Scalable Decentralized Algorithms for Online Personalized Mean Estimation

# Scalable Decentralized Algorithms for Online Personalized Mean Estimation

**Franco Galante**[1], **Giovanni Neglia**[2], **Emilio Leonardi**[1]

[1]Politecnico di Torino
[2]INRIA
franco.galante@polito.it, giovanni.neglia@inria.fr, emilio.leonardi@polito.it

**Abstract**

In numerous settings, agents lack sufficient data to learn a model directly. Collaborating with other agents may help, but introduces a bias-variance trade-off when local data distributions differ. A key challenge is for each agent to identify clients with similar distributions while learning the model, a problem that remains largely unresolved. This study focuses on a particular instance of the overarching problem, where each agent collects samples from a real-valued distribution over time to estimate its mean. Existing algorithms face impractical per-agent space and time complexities (linear in the number of agents $|\mathcal{A}|$). To address scalability challenges, we propose a framework where agents self-organize into a graph, allowing each agent to communicate with only a selected number of peers $r$. We propose two collaborative mean estimation algorithms: one employs a consensus-based approach, while the other uses a message-passing scheme, with complexity $\mathcal{O}(r)$ and $\mathcal{O}(r \cdot \log |\mathcal{A}|)$, respectively. We establish conditions for both algorithms to yield asymptotically optimal estimates and we provide a theoretical characterization of their performance.

**Code** — https://github.com/Franco-Galante/scalable-decentralized-algorithms-AAAI25

## 1 Introduction

Users' devices have become increasingly sophisticated and generate vast amounts of data. This wealth of data has enabled the development of accurate and complex models. However, it has also introduced challenges related to security, privacy, real-time processing, and resource management. In response, Federated Learning (FL) has emerged as a key privacy-preserving approach for collaborative model training (Kairouz et al. 2021; Li et al. 2020). While traditional FL methods aim to develop a single model for all clients, the statistical diversity of clients' datasets has led to the development of *personalized* models, designed to better align with the data distributions of individual clients (e.g., Ghosh et al. 2020; Fallah, Mokhtari, and Ozdaglar 2020; Li et al. 2021; Marfoq et al. 2021; Ding and Wang 2022).

Many personalized FL strategies group clients into clusters and then tailor a model for each cluster (e.g., Ghosh et al. 2020; Sattler, Müller, and Samek 2021; Ding and Wang 2022). Ideally, clustering would group clients with similar local optimal models. However, since the optimal models are unknown a priori, model learning and cluster identification become deeply interconnected tasks. Various studies have suggested empirical measures of similarity as a workaround (e.g., Ghosh et al. 2020; Sattler, Müller, and Samek 2021), while others rely on presumed knowledge of distances across data distributions (e.g. Ding and Wang 2022; Even, Massoulié, and Scaman 2022). Nonetheless, accurately estimating these distances, especially within an FL framework where clients may possess limited data, proves to be particularly challenging. These estimation difficulties are well documented in the literature—e.g., by Even, Massoulié, and Scaman (2022) [Sec. 6]—highlighting the problem of identifying similar clients for collaborative model learning as a significant yet unresolved issue.

In this paper, we focus on a fundamental aspect of the broader challenge: estimating the mean of an $\mathbb{R}^K$-valued distribution. This problem is often regarded as the archetypal federated learning problem (Dorner et al. 2024; Tsoy et al. 2024; Grimberg et al. 2021), but it also holds significant practical relevance across various fields, such as smart agriculture, grid management, and healthcare, where multiple sensors collect private, noisy data on identical or related variables (Adi et al. 2020).

We consider an online, decentralized scenario where, at each time slot, clients receive new samples and exchange information with a limited number of peers. To the best of our knowledge, the state-of-the-art method in this setting is the Collaborative Mean Estimation algorithm (ColME) by Asadi et al. (2022). Unfortunately, ColME faces scalability issues in large systems, as both its per-agent space and time complexities are linear in the number of clients $|\mathcal{A}|$. Moreover, in its current form, ColME is applicable only to scalar mean estimation problems ($K = 1$) and its convergence guarantees only hold for sub-Gaussian data distributions.

We extend the methodology proposed by Asadi et al. (2022) to accommodate multidimensional data drawn from the broader class of distributions with bounded fourth moment. To address ColME's scalability challenges, we propose that clients self-organize into a network where each client communicates with at most $r$ neighbors. Over time, this set of neighbors is pruned as clients progressively ex-

clude the less similar ones. In this framework, we introduce two collaborative mean estimation algorithms: one based on consensus and the other on a message-passing scheme. The complexities of these algorithms are $\mathcal{O}(r)$ and $\mathcal{O}(r \cdot \log |\mathcal{A}|)$, respectively. We demonstrate that, despite each client exchanging information with only $r \ll |\mathcal{A}|$ neighbors, it is possible to achieve a convergence speedup for mean estimates by a factor of $\Omega(|\mathcal{A}|^{1/2-\phi})$, where $\phi$ can be made arbitrarily close to 0.

Lastly, we conduct preliminary experiments demonstrating how our algorithms can be adapted to federatedly learn more general machine learning models.

## 2 Related Work

For an overview of personalized federated learning, see the recent survey by Tan et al. (2023). Here, we highlight only the most relevant approaches related to this paper.

Ghosh et al. (2020) and Sattler, Müller, and Samek (2021) were the first to propose clustered FL algorithms, which divide the clients based on the similarity of their data distributions. Similarity is empirically evaluated by the Euclidean distance between local models and by the cosine similarity of their updates. Ding and Wang (2022) study more sophisticated clustering algorithms assuming that clients can efficiently estimate some specific (pseudo)-distances across local distributions (i.e., the integral probability metrics).

Beaussart et al. (2021), Chayti et al. (2022), Grimberg et al. (2021), and Even, Massoulié, and Scaman (2022) consider decentralized approaches, which allow each client to learn a personal model relying on a specific convex combination of information (gradients) from other clients. In particular, Even, Massoulié, and Scaman (2022) prove that collaboration can at most speed up the convergence time linearly in the number of similar agents and provide algorithms, which, under a priori knowledge of pairwise client distributions' distances, achieve such speedup. The authors recognize the complexity of estimating these distances and provide practical estimation algorithms for linear regression problems, which asymptotically achieve the same speedup, scaling the number of clients but maintaining the number of clusters fixed. The generalization properties of personalized models obtained by convex combinations of clients' models are studied in Mansour et al. (2020), while Donahue and Kleinberg (2021) look at the problem through the lens of game theory. The work most similar to ours is Asadi et al. (2022), which we describe in detail in the next section.

## 3 Model and Background

Table 1 lists the most important symbols used throughout the paper. Superscripts are added to variables to indicate whether they pertain to C-ColME (C), B-ColME (B), or both approaches (D).

We consider a set $\mathcal{A}$ of agents (computational units). At each time instant $t$, an agent $a \in \mathcal{A} = \{1, 2, \cdots, |\mathcal{A}|\}$ generates a new sample $\boldsymbol{x}_a^t \in \mathbb{R}^K$, with $K \in \mathbb{N}$, drawn i.i.d. from a distribution $D_a$ with expected value $\boldsymbol{\mu}_a = \mathbb{E}[\boldsymbol{x}_a^t]$. Expected values are not necessarily distinct across agents. Indeed, given two agents $a$ and $a'$, and a norm $||\cdot||$ in $\mathbb{R}^K$, we denote the gap between the agents' *true* means by $\Delta_{a,a'} := ||\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'}||$. Let $\mathcal{C}_a$ be the group of agents with the same *true* mean as $a$, (i.e., those for which $\Delta_{a,a'} = 0$). In the following, we will refer to $\mathcal{C}_a$ as the 'similarity class of $a$.'

The goal of each agent $a \in \mathcal{A}$ is to estimate its mean $\boldsymbol{\mu}_a$. To this end, at each time $t$ the agent can compute its *local* mean estimate $\bar{\boldsymbol{x}}_{a,a}^t = \frac{1}{t}\sum_{t'=1}^{t} \boldsymbol{x}_a^{t'}$ over the $t$ available samples. Additionally, the agent can obtain a more accurate estimate by leveraging information from other agents in $\mathcal{A}$, provided it can identify those who share the same true mean.

