assumption 3 and from Lemmas 3 and 8 and the fact that $x_i^{(t,0)} = \beta^{(t,0)} = x^{(t)}$, and (b) follows from Assumption 2.

In addition, we have

$$\begin{aligned}
a_i^{(t,1)} &= \mathbb{E}\left\|\eta_t \widetilde{\nabla} f_i(x^{(t)}) - \eta_t \nabla f(x^{(t)})\right\|^2 \\
&\overset{(a)}{=} \eta_t^2 \left(\mathbb{E}\left\|\widetilde{\nabla} f_i(x^{(t)}) - \nabla f_i(x^{(t)})\right\|^2 + \mathbb{E}\left\|\nabla f_i(x^{(t)}) - \nabla f(x^{(t)})\right\|^2\right) \\
&\overset{(b)}{\leq} \eta_t^2\varrho^2 + \eta_t^2(2\delta^2 + 8\beta^2 \mathbb{E}\left\|x^{(t)} - x^*\right\|^2), \qquad (23)
\end{aligned}$$

where $(a)$ holds because of Triangle inequality and Assumption 3, and $(b)$ holds because of Lemma 4.1 and Lemma 3.

Now, as shown in Section 7.5 of [27], one can easily use induction to show that (22) implies that $a_i^{(t,k)} \leq 2Ta_i^{(t,1)}(1+2T\beta^2\eta_t^2)^{k-1}$. As a result, we have

$$\sum_{k=0}^{T-1} a_i^{(t,k)} \leq 2Ta_i^{(t,1)} \sum_{k=0}^{T-1}(1+2T\beta^2\eta_t^2)^T \overset{(a)}{\leq} 2T^2 a_i^{(t,1)} e^{2T^2\beta^2\eta_t^2} \overset{(b)}{\leq} 2eT^2 a_i^{(t,1)}$$

where $(a)$ is a consequence of the inequality $1+x \leq e^x$ and $(b)$ follows from $\eta_t < \frac{1}{\sqrt{2}T\beta}$. Invoking (23) now completes the proof. □

**Lemma 10.** *Suppose* $\eta < \frac{\mu}{\beta^2}$*. Then* $\mathbb{E}\left\|\beta^{(t,k)} - x^*\right\|^2 \leq (1-\mu\eta_t)^k \mathbb{E}\left\|x^{(t)} - x^*\right\|^2$ *for all* $0 \leq k \leq T-1$ *and all* $t \geq 1$*.*

*Proof.* We can repeat the steps used to prove Eq. (37) in [27] and then take expectations on both sides to obtain $\mathbb{E}\left\|\beta^{(t,k)} - x^*\right\|^2 \leq (1-\mu\eta_t)\mathbb{E}\left\|\beta^{(t,k-1)} - x^*\right\|^2$ for all $k \in [T-1]$ and $t \geq 1$. Since $\beta^{(t,0)} = x^{(t)}$, the required inequality is now easily proved by induction. □

**Lemma 11.** *The following holds for all* $i \in [n]$*,* $t \in \mathbb{N}$ *and* $0 \leq k \leq T$*:*

$$\mathbb{E}\left\|x_i^{(t,k)} - x^{(t)}\right\|^2 \leq 2T\varrho^2\eta_t^2 + 4eT(\varrho^2+2\delta^2)\eta_t^2 + 6\delta^2T^2\eta_t^2 + (27+4e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2. \tag{24}$$

*Proof.* We assume $k = T$ and omit the case $k < T$, which is handled using similar arguments. We have the following chain of inequalities:

$$\begin{aligned}
&\mathbb{E}\left\|x_i^{(t,T)} - x^{(t)}\right\|^2 \\
&= \eta_t^2 \mathbb{E}\left\|\sum_{q=0}^{T-1} \widetilde{\nabla} f_i(x_i^{(t,q)})\right\|^2 \\
&= \eta_t^2 \mathbb{E}\Bigg\|\sum_{q=0}^{T-1}(\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)})) + \sum_{q=0}^{T-1}(\nabla f_i(x_i^{(t,q)}) - \nabla f_i(\beta^{(t,q)})) \\
&\qquad + \sum_{q=0}^{T-1}(\nabla f_i(\beta^{(t,q)}) - \nabla f(\beta^{(t,q)})) + \sum_{q=0}^{T-1}(\nabla f(\beta^{(t,q)}) - \nabla f(x^*))\Bigg\|^2 \\
&\overset{(a)}{\leq} \eta_t^2 \Bigg(2\mathbb{E}\left\|\sum_{q=0}^{T-1} \widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)}))\right\|^2 + 4\mathbb{E}\left\|\sum_{q=0}^{T-1}(\nabla f_i(x_i^{(t,q)}) - \nabla f_i(\beta^{(t,q)}))\right\|^2 \\
&\qquad + 3\mathbb{E}\left\|\sum_{q=0}^{T-1}(\nabla f_i(\beta^{(t,q)}) - \nabla f(\beta^{(t,q)}))\right\|^2 + 3\mathbb{E}\left\|\sum_{q=0}^{T-1}(\nabla f(\beta^{(t,q)}) - \nabla f(x^*))\right\|^2\Bigg)
\end{aligned} \tag{25}$$

where $(a)$ follows from Lemma 7 and the fact that $\{\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x^{(t,q)})\}_{i=1}^n$ are zero-mean random vectors.

Next, note that $\mathbb{E}\left\|\sum_{q=0}^{T-1}(\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)}))\right\|^2 = \sum_{q=0}^{T-1} \mathbb{E}\left\|\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)})\right\|^2$ because of Lemma 8. Since $\mathbb{E}\left\|\widetilde{\nabla} f_i(x_i^{(t,q)}) - \nabla f_i(x_i^{(t,q)})\right\|^2 \le \varrho^2$, (25) implies that

$$\begin{aligned}
&\mathbb{E}\left\|x_i^{(t,T)} - x^{(t)}\right\|^2 \\
&\le \eta_t^2 (2T\varrho^2 + 4\mathbb{E}\left\|\sum_{q=0}^{T-1} (\nabla f_i(x_i^{(t,q)}) - \nabla f_i(\beta^{(t,q)}))\right\|^2 \\
&\qquad + 3\mathbb{E}\left\|\sum_{q=0}^{T-1} (\nabla f_i(\beta^{(t,q)}) - \nabla f(\beta^{(t,q)}))\right\|^2 + 3\mathbb{E}\left\|\sum_{q=0}^{T-1} (\nabla f(\beta^{(t,q)}) - \nabla f(x^*))\right\|^2 \Bigg) \\
&\overset{(a)}{\le} 2T\varrho^2\eta_t^2 + T\eta_t^2 \sum_{q=0}^{T-1} \left( 4\beta^2 \mathbb{E}\left\|x_i^{(t,q)} - \beta^{(t,q)}\right\|^2 + 3\left(2\delta^2 + 8\beta^2 \mathbb{E}\left\|\beta^{(t,q)} - x^*\right\|^2\right) \right. \\
&\qquad\qquad\qquad \left. + 3\beta^2 \left\|\beta^{(t,q)} - x^*\right\|^2 \right) \\
&\le 2T\varrho^2\eta_t^2 + T\eta_t^2 \sum_{q=0}^{T-1} \left( 4\beta^2 \mathbb{E}\left\|x_i^{(t,q)} - \beta^{(t,q)}\right\|^2 \right. \\
&\qquad\qquad\qquad \left. + 6\delta^2 T + 27\beta^2 \sum_{q=0}^{T-1} \mathbb{E}\left\|\beta^{(t,q)} - x^*\right\|^2 \right) \\
&= 2T\varrho^2\eta_t^2 + 4T\beta^2\eta_t^2 \sum_{q=0}^{T-1} \mathbb{E}\left\|x_i^{(t,q)} - \beta^{(t,q)}\right\|^2 \\
&\qquad\qquad\qquad + 6\delta^2 T^2\eta_t^2 + 27T\beta^2\eta_t^2 \sum_{q=0}^{T-1} \mathbb{E}\left\|\beta^{(t,q)} - x^*\right\|^2 \\
&\overset{(d)}{\le} 2T\varrho^2\eta_t^2 + 8eT^3\beta^2\eta_t^4 \left(\varrho^2 + 2\delta^2 + 8\beta^2 \mathbb{E}\left\|x^{(t)} - x^*\right\|^2\right) \\
&\quad + 6\delta^2 T^2\eta_t^2 + 27T\beta^2\eta_t^2 \mathbb{E}\left\|x^{(t)} - x^*\right\|^2 \sum_{q=0}^{T-1} (1-\mu\eta_t)^q \\
&\overset{(e)}{\le} 2T\varrho^2\eta_t^2 + 4eT(\varrho^2 + 2\delta^2)\eta_t^2 + 6\delta^2 T^2\eta_t^2 + (27+4e)T^2\beta^2\eta_t^2 \mathbb{E}\left\|x^{(t)} - x^*\right\|^2,
\end{aligned}$$