For the scalar case, (i.e., when $K = 1$) Asadi et al. (2022) proposed ColME as a collaborative algorithm for mean estimation. It relies on two key steps, executed concurrently: i) the identification of the similarity classes; ii) the collaboration with agents *believed to belong* to the same class to improve the local estimate. To ensure direct comparability with the results in (Asadi et al. 2022) and simplify notation, in what follows, we focus on the scalar case. Readers interested in the general case can find the analysis in Appendix C. All the appendices referred to in this work are accessible at (Galante, Neglia, and Leonardi 2024).

| Symbol | Description |
|---|---|
| $\mathcal{A}$ | Set of agents (computational units) |
| $D_a$ | Distribution of agent $a \in \mathcal{A}$ |
| $\boldsymbol{\mu}_a$ | True mean of distribution $D_a$, i.e., $\boldsymbol{\mu}_a = \mathbb{E}[\boldsymbol{x}_a^t]$ |
| $\bar{\boldsymbol{x}}_a^t$ | Vectorial sample $\bar{\boldsymbol{x}}_a^t \in \mathbb{R}^K$ drawn from $D_a$ |
| $\Delta_{a,a'}$ | Gap between agents $a$ and $a'$ true means |
| $\mathcal{C}_a$ | Similarity class of $a$, $\mathcal{C}_a = \{a' \in \mathcal{A} \mid \Delta_{a,a'} = 0\}$ |
| $\mathcal{C}_a^t$ | Estimated similarity class of $a$ at time $t$ |
| $\bar{x}_{a,a}^t$ | Local mean estimate of agent $a$ at $t$ |
| $\bar{x}_{a,a'}^t$ | Local mean estimate of agent $a'$ by agent $a$ at $t$ |
| $n_{a,a'}^t$ | Number of samples used to compute $\bar{x}_{a,a'}^t$ at $t$ |
| $\hat{\mu}_a^t$ | Collaborative mean estimate at $t$ |
| $\beta_\gamma(n)$ | Width of the confidence interval |
| $d_\gamma^t(a,a')$ | Optimistic distance between agents $a$ and $a'$ |
| $\zeta_a$ | Time to identify same-class neighbors w.h.p. |
| $\tau_a$ | Time to obtain $(\epsilon, \delta)$ convergence |
| $\mathcal{G}$ | Collaborative graph $\mathcal{G}(\mathcal{A}, \mathcal{E})$ |
| $\mathcal{G}^t$ | Pruned collaborative graph $\mathcal{G}^t(\mathcal{A}, \mathcal{E}^t)$ |
| $\mathcal{N}_a$ | Neighborhood of agent $a$ (or up to distance $d$: $\mathcal{N}_a^d$) |
| $r$ | Upper bound on $a$'s neighborhood size $\lvert\mathcal{N}_a\rvert$ |
| $\mathcal{CC}_a$ | Agents in the connected component of the subgraph induced by agents in $\mathcal{C}_a$ to which agent $a$ belongs |
| $\mathcal{CC}_a^d$ | Same as $\mathcal{CC}_a$ but with agents up to $d$-hops from $a$ |

Table 1: Notation Summary

**1) Identifying Similarity Classes.** We denote $\mathcal{C}_a^t$ as the set of agents that, at time $t$, agent $a$ *estimates* to belong to its similarity class $\mathcal{C}_a$. Specifically, agent $a$ includes agent $a'$ in its *estimated* similarity class $\mathcal{C}_a^t$ if their local mean estimates are sufficiently close. In general, at time $t$, agent $a$ does not have access to the most recent local mean estimate of agent $a'$ (computed over $t$ samples), but rather to a *stale* value $\bar{x}_{a,a'}^t$ computed over $n_{a,a'}^t$ samples, where $n_{a,a'}^t$ corresponds to the time when $a$

and $a'$ last communicated. Agent $a$ can then estimate its true mean and the true mean of agent $a'$ to belong to the confidence intervals $I_{a,a} = \left[\bar{x}^t_{a,a} - \beta_\gamma(t), \bar{x}^t_{a,a} + \beta_\gamma(t)\right]$, $I_{a,a'} = \left[\bar{x}^t_{a,a'} - \beta_\gamma(n^t_{a,a'}), \bar{x}^t_{a,a'} + \beta_\gamma(n^t_{a,a'})\right]$, respectively. As expected, the interval amplitude $\beta_\gamma(n)$ depends on the number of samples $n$ on which the empirical average is computed and on the target level of confidence 1-2$\gamma$ associated with the interval. Agent $a$ will then consider $a'$ to belong to its *estimated* similarity class $\mathcal{C}^t_a$ if the two intervals overlap, i.e., $I_{a,a} \cap I_{a,a'} \neq \emptyset$, or equivalently if the optimistic distance $d^t_\gamma(a, a')$ is zero or less:

$$d^t_\gamma(a, a') := |\bar{x}_{a,a} - \bar{x}_{a,a'}| - \beta_\gamma(t) - \beta_\gamma(n^t_{a,a'}) \leqslant 0. \quad (1)$$

To achieve the most accurate estimation of its similarity class, agent $a$ could retrieve from each other peer $a'$ the most recent estimate $\bar{x}^t_{a,a'}$ at each time $t$. However, this approach would result in a per-agent communication burden of $\mathcal{O}(|\mathcal{A}|)$. To mitigate this effect, node $a$ cyclically queries a *single* node from $\mathcal{C}^{t-1}_a$ at each time instant $t$, according to a *Round-Robin* scheme. This leads to the following update rule for $\mathcal{C}^t_a$:

$$\mathcal{C}^t_a = \{a' \in \mathcal{C}^{t-1}_a : d^t_\gamma(a, a') \leqslant 0\}, \quad (2)$$

and we observe that by construction $\mathcal{C}^t_a \subseteq \mathcal{C}^{t-1}_a$. Initially, the agent sets $\mathcal{C}^0_a = \mathcal{A}$ and progressively makes irreversible decisions to remove agents that are deemed too dissimilar.

**2) Estimating the Mean.** Each node $a$ computes an *estimate* $\hat{\mu}^t_a$ of its true mean $\mu_a$ combining the available estimates according to a *simple weighting* scheme, where the number of samples $n^t_{a,a'}$ are the weights:

$$\hat{\mu}^t_a = \sum_{a' \in \mathcal{C}^t_a} \frac{n^t_{a,a'}}{\sum_{a' \in \mathcal{C}^t_a} n^t_{a,a'}} \bar{x}^t_{a,a'}.$$

**ColME's theoretical guarantees.** Asadi et al. (2022) prove that if data distributions $\{D_a\}_{a \in \mathcal{A}}$ are sub-Gaussian and the amplitude of the confidence intervals is selected as:

$$\beta_\gamma(n) = \sigma \sqrt{\frac{2}{n}\left(1 + \frac{1}{n}\right) \ln(\sqrt{(n+1)}/\gamma)}, \quad (3)$$

then, with high probability, client $a$ correctly identifies its similarity class ($\mathcal{C}^t_a = \mathcal{C}_a$) and obtains an $\epsilon$-accurate estimate for $t$ larger than two opportune constants $\zeta_a$ and $\tau_a$, respectively (see Appendix A for the expressions and the results' statements).

**ColME's limitations.** Asadi et al. (2022) acknowledge in their paper the main limitations of ColME: each agent requires a memory footprint and computational complexity proportional to $|\mathcal{A}|$. Indeed, each agent $a$ must store all neighbors' local estimates $\bar{x}^t_{a,a'}$ and the corresponding sample counts $n^t_{a,a'}$. Additionally, as agent $a$ receives a new sample at each time $t$, it updates its local estimate $\bar{x}^t_{a,a}$ (and also $n^t_{a,a}$). This update affects the distance $d^t_\gamma(a, a')$, which thus must be recomputed for all $a' \in \mathcal{C}^t_a$. This leads to a per-agent time and space complexity of $\mathcal{O}(|\mathcal{A}|)$ per time slot, which becomes impractical in large-scale systems. Moreover, while the *Round-Robin* query scheme reduces the communication burden, it introduces a significant delay in the estimation as $\zeta_a \in \mathcal{O}(|\mathcal{A}|)$.

## 4 Scalable Algorithms over a Graph $\mathcal{G}$

Online mean estimation can be made $\tilde{\mathcal{O}}(1)$ with our *scalable* approaches: C-ColME (Sec. 4.1), and B-ColME (Sec. 4.2). Both algorithms consider agents $\mathcal{A}$ organized in a fixed graph $\mathcal{G}(\mathcal{A}, \mathcal{E})$ and restrict communication to pairs of agents adjacent in $\mathcal{G}$. Let $\mathcal{N}_a$ and $\mathcal{N}^d_a$ denote the set of neighbors of agent $a$ and the set of agents at distance at most $d$ from $a$, respectively. Let $r$ represent the maximum size of any agent's neighborhood in $\mathcal{G}$, i.e., $|\mathcal{N}_a| \leq r$ for all $a \in \mathcal{A}$.

Consider the subgraph $\mathcal{G}'$ of $\mathcal{G}$ induced by the agents in $\mathcal{C}_a$. Let $\mathcal{CC}_a$ denote the (initially unknown) set of agents in the connected component of $\mathcal{G}'$ to which $a$ belongs, and let $\mathcal{CC}^d_a \subset \mathcal{CC}_a$ represent the subset of agents within $\mathcal{CC}_a$ that are at most $d$ hops away from $a$.

Each agent $a$ aims to identify which nodes in its neighborhood $\mathcal{N}_a$ belong to its similarity class $\mathcal{C}_a$. To achieve this, agent $a$ receives at time $t$ an updated local mean estimate $\bar{x}^{t'}_{a,a'}$ from each neighbor $a' \in \mathcal{N}_a$. We denote with $\mathcal{C}^t_a \subseteq \mathcal{N}_a$ the set of neighbors that agent $a$ deems to belong to its own similarity class at time $t$, initially: $\mathcal{C}^0_a = \mathcal{N}_a$. Similarly to ColME, at each time $t$, agent $a$ first computes the distance $d^t_\gamma(a, a')$ for every $a' \in \mathcal{C}^{t-1}_a$ according to (1) and then updates $\mathcal{C}^t_a$ according to (2). As for ColME, $\mathcal{C}^t_a \subseteq \mathcal{C}^{t-1}_a$ and as soon as $a$ removes $a'$ from $\mathcal{C}^t_a$, it stops communicating with $a'$. Subsequently, communication occur over the pruned graph $\mathcal{G}^t = (\mathcal{A}, \mathcal{E}^t)$, where $\mathcal{E}^t = \{(a, a') \in \mathcal{E} : a' \in \mathcal{C}^t_a\}$.

The theoretical guarantees of our algorithms hold under more general settings than those in (Asadi et al. 2022). In particular, they apply to any set of distributions $\{D_a\}_{a \in \mathcal{A}}$ for which the following assumption is satisfied:

**Assumption 1.** *There exists a positive function $\beta_\gamma(\cdot) \in o(1)$ such that the true mean belongs to all intervals centered in $\bar{x}^t_{a,a}$ of width $\pm\beta_\gamma(t)$ for $t \in \mathbb{N}$ with confidence $1 - 2\gamma$, namely:*

$$\mathbb{P}\left(\forall t \in \mathbb{N}, \left|\bar{x}^t_{a,a} - \mu\right| < \beta_\gamma(t)\right) \geq 1 - 2\gamma, \forall a \in \mathcal{A}. \quad (4)$$

Assumption 1 is satisfied by sub-Gaussian distributions (SGD) with parameter $\sigma^2$, by selecting $\beta_\gamma(\cdot)$ as in (3). In Appendix B, we show that the assumption also holds for bounded fourth-moment distributions (BFMD) for $\beta_\gamma(\cdot)$ chosen as follows:

$$\beta_\gamma(n) = \left(2\frac{(\kappa+3)\sigma^4}{\gamma}\right)^{\frac{1}{4}} \left(\frac{1 + \ln^2 n}{n}\right)^{\frac{1}{4}}, \quad (5)$$

where $\sigma^2$ bounds the variance of the distributions $\{D_a, \forall a \in \mathcal{A}\}$ and $\kappa\sigma^4$ their fourth moment. When all the variables are identically distributed, $\kappa$ corresponds to the kurtosis. Moreover, for distributions with a larger number of bounded moments, tighter expressions can be derived for $\beta_\gamma(\cdot)$ (see Remark 1 in Appendix B). In what follows, we assume that Assumption 1 is always satisfied.

We aim first to determine the time needed for all agents in the connected component $\mathcal{CC}_a$ to identify the subset of neighbors residing in their similarity class, i.e., $\mathcal{C}_{a'}^t = \mathcal{C}_{a'} \cap \mathcal{N}_{a'}, \forall a' \in \mathcal{CC}_a$. Following a similar approach to Asadi et al. (2022, Theorem 1) we can prove that:

**Theorem 1.** [Proof in Appendix B] *Considering an arbitrarily chosen agent $a$ in $\mathcal{A}$, for any $\delta \in (0,1)$, employing either B-ColME or C-ColME we have:*

$$\mathbb{P}\left(\exists t > \zeta_a^D, \exists a' \in \mathcal{CC}_a : \mathcal{C}_{a'}^t \neq \mathcal{C}_{a'} \cap \mathcal{N}_{a'}\right) \leqslant \frac{\delta}{2}, \qquad (6)$$

*with* $\zeta_a^D = n_\gamma^\star\left(\frac{\Delta_a}{4}\right) + 1$, $\Delta_a = \min_{a' \in \mathcal{A} \setminus \mathcal{C}_a} \Delta_{a,a'}$, $\gamma = \frac{\delta}{4r|\mathcal{CC}_a|}$. $n_\gamma^\star(x)$ *denotes the minimum number of samples that are needed to ensure* $\beta_\gamma(n) < x$, *i.e.,* $n_\gamma^\star(x) = \lceil \beta_\gamma^{-1}(x) \rceil$.

This result demonstrates that the time required for all agents $a'$ in $\mathcal{CC}_a$ (the connected component to which $a$ belongs) to correctly identify their neighbors within the same similarity class $\mathcal{C}_a$ is bounded by $n_\gamma^\star(\frac{\Delta_a}{4})+1$. Here, $n_\gamma^\star(\frac{\Delta_a}{4})$ represents the number of samples needed to distinguish (with confidence $1-2\gamma$) the true mean of agent $a$ from that of an agent belonging to the 'closest' similarity class (i.e., the one with the closest true mean). The additional 1 accounts for the unit delay in communicating with the neighbors.

When comparing performance of B-ColME and C-ColME (Theorem 1) with ColME (Asadi et al. 2022) [Theorem 1] (reported in Appendix A as Theorem 7 for completeness), we observe that for large systems, if $r|\mathcal{CC}_a| \in \Theta(|\mathcal{A}|)$, $\zeta_a \approx |\mathcal{A}| + \zeta_a^D$, showing that, as expected, agents can identify much faster similar agents in their neighborhood than in the whole population $\mathcal{A}$.[1] See Sec. 5 for a detailed comparison of ColME, C-ColME, and B-ColME.

### 4.1 Consensus-based Algorithm: C-ColME

This section introduces the first collaborative mean estimation approach, inspired by consensus algorithms in dynamic settings, as in (Montijano et al. 2014; Franceschelli and Gasparri 2019). The basic idea is that each agent maintains two metrics: the empirical average of its local samples $\bar{x}_{a,a}^t$, and the 'consensus' estimate $\hat{\mu}_a^t$. The consensus variable is updated at time $t$ by computing a convex combination of the local empirical average $\bar{x}_{a,a}^t$ and a weighted sum of the consensus estimates in its (close) neighborhood $\{\hat{\mu}_{a'}^{t-1}, a' \in \mathcal{C}_a^{t-1} \cup \{a\}\}$, see Algorithm 1.

The dynamics of all estimates are captured by:

$$\hat{\boldsymbol{\mu}}^{t+1} = (1-\alpha_t)\,\bar{\mathbf{x}}^{t+1} + \alpha_t W_t \hat{\boldsymbol{\mu}}^t, \qquad (7)$$

where $(W_t)_{a,a'} = 0$ if $a' \notin \mathcal{C}_a^t$ and $\alpha_t \in (0,1)$ is the memory parameter. Once the agents cease pruning their neighbors, say at time $\tau$, the matrix $W_t$ does not need to change anymore, i.e., $W_t = W$ for any $t \geqslant \tau$ with $W_{a,a'} > 0$ if and only if $a' \in \mathcal{C}_a \cap \mathcal{N}_a$. In order to achieve consensus, the matrix $W$ needs to be doubly stochastic (Xiao and Boyd 2004) and we also require it to be symmetric. By time $\tau$, the original communication graph is split into $C$ connected components, where component $c$ includes $n_c$ agents. By an opportune permutation of the agents, we can write the matrix $W$ as follows

$$W = \begin{pmatrix} {}_1W & 0_{n_1 \times n_2} & \cdots & 0_{n_1 \times n_C} \\ 0_{n_2 \times n_1} & {}_2W & \cdots & 0_{n_2 \times n_C} \\ \cdots & \cdots & \cdots & \cdots \\ 0_{n_C \times n_1} & 0_{n_C \times n_2} & \cdots & {}_CW \end{pmatrix}, \qquad (8)$$

where each matrix ${}_cW$ is an $n_c \times n_c$ symmetric stochastic matrix. For $t \geqslant \tau$, the estimates in the different components evolve independently. We can then focus on a given component $c$. All agents in the same component share the same expected value, which we denote by $\mu(c)$. Moreover, let ${}_c\boldsymbol{\mu} = \mu(c)\mathbf{1}_c$. We denote by ${}_c\mathbf{x}^t$ and ${}_c\hat{\boldsymbol{\mu}}^t$ the $n_c$-dimensional vectors containing the samples' empirical averages and the consensus estimates for the agents in component $c$ and by $\lambda_{2,c}$ the second largest module of the eigenvalues of ${}_cW$.

Note that the actual evolution of $W_t$ is challenging to characterize due to topology changes during the graph pruning phase. However, our main results (Theorems 2 and 3) remain applicable to any system where the sequence of matrices $W_t$ for $t \leq \tau$ is arbitrarily set.

Algorithm 1: C-ColME over a Time Horizon $H$

**Input:** $\mathcal{G} = (\mathcal{A}, \mathcal{E})$, $(D_a)_{a \in \mathcal{A}}$, $\epsilon \in \mathbb{R}^+$, $\delta \in (0,1]$
**Output:** $\hat{\mu}_a$, $\forall a \in \mathcal{A}$ with $\mathbb{P}\left(|\hat{\mu}_a - \mu_a| < \epsilon\right) \geqslant 1-\delta$
$\mathcal{C}_a^0 \leftarrow \mathcal{N}_a, \forall a \in \mathcal{A}$
**for** time $t$ in $\{1, .., H\}$ **do**
  In parallel for all nodes $a \in \mathcal{A}$
  Draw $x_a^t \sim D_a$
  $\bar{x}_a^t \leftarrow \frac{t-1}{t}\bar{x}_a^{t-1} + \frac{1}{t}x_a^t$
  Compute $\beta_\gamma(t)$ with Eq. (3) or Eq. (5)
  **for** neighbor $a'$ in $\mathcal{N}_a \cap \mathcal{C}_a^{t-1}$ **do**
    $d_\gamma^t(a,a') \leftarrow \left|\bar{x}_a^t - \bar{x}_{a'}^{t-1}\right| - \beta_\gamma(t) - \beta_\gamma(t-1)$
  **end for**
  $\mathcal{C}_a^t \leftarrow \left\{a' \in \mathcal{N}_a \cap \mathcal{C}_a^{t-1} \text{ s.t. } d_\gamma^t(a,a') \leqslant 0\right\}$
  $\hat{\mu}_a^t \leftarrow (1-\alpha_t)\bar{x}_a^t + \alpha_t \sum_{a' \in \mathcal{C}_a \cup \{a\}} (W_t)_{a,a'} \hat{\mu}_{a'}^{t-1}$
**end for**