where $(a)$ follows from Lemmas 3 and 4.1 and Assumption 2, $(d)$ follows from Lemmas 9 and 10 and from the assumption that $\eta_t < \frac{1}{\sqrt{2}T\beta}$, and $(e)$ follows from the observations that $2T^2\eta_t^2\beta^2 < 1$ and

$$\sum_{k=0}^{T-1} (1-\mu\eta_t)^k = \frac{1-(1-\mu\eta_t)^T}{1-(1-\mu\eta_t)} \le \frac{1-(1-T\mu\eta_t)}{\mu\eta_t} = T. \tag{26}$$

□

We are now ready to prove Proposition 4.3.

## Proof of Proposition 4.3

*Proof.* The definitions of $x^{(t)}$, $\bar{x}^{(t)}$, $X_{\text{diff}}(t)$, and $v(t)$ imply that $x^{(t+1)} = x^{(t)} + X_{\text{diff}}(t)A^T(t)v(t)$ and $\bar{x}^{(t+1)} = x^{(t)} + \frac{1}{n}X_{\text{diff}}(t)\mathbf{1}$. Consequently,

$$
\begin{aligned}
&\mathbb{E}\left\|x^{(t+1)} - \bar{x}^{(t+1)}\right\|^2 \\
&= \mathbb{E}\left\|X_{\text{diff}}(t)\left(A^T(t)v(t) - \frac{1}{n}\mathbf{1}\right)\right\|^2 \\
&\overset{(a)}{\le} \mathbb{E}\left\|X_{\text{diff}}(t)\right\|^2 \mathbb{E}\left\|A^T(t)v(t) - \frac{1}{n}\mathbf{1}\right\|^2 \\
&\overset{(b)}{\le} \left(\sum_{i=1}^n \mathbb{E}\left\|x_i^{(t,T)} - x^{(t)}\right\|^2\right)\left(\frac{1}{m(t)} - \frac{1}{n}\right)\sum_{\ell=1}^c \frac{n_\ell}{n}\phi_\ell(t) \\
&\overset{(c)}{\le} n\left(2T\varrho^2\eta_t^2 + 4eT(\varrho^2 + 2\delta^2)\eta_t^2 + 6\delta^2T^2\eta_t^2 + (27+4e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2\right)\left(\frac{1}{m(t)} - \frac{1}{n}\right)\sum_{\ell=1}^c \frac{n_\ell}{n}\phi_\ell(t) \\
&= \left(2T\varrho^2\eta_t^2 + 4eT(\varrho^2 + 2\delta^2)\eta_t^2 + 6\delta^2T^2\eta_t^2 + (27+4e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2\right)\phi(t),
\end{aligned}
$$

where $\phi(t) = \left(\frac{n}{m(t)} - 1\right)\sum_{\ell=1}^c \frac{n_\ell}{n}\phi_\ell(t)$ as defined in (5), $(a)$ holds because $X_{\text{diff}}(t)$, which is determined by $\{x^{(i,T)} - x^{(t)}\}_{i=1}^n$, is independent of the random subset of nodes sampled by the PS and hence also independent of $v(t)$, $(b)$ follows from Lemmas 4 and 6, and $(c)$ follows from Lemma 11. □