**Theorem 2.** [Proof in Appendix F] *Consider a system which evolves according to* (7) *with $W_t = W$ in* (8), *for $t \geqslant \tau$. Let ${}_cP = 1/n_c \mathbf{1}_c \mathbf{1}_c^\intercal$. For $\alpha_t = \frac{t}{t+1}$, it holds:*

$$\mathbb{E}\left[\|{}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu}\|^4\right] \in \mathcal{O}\left(\sup_{W_1, \cdots, W_{\zeta_D}} \frac{\mathbb{E}\left[\|{}_c\hat{\boldsymbol{\mu}}^{\zeta_D} - {}_c\boldsymbol{\mu}\|^4\right]}{(t+1)^4}\right)$$
$$+ \mathcal{O}\left(\frac{(1-1/\ln\lambda_{2,c})^2}{(1-\lambda_{2,c})^2} \frac{\mathbb{E}\left[\|{}_c\mathbf{x} - {}_cP\,{}_c\mathbf{x}\|^4\right]}{(t+1)^4}\right)$$
$$+ \mathcal{O}\left(\mathbb{E}\left[\|{}_cP\,{}_c\mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]\left(\frac{1+\ln t}{1+t}\right)^2\right).$$

The theorem shows that the error, quantified through the fourth moment, can be decomposed into three terms decreasing over time. The first term depends on the estimates' error at time $\tau$. The second term captures the effect of the consensus averaging, i.e., how effective is the algorithm in bringing

[1] For a fairer comparison, we should let ColME query $r$ other agents at each time $t$, where $r$ is the average degree of $\mathcal{G}$. In this case, $\zeta_a \approx |\mathcal{A}|/r + \zeta_a^D$ and the conclusion does not change.

the local estimates ${}_c\mathbf{x}$ close to their empirical value ${}_cP\,{}_c\mathbf{x}$ (for example it is minimized if $\lambda_2 = 0$, which corresponds to ${}_cW = {}_cP$, the ideal choice for the matrix ${}_cW$). Finally, the third term represents the minimum possible error, which would be obtained by averaging the estimates of all agents in the component using the matrix ${}_cP$.

Theorem 3 shows that C-ColME achieves a speedup proportional to the size of the connected component $|\mathcal{CC}_a|$.

**Theorem 3.** [Proof in Appendix F] *Consider a graph component $c$ and pick uniformly at random an agent $a$ in $c$. Let $g(x) \coloneqq x\ln^2(ex)$ and $\alpha_t = \frac{t}{t+1}$. Under BFMD, it holds:*

$$\mathbb{P}\left(\forall t > \tau_a^C,\ |\hat{\mu}_a^t - \mu_a| < \epsilon\right) \geqslant 1-\delta$$

*where* $\tau_a^C = \max\left\{\zeta_a^D, g\left(C\frac{\mathbb{E}\left[\|{}_cP\,{}_c\mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{|\mathcal{CC}_a|\epsilon^4\delta}\right) \in \tilde{\mathcal{O}}\left(\frac{\tilde{n}_{\frac{\delta}{2}}(\varepsilon)}{|\mathcal{CC}_a|}\right)\right\}$ *and* $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon) = \left\lceil \frac{2(\kappa+3)\sigma^4}{\delta\varepsilon^4}\right\rceil$.

The theorem shows that the time to reach an $\epsilon$-accurate estimate with high probability is the maximum of the time for the agents in $\mathcal{CC}_a$ to identify their neighbors in the same similarity class and the time required for those agents to obtain and $\epsilon$-accurate estimate if they could share their own samples. Indeed, we observe that $n^\star_{\delta/2}(\epsilon)$ is the number of samples sufficient to ensure that $\mathbb{P}(|\hat{\mu}_a - \mu_a| > \epsilon) < \delta/2$ (see details in Appendix E) and that the nodes in $\mathcal{CC}_a$ collectively gather this number of samples by time $t = \left\lceil \frac{n^\star_{\delta/2}(\varepsilon)}{|\mathcal{CC}_a|}\right\rceil$.

Appendix F also presents convergence results for the case $\alpha_t = \alpha$, but they do not enjoy the same speedup factor.

### 4.2 Message-passing Algorithm: B-ColME

In B-ColME, each node $a \in \mathcal{A}$ continuously exchanges *messages* with its direct neighbors $a' \in \mathcal{C}_a^t$. This enables node $a$ to acquire not only the neighbor's local estimates $\{\bar{x}^t_{a,a'}, a' \in \mathcal{C}_a^t\}$, but also aggregated estimates from nodes up to a distance $d$ in the graph $\mathcal{G}^t$ (where $d$ is a tunable parameter). Indeed, each neighbor $a'$ acts as a *forwarder*, granting node $a$ access to the records from its own neighbors $a'' \in \mathcal{C}^t_{a'} \setminus \{a\}$. Provided each agent correctly identifies all similar nodes in its neighborhood, agent $a$ can potentially access the (delayed) local estimates of all agents in $\mathcal{CC}_a^d$.

In our message-passing scheme, at time $t$, agent $a \in \mathcal{A}$ receives a message $M^{t,a'\to a}$ from all neighbors $a' \in \mathcal{C}_a^t$. The message $M^{t,a'\to a}$ is a $d \times 2$ table whose elements $m_{h,1}^{t,a'\to a}$ contain a sum of samples, while $m_{h,2}^{t,a'\to a}$ indicates the number of samples contributing to this sum. In particular, at each time $t$, the first row of the table is set as: $m_{1,1}^{t,a'\to a} = \sum_{\tau=1}^t x_{a'}^\tau$ and $m_{1,2}^{t,a'\to a} = t$, i.e., the immediate neighbors' sum of local samples. The remaining entries are computed through the following recursion:

$$m_{h,i}^{t,a'\to a} = \sum_{a''\in\mathcal{C}^t_{a'},\ a''\neq a} m_{h-1,i}^{t-1,a''\to a'},$$

for $h \in \{2,\dots,d\}$ and $i \in \{1,2\}$. This captures information extending beyond immediate neighbors. For additional details on B-ColME see Algorithm 2 and Fig. 5 in Appendix D.

If $\mathcal{G}^t \cap \mathcal{N}_a^d$ is a tree, then $m_{h,1}^{t,a'\to a}$ contains the sum of all samples generated within time $t-h+1$ by agents $a'' \in \mathcal{G}^t$ at distance $h-1$ from $a'$ and distance $h$ from $a$, while $m_{h,2}^{t,a'\to a}$ contains the corresponding number of samples (the proof is by induction on $h$). Agent $a$ can estimate its mean as:

$$\hat{\mu}_a^t = \frac{\sum_{\tau=1}^t x_a^\tau + \sum_{a'\in\mathcal{C}_a^t}\sum_{h=1}^d m_{h,1}^{t,a'\to a}}{t + \sum_{a'\in\mathcal{C}_a^t}\sum_{h=1}^d m_{h,2}^{t,a'\to a}}. \tag{9}$$

Under the local tree structure assumption, this corresponds to performing an empirical average over all the samples generated by all agents in $\mathcal{G}^t$ at distance $0 \leqslant h \leqslant d$ from $a$ up to time $t-h$. If $\mathcal{G}^t \cap \mathcal{N}_a^d$ is not a tree, samples collected by a given agent $a''$ may be included in messages received by $a$ through different parallel paths (from $a$ to $a''$). As a result, these samples are erroneously counted multiple times in (9). The parameter $d$ must be chosen to prevent this issue with high probability, as discussed in Sec. 4.3.

Algorithm 2: B-ColME over a Time Horizon $H$

**Input:** $\mathcal{G} = (\mathcal{A}, \mathcal{E})$, $(D_a)_{a\in\mathcal{A}}$, $\varepsilon \in \mathbb{R}^+$, $\delta \in (0,1]$
**Output:** $\hat{\mu}_a$, $\forall a \in \mathcal{A}$ with $\mathbb{P}\left(|\hat{\mu}_a - \mu_a| < \varepsilon\right) \geqslant 1-\delta$
$\mathcal{C}_a^0 \leftarrow \mathcal{N}_a, \forall a \in \mathcal{A}$
**for** time $t$ in $\{1,..,H\}$ **do**
  In parallel for all nodes $a \in \mathcal{A}$
  Draw $x_a^t \sim D_a$
  $\bar{x}_a^t \leftarrow \frac{t-1}{t}\bar{x}_a^{t-1} + \frac{1}{t}x_a^t$
  Compute $\beta_\gamma(t)$ with Eq. (3) or Eq. (5)
  **for** neighbor $a'$ in $\mathcal{C}_a^{t-1}$ **do**
    $d_\gamma^t(a,a') \leftarrow \left|\bar{x}_a^t - \bar{x}_{a'}^{t-1}\right| - \beta_\gamma(t) - \beta_\gamma(t-1)$
    **if** $d_\gamma^t(a,a') > 0$ **then**
      $\mathcal{C}_a^t \leftarrow \mathcal{C}_a^t \setminus \{a'\}$
    **end if**
  **end for**
  **for** neighbor $a'$ in $\mathcal{C}_a^t$ **do**
    Compute $M^{t,a\to a'}$ and send it to $a'$
  **end for**
  Wait for messages $M^{t,a'\to a}$ $\forall a' \in \mathcal{C}_a^t$
  $\hat{\mu}_a^t \leftarrow \frac{\sum_{\tau=1}^t x_a^\tau + \sum_{a'\in\mathcal{C}_a^t}\sum_{h=1}^d m_{h,1}^{t-1,a'\to a}}{t + \sum_{a'\in\mathcal{C}_a^t}\sum_{h=1}^d m_{h,2}^{t,a'\to a}}$
**end for**

Theorem 4 presents the $(\epsilon,\delta)$ convergence result for B-ColME, which enjoys a speedup proportional to $|\mathcal{CC}_a^d|$.