## Some Other Auxiliary Lemmas

**Lemma 12.** *Let $\bar{x}^{(t,k)} := \frac{1}{n}\sum_{i=1}^n x_i^{(t,k)}$ for $0 \le k \le T$ and $t \ge 1$. Then*

$$
\begin{aligned}
\mathbb{E}\left\|\bar{x}^{(t,k+1)} - x^*\right\|^2 \le (1 - \mu\eta_t)\mathbb{E}\left\|\bar{x}^{(t,k)} - x^*\right\|^2 + \eta_t^2\frac{\varrho^2}{n} + 6\beta\Gamma\eta_t^2 + 4T\varrho^2\eta_t^2 \\
+ 8eT(\varrho^2 + 2\delta^2)\eta_t^2 + 12\delta^2T^2\eta_t^2 + (54+8e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2
\end{aligned}
$$

*for all $0 \le k \le T$ and $t \ge 1$, where $\Gamma := \min_{x\in\mathbb{R}^p} f(x) - \frac{1}{n}\sum_{i=1}^n \min_{x\in\mathbb{R}^p} f_i(x)$.*

*Proof.* We first note that

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|x_i^{(t,k)} - \bar{x}^{(t,k)}\right\|^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|x_i^{(t,k)} - \frac{1}{n}\sum_{j=1}^n x_j^{(t,k)}\right\|^2\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|x_i^{(t,k)} - x^{(t)}\right\|^2\right] \\
&\overset{(b)}{\le} 2T\varrho^2\eta_t^2 + 4eT(\varrho^2 + 2\delta^2)\eta_t^2 + 6\delta^2T^2\eta_t^2 + (27+4e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2, \qquad (27)
\end{aligned}
$$

where $(a)$ follows from the fact that the arithmetic mean of a finite set of vectors $\{z_i\}_{i=1}^n$ is the minimizer of the unweighted mean-square loss function $\mathbb{R} \ni y \to \tilde{\ell}(y) = \frac{1}{n}\sum_{i=1}^n\|z_i - y\|^2$, and $(b)$ follows from Lemma 11.

An application of Lemma 1 of [17] now yields

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(t,k+1)} - x^*\right\|^2 \\
&\leq (1-\mu\eta_t)\mathbb{E}\left\|\bar{x}^{(t,k)} - x^*\right\|^2 + \eta_t^2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \widetilde{\nabla} f_i(x_i^{(t,k)}) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^{(t,k)})\right\|^2 \\
&\quad + 6\beta\Gamma\eta_t^2 + 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|x_i^{(t,k)} - x^{(t)}\right\|^2\right] \\
&\overset{(a)}{\leq} (1-\mu\eta_t)\mathbb{E}\left\|\bar{x}^{(t,k)} - x^*\right\|^2 + \eta_t^2\frac{\varrho^2}{n} + 6\beta\Gamma\eta_t^2 \\
&\quad + 2(2T\varrho^2\eta_t^2 + 4eT(\varrho^2+2\delta^2)\eta_t^2 + 6\delta^2T^2\eta_t^2 + (27+4e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2),
\end{aligned}$$

where $(a)$ holds because of (27) and the fact that the data mini-batches used to compute stochastic gradients are independent across clients. □

**Lemma 13.** *Let $\bar{x}^{(t)} := \frac{1}{n}\sum_{i=1}^n x_i^{(t-1,T)}$ for all $t \geq 1$. Then*

$$\mathbb{E}\left\|\bar{x}^{(t+1)} - x^*\right\|^2 \leq \left((1-\mu\eta_t)^T + (54+8e)T^3\beta^2\eta_t^2\right)\mathbb{E}\left\|x^{(t)} - x^*\right\|^2 + Th(\eta_t),$$

*where $h(\eta_t) := \left(\frac{\varrho^2}{n} + 6\beta\Gamma + 4T\varrho^2 + 8eT(\varrho^2+2\delta^2) + 12\delta^2T^2\right)\eta_t^2$.*