**Theorem 4.** [Proof in Appendix E] *Provided that $\mathcal{CC}_a^d$ is a tree, for any $\delta \in (0,1)$, employing B-ColME, we have:*

$$\mathbb{P}\left(\forall t > \tau_a^B,\ |\hat{\mu}_a^t - \mu_a| < \varepsilon\right) \geqslant 1-\delta$$

*where* $\tau_a^B = \max\left[\zeta_a^D + d, \frac{\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)}{|\mathcal{CC}_a^d|} + d\right]$ *and* $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon) = \left\lceil -\frac{2\sigma^2}{\varepsilon^2}\ln\left(\frac{\delta}{4}(1-\mathrm{e}^{-\frac{\varepsilon^2}{\sigma^2}})\right)\right\rceil$ *for SGD and as in Theorem 3 for BFMD.*

Similar considerations to those for Theorem 3 apply. The additional term $d$ accounts for the delay introduced by the message-passing scheme.

**Corollary 5.** *Let $\mathbb{P}(\mathcal{CC}_a^d$ is not a tree$) = \delta'$, then for any $\delta \in (0,1)$, employing B-ColME, we have:*

$$\mathbb{P}\left(\forall t > \tau_a^B \,:\, |\hat{\mu}_a^t - \mu_a| < \varepsilon\right) \geqslant 1 - \delta - \delta'.$$

Theorem 23 (Appendix G) provides upper bounds for $\delta'$ when $\mathcal{G}$ is a random regular graph. In particular, as long we set $d$ as in Proposition 6 (Sec. 4.3), $\delta'$ converges to 0 as the number of agents $|\mathcal{A}|$ increases.

### 4.3 Choice of the Graph and other Parameters

The selection of the graph $\mathcal{G}(\mathcal{A}, \mathcal{E})$ is crucial for the effectiveness of our algorithms. Here we state the key desirable properties of $\mathcal{G}(\mathcal{A}, \mathcal{E})$. First, Theorems 3 and 4 show that learning timescales, $\tau_a^B$ and $\tau_a^C$, decrease as the size of the collaborating agent groups, $\mathcal{CC}_a^d$ and $\mathcal{CC}_a$, increase. Therefore, a highly connected graph is preferred to promote the formation of large clusters of agents belonging to the same similarity class after the disconnection of inter-class edges. Second, the spatial and temporal complexities of B-ColME and C-ColME are directly proportional to the agents' degree within the graph. Hence, we want the degree to be small and possibly uniform across the agents to balance computation across agents. Note that the first two criteria partially conflict, as a higher degree generally leads to larger groups $\mathcal{CC}_a^d$ and $\mathcal{CC}_a$, while a smaller degree ensures better spatial and temporal complexities. A third criterion, specific to B-ColME, is that each agent's neighborhood should have a tree-like structure extending up to $d$ hops, with $d$ as large as possible.

Considering these criteria, we opt for the class of *simple* random regular graphs $\mathcal{G}_0(N, r)$. These graphs are sampled uniformly at random from the set of all $r$-regular simple graphs with $N$ nodes, i.e., graphs without parallel edges or self-loops, and in which every node has exactly $r$ neighbors. Note that an even product $rN$ guarantees the set is not empty. The class $\mathcal{G}_0(N, r)$ exhibits strong connectivity properties for small values of $r$. Specifically, for any $r \geqslant 3$, the probability that the sampled graph is connected approaches one as $N$ increases. Moreover, the sampled graph demonstrates a local tree-like structure with high probability (proof in Appendix G). The choice of $r$ (agents' degree) illustrates the trade-off discussed above between reducing complexity (low $r$) and having large connected components (high $r$). A sensible rule is to select $r$ sufficiently large to guarantee that most agents in the smallest (most critical) class belong to the same connected component. Consider a class including a fraction $p_{k_a}$ of agents, Table 3 in Appendix G shows the average fraction of agents in this class that is not connected to the main connected component as a function of $r$. To keep this fraction below e.g. $10^{-2}$ a good rule of thumb is $r = 4/p_{k_a}$.

A final key parameter for B-ColME is the maximum distance $d$ over which local estimates from agents are propagated. This parameter must be carefully calibrated: it should be small enough to ensure that $\mathcal{CC}_a^d$, for a randomly chosen $a \in \mathcal{A}$, has a tree-like structure with high probability. However, choosing a $d$ that is too small could unnecessarily restrict the size of $\mathcal{CC}_a^d$, thereby undermining the effectiveness of the estimation process (Theorem 4). A comprehensive analysis of how the parameters $r$ and $d$ influence both the structure of $\mathcal{N}_a^d$ and the size of $\mathcal{CC}_a^d$ can be found in Appendix G. Here, we informally summarize the main result:

**Proposition 6.** *By selecting $d = \left\lfloor \frac{1}{2} \log_{r-1} \frac{|\mathcal{A}|}{\log_{r-1} ||\mathcal{A}|} \right\rfloor$ the number of nodes $a \in \mathcal{A}$, whose $d$-neighborhood is not a tree, is $o(|\mathcal{A}|)$ with a probability tending to 1 as $|\mathcal{A}|$ increases. For the same $d$ and $r \in \Theta(\log(1/\delta))$, $|\mathcal{CC}_a^d|$ is in $\Omega(|\mathcal{A}|^{\frac{1}{2}-\phi})$ for any arbitrarily small $\phi > 0$ with probability arbitrarily close to 1.*

Finally, for C-ColME, the consensus matrix $W$ could be chosen to minimize the second largest module $\lambda_{2,c}$ of the eigenvalues of each block ${}_cW$ in order to minimize the bound in Theorem 2. This optimization problem has been studied by Xiao and Boyd (2004) and requires in general a centralized solution. In what follows, we consider the following simple, decentralized configuration rule: $(W_t)_{a,b} = \frac{1}{\max\{|\mathcal{C}_a^t|, |\mathcal{C}_b^t|\}+1}, \forall b \in \mathcal{C}_a^t$ and $(W_t)_{a,a} = 1 - \sum_{b \in \mathcal{C}_a^t} (W_t)_{a,b}$, making $W$ symmetric and doubly stochastic.

## 5 Algorithms' Comparison

Table 2 presents a comparative analysis of the three algorithms: ColME, C-ColME, and B-ColME. For a fair comparison, we consider a variant of ColME, where each agent can communicate with $r$ agents at each time $t$, so that all three algorithms incur the same communication overhead.

The second column of Table 2 outlines the space and time complexities of the algorithms. Notably, even when $r$ and $d$ are allowed to increase logarithmically with the number of agents $|\mathcal{A}|$, B-ColME retains its efficiency advantage over ColME. C-ColME demonstrates even greater improvements, further reducing the per-agent burden compared to the savings achieved by B-ColME.

The third and fourth columns detail the characteristic times required to achieve $(\epsilon, \delta)$ convergence for the estimates generated by the three algorithms, considering both sub-Gaussian local data distributions and distributions with bounded fourth moment. The characteristic times correspond to $\tau_a$, $\tau_a^C$, and $\tau_a^B$ in Theorem 7 in Appendix A, Theorem 3, and 4, respectively. The table reports their asymptotic behavior as the number of agents $|\mathcal{A}|$ increases ignoring logarithmic factors. The detailed derivations of these results are provided in Appendix H.

Three factors contribute to the characteristic times. The first factor is the time required to correctly identify potential collaborators. For ColME, this involves each agent classifying the other $|\mathcal{A}| - 1$ agents, leading to a term that scales as $\log|\mathcal{A}|$ or $|\mathcal{A}|$, depending on the assumed properties of the local distribution. For B-ColME and C-ColME, $|\mathcal{A}|$ is replaced by $|\mathcal{CC}_a|r$, which represents an upper bound on the number of connections the agents in $\mathcal{CC}_a$ may have initially established with agents from different classes. This substitution may not be immediate, as one might initially expect the relevant scale to be simply $r$. However, this adjustment accounts for the potential ripple effect of classification er-

| | Per-agent space/time complexity | Convergence time | |
|---|---|---|---|
| | | sub-Gaussian | bounded 4-th moment |
| ColME ($r$ communications) | $\lvert\mathcal{A}\rvert$ | $\frac{1}{\Delta_a^2}\log\frac{\lvert\mathcal{A}\rvert}{\Delta_a\delta}+\frac{\lvert\mathcal{A}\rvert}{r}+\frac{1}{\lvert\mathcal{C}_a\rvert}\frac{1}{\varepsilon^2}\log\frac{1}{\delta\varepsilon^2}$ | $\frac{1}{\Delta_a^4}\frac{\lvert\mathcal{A}\rvert}{\delta}+\frac{\lvert\mathcal{A}\rvert}{r}+\frac{1}{\lvert\mathcal{C}_a\rvert}\frac{1}{\delta\varepsilon^4}$ |
| C-ColME [Thm. 3] ($\alpha_t=\frac{t}{t+1}$) | $r$ | $-$ | $\frac{1}{\Delta_a^4}\frac{\lvert\mathcal{CC}_a\rvert r}{\delta}+\frac{1}{\lvert\mathcal{CC}_a\rvert}\frac{1}{\delta\varepsilon^4}$ |
| B-ColME [Thm. 4] | $rd$ | $\frac{1}{\Delta_a^2}\log\frac{\lvert\mathcal{CC}_a\rvert r}{\Delta_a\delta}+d+\frac{1}{\lvert\mathcal{CC}_a^d\rvert}\frac{1}{\varepsilon^2}\log\frac{1}{\delta\varepsilon^2}$ | $\frac{1}{\Delta_a^4}\frac{\lvert\mathcal{CC}_a\rvert r}{\delta}+d+\frac{1}{\lvert\mathcal{CC}_a^d\rvert}\frac{1}{\delta\varepsilon^4}$ |

Table 2: Comparison of collaborative estimation algorithms. The convergence time is provided in order sense.

rors: a mistake by any agent $a$ can impact the estimates of all agents within the same connected component $\mathcal{CC}_a$.