*Proof.* Let $\theta^{(t,k)} := \mathbb{E}\left\|\bar{x}^{(t,k)} - x^*\right\|^2$. Then we know from Lemma 12 that

$$\theta^{(t,k+1)} \leq (1-\mu\eta_t)\theta^{(t,k)} + \tilde{h}(\eta_t),$$

where $\tilde{h}(\eta_t) = h(\eta_t) + (54+8e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2$. By induction on $k$, we can now show that $\theta^{(t,k)} \leq (1-\mu\eta_t)^k\theta^{(t,0)} + \sum_{q=0}^{k-1}(1-\mu\eta_t)^q\tilde{h}(\eta_t)$. As a result,

$$\theta^{(t,T)} \leq (1-\mu\eta_t)^T\theta^{(t,0)} + \tilde{h}(\eta_t)\sum_{q=0}^{T-1}(1-\mu\eta_t)^q \leq (1-\mu\eta_t)^T\theta^{(t,0)} + T\tilde{h}(\eta_t),$$

where the second inequality follows from (26). Noting that $\bar{x}^{(t,T)} = \bar{x}^{(t+1)}$ and $\bar{x}^{(t,0)} = x^{(t)}$ now yields the required inequality. □

## Proof of Proposition 4.4

*Proof.* Lemma 13 implies that

$$\begin{aligned}
\mathbb{E}\left\|\bar{x}^{(t+1)} - x^*\right\|^2 &\leq \left((1-\mu\eta_t)^T + (54+8e)T^3\beta^2\eta_t^2\right)\mathbb{E}\left\|x^{(t)} - x^*\right\|^2 \\
&\quad + T\left(\frac{\varrho^2}{n} + 6\beta\Gamma + 4T\varrho^2 + 8eT(\varrho^2+2\delta^2) + 12\delta^2T^2\right)\eta_t^2
\end{aligned} \tag{28}$$

On the other hand, Proposition 4.3 implies

$$\begin{aligned}
&\mathbb{E}\left\|x^{(t+1)} - \bar{x}^{(t+1)}\right\|^2 \\
&\leq \left(2T\varrho^2\eta_t^2 + 4eT(\varrho^2+2\delta^2)\eta_t^2 + 6\delta^2T^2\eta_t^2 + (27+4e)T^2\beta^2\eta_t^2\mathbb{E}\left\|x^{(t)} - x^*\right\|^2\right)\phi(t).
\end{aligned} \tag{29}$$

In view of Lemma 4.2, the inequalities (28) and (29) collectively imply

$$\begin{aligned}\mathbb{E}\left\|x^{(t+1)}-x^*\right\|^2 &\le \left((1-\mu\eta_t)^T+(27+4e)T^2\beta^2\eta_t^2(2T+\phi(t))\right)\mathbb{E}\left\|x^{(t)}-x^*\right\|^2\\ &\quad + T\left(\frac{\varrho^2}{n}+6\beta\Gamma+4T\varrho^2+8eT(\varrho^2+2\delta^2)+12\delta^2T^2\right)\eta_t^2\\ &\quad + \left(2T\varrho^2+4eT(\varrho^2+2\delta^2)+6\delta^2T^2\right)\phi(t)\eta_t^2\end{aligned}$$

as required. □

## Lemma Used in the Proof of Theorem 4.5

**Lemma 14.** *For any $y \in \mathbb{N}$ and $q \in (0, \frac{4}{y-3})$, we have $(1-q)^y \le 1 - qy + \frac{y(y-1)}{2}q^2$.*

*Proof.* We can easily verify that the inequality in question reduces to an equality for $y \in \{1,2\}$. Also, for $y = 3$ we can use the identity $(1-q)^3 = 1-3q+3q^2-q^3$ to verify the given inequality. So, we assume $y \ge 4$ for the remainder of this proof.