The second factor contributing to the characteristic times is the time each agent needs to collect all relevant information. For ColME, this time is proportional to $|\mathcal{A}|/r$, as an agent queries all other agents. For B-ColME, the time is specifically tied to $d$, the maximum number of hops messages propagate. Notably, this term does not appear for C-ColME, as it is dominated by the final term.

The final term represents the time needed for accurate mean estimation after collaborators have been identified, highlighting the benefits of collaboration. In ColME, the collaboration's benefit is particularly striking, as all agents within the same class work together to improve their estimates. This collective effort effectively reduces the convergence time by a factor proportional to the size of the collaborating group, $|\mathcal{C}_a|$. For B-ColME and C-ColME, although the speed-up remains proportional to the number of collaborating agents, the actual numbers of collaborators, $|\mathcal{CC}_a^d|$ for B-ColME and $|\mathcal{CC}_a|$ for C-ColME, are in general smaller.

In conclusion, while ColME potentially offers the most accurate estimates, it requires longer convergence times and greater memory and computational resources. In contrast, B-ColME and C-ColME present more efficient alternatives, achieving faster convergence with reduced per-agent resource demands. However, this efficiency may come at the expense of the maximum attainable accuracy. The next section quantifies this trade-off experimentally.

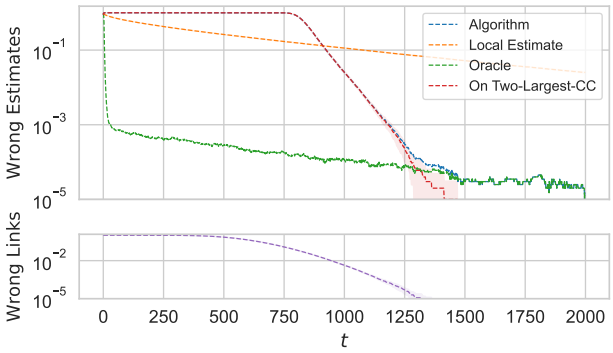
## 6 Numerical Experiments

We evaluate the proposed algorithms on the class $\mathcal{G}_0(N,r)$ of simple regular graphs (see Sec. 4.3). In this setting, each agent connects to $r$ other agents chosen uniformly at random in $\mathcal{A}$. This setup also provides the tree-like local structure required for B-ColME. Agents belong to one of two classes, associated with Gaussian distributions $D_1 \sim \mathcal{N}(\mu_1 = 0, \sigma^2 = 4)$ and $D_2 \sim \mathcal{N}(\mu_2 = 1, \sigma^2 = 4)$. Each node is assigned to one of the two classes with equal probability. Unless otherwise stated, in the experiments $|\mathcal{A}| = N = 10000$, $r = 10$, $d = 4$, $\varepsilon = 0.1$, $\delta = 0.1$, and $\beta_\gamma(n)$ as in (3). In Appendix C and I, we provide additional experiments for the multidimensional case and varying the system's parameters.

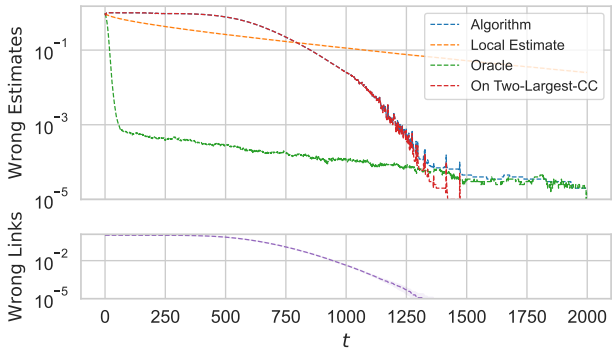
Figure 1 showcases the performance of B-ColME and C-ColME using two key metrics: the fraction of agents with *incorrect estimates* ($\hat{\mu}_a^t$ more than $\varepsilon$ away from the true mean $\mu_a$), and the fraction of *wrong links* still in use (a wrong link connects agents from different classes). We compare our algorithms against two benchmarks. The first benchmark has each agent independently relying on its *local* estimate $\bar{x}_{a,a}^t$. In the second benchmark, an *oracle* provides each agent with precise knowledge of which neighbors belong to the same similarity class (i.e., $\mathcal{C}_a^t = \mathcal{C}_a \cap \mathcal{N}_a, \forall a, t$). The figure reveals that B-ColME has a longer transient phase but then exhibits a slightly steeper convergence than C-ColME. Notably, B-ColME's estimates show no apparent improvement until about 90% of the wrong links have been removed, whereas C-ColME's estimates begin to improve as soon as the first edges are eliminated. This phenomenon can be explained as follows. In B-ColME, the estimates at agent $a$ are not influenced by the removal of some wrong links as long as its $d$-hop neighborhood $\mathcal{N}_a^d$ remains unchanged. For instance, a given node $a' \notin \mathcal{C}_a$ is removed from $\mathcal{N}_a^d$ only when *all* paths of length at most $d$ between agent $a$ and agent $a'$ are eliminated. In contrast, in C-ColME, agent $a'$ contributes to the weighted estimate at agent $a$ with a weight equal to the sum over all paths between $a$ and $a'$ of the product of the consensus weights along the path. As paths are progressively removed, the negative impact of $a'$ on agent $a$'s estimate is gradually reduced. However, once all wrong links are removed, B-ColME benefits from its estimates being computed solely on agents belonging to the same class, while C-ColME requires some additional time for the effect of past estimates to fade away.

We also compare the proposed algorithms with ColME and a simplified version (s-ColME) where the optimistic distance $d_\gamma^t(a,a')$ is recomputed only for the $r$ agents queried at time $t$, achieving an $\mathcal{O}(r)$ per-agent computational cost (the memory cost remains $\mathcal{O}(|\mathcal{A}|)$). As predicted by the theoretical analysis, B-ColME and C-ColME are faster than ColME, but at the cost of a higher asymptotic error because agent $a$ collaborates only with the smaller group of nodes in $\mathcal{CC}_a^d$ for B-ColME, and $\mathcal{CC}_a$ for C-ColME. ColME pays for this asymptotic improvement with a $\mathcal{O}(|\mathcal{A}|)$ space-time complexity per agent, impractical for large-scale systems. Note that s-ColME improves ColME's complexity at the cost of a much slower discovery of same-class neighbors.

While we focused on online mean estimation, our approach can be adapted to decentralized federated learning. To illustrate this possibility, we adapt the consensus-based decentralized federated learning algorithm, by letting agents progressively exclude neighbors they identify as belonging to a different class. The cosine dissimilarity of agents' updates, the same metric used in Clus-




(a) B-ColME




(b) C-ColME

Figure 1: Fraction of agents with estimate deviates by more than $\epsilon$ from the true value, i.e., $|\{a \in \mathcal{A} : |\hat{\mu}_a^t - \mu_a| > \varepsilon\}|/|\mathcal{A}|$ (top) and fraction of *wrong links* (bottom) for B-ColME (a) and C-ColME (b), over 20 realizations with 95% confidence intervals.

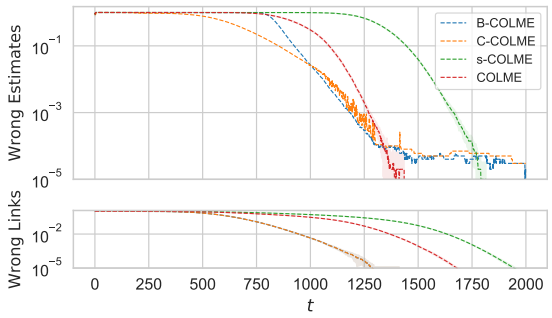



Figure 2: Comparison of our algorithms and two versions of ColME, over 10 realizations.