Observe that the Taylor series expansion of $(1-q)^y$ yields

$$\begin{aligned}(1-q)^y &= 1 - qy + \frac{y(y-1)}{2}q^2\\ &\quad - \sum_{\ell=1}^{\lceil\frac{y-1}{2}\rceil}\left(\frac{y(y-1)\cdots(y-2\ell)q^{2\ell+1}}{(2\ell+1)!} - \frac{y(y-1)\cdots(y-(2\ell+1))q^{2\ell+2}}{(2\ell+2)!}\right).\end{aligned}$$

Therefore, it suffices to show that

$$\frac{y(y-1)\cdots(y-2\ell)q^{2\ell+1}}{(2\ell+1)!} \ge \frac{y(y-1)\cdots(y-(2\ell+1))q^{2\ell+2}}{(2\ell+2)!} \tag{30}$$

for each $\ell \in \{1,2,\ldots,\lceil\frac{y-1}{2}\rceil\}$ whenever $q < \frac{4}{y-3}$. Note that (30) is equivalent to $q \le \frac{2\ell+2}{y-(2\ell+1)}$ for $\ell$ in the appropriate range. Since the right hand side of this simpler inequality is increasing in $\ell$ on the set $\{1,2,\ldots,\lceil\frac{y-1}{2}\rceil\}$, it suffices to prove the inequality for $\ell = 1$, i.e., we need to show that $q \le \frac{4}{y-3}$, which is precisely the assumption made by the lemma. □

## Proof of Theorem 4.5

*Proof.* The first broad step is to examine the coefficient of $\mathbb{E}\left\|x^{(t)}-x^*\right\|^2$ in (9) and bound it by a quadratic function of $\eta_t$. To this end, we invoke Lemma 14 in light of the condition $\eta_t \le \frac{4}{\mu(T-3)}$ (which is already satisfied due to our assumption $\eta_t = \frac{4}{T\mu(t+t_1)}$) to obtain

$$(1-\mu\eta_t)^T \le 1 - T\mu\eta_t + \frac{T(T-1)}{2}\mu^2\eta_t^2. \tag{31}$$

As a result, we have

$$(1-\mu\eta_t)^T + (27+4e)T^2\beta^2\eta_t^2(2T+\phi(t)) \le 1 - a\eta_t + b\eta_t^2, \tag{32}$$

where $a := T\mu$ and $b := \frac{T(T-1)\mu}{2}\mu^2 + (2T+\phi_{\max})(27+4e)T^2\beta^2\eta_t^2$. Since $1 - a\eta_t + b\eta_t^2 \le 1 - \frac{a}{2}\eta_t$ for $0 \le \eta_t \le \frac{a}{2b}$, (32) further results in $(1-\mu\eta_t)^T + (27+4e)T^2\beta^2\eta_t^2(2T+\phi(t)) \le 1 - \frac{a}{2}\eta_t$. For $\eta_t = \frac{4}{T\mu(t+t_1)} = \frac{4}{a(t+t_1)}$, this is equivalent to

$$(1-\mu\eta_t)^T + (27+4e)T^2\beta^2\eta_t^2(2T+\phi(t)) \le 1 - \frac{2}{t+t_1}, \tag{33}$$

where $t_1$ satisfies $\frac{4}{a(t+t_1)} \leq \frac{a}{2b}$, i.e.,

$$t + t_1 \geq \frac{4T(T-1)\mu^2 + (16T + 8\phi_{\max})(27+4e)T^2\beta^2}{T^2\mu^2} = 4\left(1 - \frac{1}{T}\right) + (16T + 8\phi_{\max})\left(\frac{\beta}{\mu}\right)^2,$$

a condition satisfied by our choice of $t_1$ in the theorem.

We now combine (33) with (9) and replace $\eta_t$ with $\frac{4}{T\mu(t+t_1)}$ so as to obtain

$$\mathbb{E}\left\|x^{(t+1)} - x^*\right\|^2 \leq \left(1 - \frac{2}{t+t_1}\right)\mathbb{E}\left\|x^{(t)} - x^*\right\|^2 + \frac{\alpha}{(t+t_1)^2}, \tag{34}$$

where $\alpha := \frac{16}{\mu^2 T}\left(\frac{\varrho^2}{n} + \beta\Gamma\right) + (32T + 16\phi_{\max})\left(\frac{2}{T}\left(\frac{\varrho}{\mu}\right)^2 + \frac{4e}{T}\left(\left(\frac{\varrho}{\mu}\right)^2 + 2\left(\frac{\delta}{\mu}\right)^2\right) + 6\left(\frac{\delta}{\mu}\right)^2\right)$.

An application of Lemma 5 of [27] to (34) now yields the desired upper bound. □

## Proof of Proposition 5.2

*Proof.* We first prove (15). Recall that $\sigma_1^2$ and $\sigma_2^2$ are, respectively, the spectral radius and the second-largest eigenvalue of $AA^\top$. Using the non-negativity of $A$ and the fact that every matrix norm is an upper bound on the spectral radius, we have

$$\begin{aligned}\sigma_1^2 &\leq \left\|AA^\top\right\|_\infty = \max_{i\in[s]}\sum_{j=1}^{s}(AA^\top)_{ij} = \max_{i\in[s]}\sum_{j=1}^{s}\sum_{k=1}^{s} a_{ik}a_{jk} \\ &= \max_{i\in[s]}\sum_{k=1}^{s} a_{ik}\left(\sum_{j=1}^{s} a_{jk}\right) \overset{(a)}{=} \max_{i\in[s]}\sum_{k\in\mathcal{N}_i^-} a_{ik} \overset{(b)}{\leq} \frac{d_{\max}^{\text{in}}}{d_{\min}^+} = 1 + \varphi,\end{aligned}$$

where $(a)$ follows from the column-stochasticity of $A$ and the fact that $a_{ik} = 0$ for all $k \notin \mathcal{N}_i^-$, and $(b)$ holds because $|\mathcal{N}_i^-| \leq d_{\max}^{\text{in}}$ and because $a_{ik} = (d_i^+)^{-1} \leq (d_{\min}^+)^{-1}$.

To prove (16), we first compute a set of quantities defined in terms of $A$. We then show that $A^\top A$ is positive. Finally, we apply Corollary 3.1.1 of [21] to $A^\top A$ using the quantities. We remark that Corollary 3.1.1 of [21] is based on Theorem 3.1 of [21] , a result that bounds the maximum entry of the Perron eigenvector of a given positive matrix in terms of the sum of the entries of the same eigenvector (since the eigenvector is unique up to scaling as per the Perron-Frobenius theorem [23]). Equivalently, the result bounds the minimum entry of the Perron eigenvector in terms of the maximum entry and the sum of the remaining entries of the eigenvector.

First, we let $\tilde{m} := \min_{i,j\in[s]}(A^\top A)_{ij}$ and bound the same:

$$\begin{aligned}\tilde{m} &= \min_{i,j\in[s]}\sum_{k=1}^{s} a_{ki}a_{kj} = \min_{i,j\in[s]}\sum_{k\in\mathcal{N}_i^+\cap\mathcal{N}_j^+} a_{ki}a_{kj} \\ &\geq \min_{i,j\in[s]}\sum_{k\in\mathcal{N}_i^+\cap\mathcal{N}_j^+}\left(\frac{1}{d_{\max}}\right)^2 \overset{(a)}{\geq} \frac{(2\alpha-1)s}{d_{\max}^2} = \frac{(2\alpha-1)(1-\varepsilon)^2}{\alpha^2 s},\end{aligned}$$

where $(a)$ holds because $|\mathcal{N}_i^+ \cap \mathcal{N}_i^-| = |\mathcal{N}_i^+| + |\mathcal{N}_j^+| - |\mathcal{N}_i^+ \cup \mathcal{N}_j^+| \geq \alpha s + \alpha s - s$.