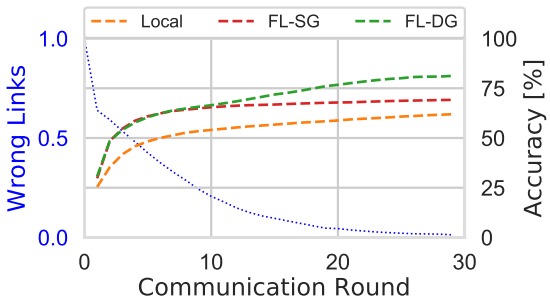



Figure 3: Accuracy of a local model (Local), a decentralized FL over a static graph (FL-SG), and our approach over a dynamic graph (FL-DG). We also show the fraction of links between classes (*wrong links*) over time for FL-DG.

teredFL (Ghosh et al. 2020; Chen et al. 2021), replaces the optimistic distance $d_\gamma^t(a, a')$ (details in Appendix J). Figure 3 shows the performance of our *decentralized FL over a dynamic graph* (FL-DG) with $|\mathcal{A}| = 100$ agents initially organized over a complete graph. Two different distributions are obtained from MNIST (Deng 2012) by swapping/maintaining some labels and each client progressively receives new data samples from one of the two distributions. As the graph is progressively split in two clusters of clients belonging to the same class, each agent's model benefits from cooperating only with similar clients and it achieves a higher accuracy.

## 7 Conclusions

In this paper, we introduced B-ColME and C-ColME, two scalable and fully distributed algorithms for collaborative local mean estimation. We thoroughly evaluated their performance through both theoretical and empirical analyses. Additionally, we adapted our approach for personalized federated learning, applying it to the task of handwritten digit recognition using the MNIST dataset.

This work points to several future research directions. Here we have allowed agents only to sever existing connections, but not to establish new ones. Investigating scenarios where agents can rewire their connections to communicate with new agents outside their original neighborhood would be an interesting extension. Additionally, we assumed that agents are partitioned into similarity classes, with agents in the same class generating data with identical true mean. Extending our approach to accommodate more general scenarios, where each agent generates data with potentially different true mean, would be a valuable avenue for further exploration.

## References

Adi, E.; Anwar, A.; Baig, Z.; and Zeadally, S. 2020. Machine learning and data analytics for the IoT. *Neural Computing and Applications*, 32(20): 16205–16233.

Amini, H. 2010. Bootstrap Percolation and Diffusion in Random Graphs with Given Vertex Degrees. *The Electronic Journal of Combinatorics*, 17(1).

Asadi, M.; Bellet, A.; Maillard, O.-A.; and Tommasi, M. 2022. Collaborative Algorithms for Online Personalized Mean Estimation. *Transactions on Machine Learning Research*.

Beaussart, M.; Grimberg, F.; Hartley, M.-A.; and Jaggi, M. 2021. WAFFLE: Weighted Averaging for Personalized Federated Learning.

Chayti, E. M.; Karimireddy, S. P.; Stich, S. U.; Flammarion,

N.; and Jaggi, M. 2022. Linear Speedup in Personalized Collaborative Learning.

Chen, M.; Yang, Z.; Saad, W.; Yin, C.; Poor, H. V.; and Cui, S. 2021. A Joint Learning and Communications Framework for Federated Learning Over Wireless Networks. *IEEE Transactions on Wireless Communications*, 20(1): 269–283.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Ding, S.; and Wang, W. 2022. Collaborative Learning by Detecting Collaboration Partners. *Advances in Neural Information Processing Systems*, 35: 15629–15641.

Donahue, K.; and Kleinberg, J. 2021. Model-Sharing Games: Analyzing Federated Learning Under Voluntary Participation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Dorner, F. E.; Konstantinov, N.; Pashaliev, G.; and Vechev, M. 2024. Incentivizing honesty among competitors in collaborative learning and optimization. *Advances in Neural Information Processing Systems*, 36: 7659–7696.

Even, M.; Massoulié, L.; and Scaman, K. 2022. On Sample Optimality in Personalized Collaborative and Federated Learning. In *Advances in Neural Information Processing Systems*.

Fallah, A.; Mokhtari, A.; and Ozdaglar, A. 2020. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. In *Advances in Neural Information Processing Systems*.

Franceschelli, M.; and Gasparri, A. 2019. Multi-stage discrete time and randomized dynamic average consensus. *Automatica*, 99: 69–81.

Galante, F.; Neglia, G.; and Leonardi, E. 2024. Scalable Decentralized Algorithms for Online Personalized Mean Estimation. arXiv:2402.12812.

Ghosh, A.; Chung, J.; Yin, D.; and Ramchandran, K. 2020. An Efficient Framework for Clustered Federated Learning. In *Advances in Neural Information Processing Systems*.

Grimberg, F.; Hartley, M.-A.; Karimireddy, S. P.; and Jaggi, M. 2021. Optimal Model Averaging: Towards Personalized Collaborative Learning.

Janson, S. 2009. The Probability That a Random Multigraph is Simple. *Combinatorics, Probability and Computing*, 18(1-2): 205–225.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Nitin Bhagoji, A.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; D'Oliveira, R. G. L.; Eichner, H.; El Rouayheb, S.; Evans, D.; Gardner, J.; Garrett, Z.; Gascón, A.; Ghazi, B.; Gibbons, P. B.; Gruteser, M.; Harchaoui, Z.; He, C.; He, L.; Huo, Z.; Hutchinson, B.; Hsu, J.; Jaggi, M.; Javidi, T.; Joshi, G.; Khodak, M.; Konecný, J.; Korolova, A.; Koushanfar, F.; Koyejo, S.; Lepoint, T.; Liu, Y.; Mittal, P.; Mohri, M.; Nock, R.; Özgür, A.; Pagh, R.; Qi, H.; Ramage, D.; Raskar, R.; Raykova, M.; Song, D.; Song, W.; Stich, S. U.; Sun, Z.; Suresh, A. T.; Tramèr, F.; Vepakomma, P.; Wang, J.; Xiong, L.; Xu, Z.; Yang, Q.; Yu, F. X.; Yu, H.; and Zhao, S. 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and Robust Federated Learning Through Personalization. In *Proceedings of the 38th International Conference on Machine Learning*.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.

Maillard, O.-A. 2019. *Mathematics of Statistical Sequential Decision Making*. Habilitation à diriger des recherches, Université de Lille Nord de France.

Mansour, Y.; Mohri, M.; Ro, J.; and Suresh, A. T. 2020. Three Approaches for Personalization with Applications to Federated Learning.

Marfoq, O.; Neglia, G.; Bellet, A.; Kameni, L.; and Vidal, R. 2021. Federated Multi-Task Learning under a Mixture of Distributions. In *Advances in Neural Information Processing Systems*.

McKay, B. D.; and Wormald, N. C. 1990. Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, 11(1): 52–67.

Meyer, C. D. 2023. *Matrix Analysis and Applied Linear Algebra*. SIAM.

Montijano, E.; Montijano, J. I.; Sagüés, C.; and Martínez, S. 2014. Robust discrete time dynamic average consensus. *Automatica*, 50(12): 3131–3138.

of Mathematical Functions, D. L. 2023. Asymptotic Expansions: Exponential and Logarithmic Integral. [Online; accessed 23-January-2024].

Rossi, W. S.; Como, G.; and Fagnani, F. 2019. Threshold Models of Cascades in Large-Scale Networks. *IEEE Transactions on Network Science and Engineering*, 6(2): 158–172.

Sattler, F.; Müller, K.-R.; and Samek, W. 2021. Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8): 3710–3722.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press.

Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2023. Towards Personalized Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12): 9587–9603.

Tsoy, N.; Mihalkova, A.; Todorova, T. N.; and Konstantinov, N. 2024. Provable Mutual Benefits from Federated Learning in Privacy-Sensitive Domains. In *International Conference on Artificial Intelligence and Statistics*.

Vershynin, R. 2018. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.

Xiao, L.; and Boyd, S. 2004. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1): 65–78.

# A Appendix A - ColME's Guarantees

We provide the two main theoretical results from Asadi et al. (2022) [Theorem 1 and Theorem 2] which were proven under the assumption that the distributions $\{D_a, a \in \mathcal{A}\}$ are sub-Gaussian. Recall that a distribution $D$ is sub-Gaussian with parameter $\sigma^2$ if $\forall \lambda \in \mathbb{R}$, $\log \mathbb{E}_{x\sim D} \exp(\lambda(x-\mu)) \leqslant \frac{1}{2}\lambda^2\sigma^2$. The proofs of the Theorems presented in this Appendix can be found in (Asadi et al. 2022).

Select the amplitude of the confidence intervals $I_{a,a}$ and $I_{a,a'}$, centered around the sample means, as:

$$\beta_\gamma(n) := \sigma\sqrt{\frac{2}{n}\left(1+\frac{1}{n}\right)\ln(\sqrt{(n+1)}/\gamma)} \tag{10}$$

and let $n^\star_\gamma(x)$ denote the minimum number of samples that are needed to ensure $\beta_\gamma(n) < x$, i.e., $n^\star_\gamma(x) = \lceil \beta^{-1}_\gamma(x) \rceil$. We also use $n^\star_\gamma(a,a')$ for $n^\star_\gamma(\Delta_{a,a'}/4)$ and $n^\star_\gamma(a)$ for $n^\star_\gamma(\Delta_a/4)$ where $\Delta_a := \min_{a'\in\mathcal{A}\setminus\mathcal{C}_a} \Delta_{a,a'}$; these values denote the minimum number of samples to distinguish (with confidence $1-2\gamma$) the true mean of agent $a$ from the true mean of agent $a'$, and from the true mean of any other agent, respectively.