On the other hand, we have

$$\tilde{m} = \min_{i,j\in[s]}\sum_{k=1}^{s} a_{ki}a_{kj} \leq \frac{1}{d_{\min}^+}\min_{i,j\in[s]}\sum_{k=1}^{s} a_{kj} = \frac{1}{d_{\min}^+} = \frac{1}{\alpha s}.$$

Next, we have

$$\begin{aligned} M_j := \max_{i\in[s]}(A^\top A)_{ij} &= \max_{i\in[s]}\sum_{k=1}^{s} a_{ki}a_{kj} \le \sum_{k=1}^{s}\left(\max_{i\in[s]} a_{ki}\right) a_{kj} \\ &\le \sum_{i=1}^{s}\frac{1}{d_{\min}^{+}}\cdot a_{kj} = \frac{1}{d_{\min}^{+}} = \frac{1}{\alpha s}. \end{aligned}$$

Arguing along similar lines results in the following for all $j \in [s]$:

$$\begin{aligned} c^{(j)} := \sum_{i=1}^{s}(A^\top A)_{ij} &= \sum_{i=1}^{s}\sum_{k=1}^{s} a_{ki}a_{kj} \ge \frac{1}{d_{\max}}\sum_{i=1}^{s}\sum_{k=1}^{s} a_{kj} \\ &= \frac{s}{d_{\max}} = \frac{1-\varepsilon}{\alpha}. \end{aligned} \tag{35}$$

Hence, $\theta_j := \min_j \left(c^{(j)} - M_j\right) \ge \frac{1-\varepsilon}{\alpha} - \frac{1}{\alpha s}$.

Using arguments similar to those used in (35), we can prove that

$$\rho_m := \min_{i\in[s]}\sum_{j=1}^{s}(A^\top A)_{ij} \ge \frac{1-\varepsilon}{\alpha}. \tag{36}$$

Likewise, we have

$$\begin{aligned} \rho_M := \max_{i\in[s]}\sum_{j=1}^{s}(A^\top A)_{ij} &= \max_{i\in[s]}\sum_{j=1}^{s}\sum_{k=1}^{s} a_{ki}a_{kj} \\ &\le \frac{1}{d_{\min}^{+}}\sum_{j=1}^{s}\sum_{k=1}^{s} a_{kj} = \frac{s}{d_{\min}^{+}} = \frac{1}{\alpha}, \end{aligned}$$

which also implies that $\rho_m \le \frac{1}{\alpha}$.

As another implication of (36), we have

$$\gamma := \sum_{i=1}^{s}\sum_{j=1}^{s}(A^\top A)_{ij} \ge s\min_{i\in[s]}\sum_{j=1}^{s}(A^\top A)_{ij} = s\rho_m \ge \frac{s(1-\varepsilon)}{\alpha}.$$

Combining all of the above bounds appropriately with Corollary 3.1.1 of [21] culminates in the following:

$$\begin{aligned} \sigma_2^2 &\le \sigma_1^2 - \tilde{m}\frac{\sigma_1^2\left(\sigma_1^2 - \rho_m + \tilde{m}s\right)}{\left(\sigma_1^2 - \rho_m\right)\left(\sigma_1^2 - \theta\right) + \tilde{m}\left(s\sigma_1^2 - \gamma + \rho_M\right)} \\ &\overset{(a)}{\le} 1 + \varphi - \left(\frac{2}{\alpha} - \frac{1}{\alpha^2}\right)\frac{(1-\varepsilon)^2}{s}\cdot\frac{1\cdot\left(1 - \frac{1}{\alpha} + \left(\frac{2}{\alpha} - \frac{1}{\alpha^2}\right)(1-\varepsilon)^2\right)}{h(\alpha,\varphi,\varepsilon,s)}, \end{aligned}$$

where

$$\begin{aligned} h(\alpha,\varphi,\varepsilon,s) := &\left(1 + \varphi - \frac{1-\varepsilon}{\alpha}\right)\left(1 + \varphi - \frac{1-\varepsilon}{\alpha} + \frac{1}{\alpha s}\right) \\ &+ \frac{1}{\alpha s}\left(s(1+\varphi) - \frac{s(1-\varepsilon)}{\alpha} + \frac{1}{\alpha}\right). \end{aligned}$$

Replacing $1 + \varphi - \frac{1-\varepsilon}{\alpha}$ with $\varepsilon_{\text{net}} - \alpha_{-1}$ and $\frac{2}{\alpha} - \frac{1}{\alpha^2}$ with $1 - \alpha_{-1}^2$ in the above inequality gives the desired bound on $\sigma_2$.

□