The first result (Theorem 7) provides a bound on the time needed to ensure that a randomly chosen agent $a \in \mathcal{A}$ correctly identifies its similarity class, i.e, $\mathcal{C}^t_a = \mathcal{C}_a$, with high probability.[2]

**Theorem 7.** *(Asadi et al. 2022, Theorem 1) Assume distributions $\{D_a, \forall a \in \mathcal{A}\}$ are sub-Gaussians with parameter $\sigma^2$. For any $\delta \in (0,1)$, and with $\gamma = \frac{\delta}{4|\mathcal{A}|}$, employing ColME, we have:*

$$\mathbb{P}\left(\exists t > \zeta_a : \mathcal{C}^t_a \neq \mathcal{C}_a\right) \leqslant \frac{\delta}{2}, \tag{11}$$

$$\textit{with } \zeta_a = n^\star_\gamma(a) + |\mathcal{A}| - 1 - \sum_{a'\in\mathcal{A}\setminus\mathcal{C}_a} \mathbb{1}_{\left\{n^\star_\gamma(a) > n^\star_\gamma(a,a') + |\mathcal{A}| - 1\right\}}.$$

Theorem 7 shows that the time $\zeta_a$ required by each agent $a$ to identify (with high probability) which other agents are in the same similarity class can be bounded by the sum of two terms. The first term, $n^\star_\gamma(a)$, is an upper-bound for the number of samples needed to conclude, with probability larger or equal than $1-2\gamma$, if the true means of two agents differ by the minimum gap $\Delta_a$. The additional term corresponds to the residual time required to acquire the estimates from other agents. It can be shown that $n^\star_\gamma(a)$ grows at most as $\log|\mathcal{A}|$, then $\zeta_a \in \mathcal{O}(|\mathcal{A}|)$ and for large systems ($|\mathcal{A}| \gg n^\star_\gamma(a)$), the need to query all agents at least once becomes the dominant factor.

Turning our attention to the estimation error, it holds:

**Theorem 8.** *(Asadi et al. 2022, Theorem 2) Assume distributions $\{D_a, \forall a \in \mathcal{A}\}$ are sub-Gaussians with parameter $\sigma^2$. For any $\delta \in (0,1)$, and with $\gamma = \frac{\delta}{4|\mathcal{A}|}$, employing ColME, we have:*

$$\mathbb{P}\left(\forall t > \tau_a, |\hat{\mu}^t_a - \mu_a| < \varepsilon\right) \geqslant 1-\delta, \tag{12}$$

*with* $\tau_a = \max\left[\zeta_a, \frac{n^\star_{\frac{\delta}{2}}(\varepsilon)}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a|-1}{2}\right]$.

Theorem 8 admits a straightforward explanation. Provided that agent $a$ has successfully estimated its similarity class at time $t$ (i.e., $\mathcal{C}^t_a = \mathcal{C}_a$), the error in the mean estimation will depend only on the available number of samples of agents in $\mathcal{C}_a$, used for the computation of $\hat{\mu}^t_a$. Now, a number of samples equal to $n^\star_{\frac{\delta}{2}}(\varepsilon)$ is sufficient to ensure that $\mathbb{P}(|\hat{\mu}_a - \mu_a| > \varepsilon) < \delta/2$. For agent $a$ such number of samples is surely available at time $t \geqslant t^*_a = \frac{n^\star_{\frac{\delta}{2}}(\varepsilon)}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a|-1}{2}$ where the second term is needed to take into account the effect of the delay introduced by the round-robin scheme. Applying the union bound, we can claim that whenever $t \geqslant \max(\zeta_a, t^*_a)$ w.p. $1-\delta$ both $\mathcal{C}^t_a = \mathcal{C}_a$ and $|\hat{\mu}^t_a - \mu_a| \leqslant \varepsilon$ hold.

[2] In this paper events that occur with a probability larger than $1-\delta$ are said to occur with high probability (w.h.p.).

# B Appendix B - Proof of Theorem 1

In this Appendix we provide the proof of Theorem 1 and, as a side result, we derive Equation (5). For the sake of clarity, we repeat the statement of the theorems.

The first theoretical result provides a bound on the probability that the nodes in the connected component $\mathcal{CC}_a$ of a certain node $a \in \mathcal{A}$ misidentify their *true* neighbors $\mathcal{C}_{a'} \cap \mathcal{N}_{a'}$, $\forall a' \in \mathcal{CC}_a$. We remark that, unlike ColME, the *goodness* of an estimate of our *scalable* algorithms depends not only on the ability of a given node to correctly identify its *true* neighborhood but also on the neighborhood estimates of all other nodes. Communication between nodes (i.e., message passing or consensus mechanism) makes error propagation possible within a connected component. Therefore, when bounding the probability of incorrect neighborhood estimation we have to take a network perspective, which influences the choice of $\gamma$.

**Theorem 9** (Incorrect neighborhood estimation)**.** *Considering an agent $a$ picked arbitrarily in $\mathcal{A}$, for any $\delta \in (0,1)$, employing B-ColME or C-ColME we have:*

$$\mathbb{P}\left(\exists t > \zeta_a^D, \exists a' \in \mathcal{CC}_a : \mathcal{C}_{a'}^t \neq \mathcal{C}_{a'} \cap \mathcal{N}_{a'}\right) \leqslant \frac{\delta}{2},$$

*with $\zeta_a^D = n_\gamma^\star(a) + 1$ and $\gamma = \frac{\delta}{4r|\mathcal{CC}_a|}$.*

*Proof.* The proof involves establishing a series of intermediate results that finally enable us to prove the theorem. We outline the steps of the proof below:

- First, we show that under the two proposed confidence interval parametrizations $\beta_\gamma(n)$ (Eq. (10) and (5)) and considering the sample mean $\bar{x}$ computed over $n$ samples, the probability that the *true* mean $\mu$, falls within the confidence interval $[\bar{x} - \beta_\gamma(n), \bar{x} + \beta_\gamma(n)]$ for every $n \in \mathbb{N}$, is at least $1 - 2\gamma$. This is completely equivalent to saying that the probability that the *true* mean value is outside the confidence interval for some $n \in \mathbb{N}$ is bounded above by $2\gamma$ (Lemma 10 and Proposition 11).
- Second, we remark that Lemma 10 and Proposition 11 consider a *local* perspective, taking one particular estimate $\bar{x}$ (i.e., $\bar{x}_{a,a'}^t$), together with its number of samples $n$ (i.e., $n_{a,a'}^t$). We extend this result by proving that the true mean falls within the confidence interval $I_{a,a'}$ (with high probability) for all the nodes $a' \in \mathcal{CC}_a$ in the connected component, for all records retrieved locally from the neighbors ($a'' \in \mathcal{N}_a$), and for every discrete time instant $t \in \mathbb{N}$ (Lemma 12). We will refer to this event as $E$. This result provides a *global* perspective over the entire connected component $\mathcal{CC}_a$ It is important to observe that only when the *true* value is in $I_{a,a'}$, we can provide guarantees about the correct estimation of the similarity class.
- Then, we consider the *optimistic distance* $d_\gamma^t(a, a'')$ for which we show that, conditionally over $E$, whenever it takes on strictly positive values (i.e., $d_\gamma^t(a, a'') > 0$), the neighbor $a''$ does not belong to the same similarity class $\mathcal{C}_a$ as agent $a$ (Lemma 13). As a byproduct, we also derive the minimal number of samples $n_\gamma^\star(a)$ needed to correctly decide whether a neighboring node belongs to the same equivalence class.
- At last, combining previous results we can easily obtain the claim.

**Lemma 10** (Interval parametrization)**.** *For any $\gamma \in (0,1)$, setting $\beta_\gamma(n) = \sigma\sqrt{\frac{2}{n}\left(1+\frac{1}{n}\right)\ln\frac{\sqrt{n+1}}{\gamma}}$ (if the random variables $x_t$ are sub-Gaussian with parameter $\sigma$) or $\beta_\gamma(n) = Hn^{-\alpha}$, with $\alpha < \frac{1}{4}$ and $H = \sqrt[4]{\frac{(\kappa+3)\sigma^4\zeta(2-4\alpha)}{\gamma}}$ (if $x_t$ have the first 4 polynomial moments bounded), it holds:*

$$\mathbb{P}\left(\exists n \in \mathbb{N}, |\bar{x} - \mu| \geq \beta_\gamma(n)\right) \leqslant 2\gamma \tag{13}$$

*Proof.* **($\sigma$-sub-Gaussian $x_t$)** We start from the theoretical guarantees on the parametrization (Eq. 10) of the confidence intervals from (Maillard 2019) [*Lemma 2.7*]: Let $\{x_t\}_{t\in\mathbb{N}}$ a sequence of independent real-valued random variables, where for each $t$, $x_t$ has mean $\mu_t$ and is $\sigma_t$-sub-Gaussian. Then, for all $\gamma \in (0,1)$ it holds:

$$\begin{aligned}
&\mathbb{P}\left(\exists n \in \mathbb{N}, \sum_{t=1}^n (x_t - \mu_t) \geq \sqrt{2\sum_{t=1}^n \sigma_t^2\left(1+\frac{1}{n}\right)\ln\frac{\sqrt{n+1}}{\gamma}}\right) \leqslant \gamma \\
&\mathbb{P}\left(\exists n \in \mathbb{N}, \sum_{t=1}^n (\mu_t - x_t) \geq \sqrt{2\sum_{t=1}^n \sigma_t^2\left(1+\frac{1}{n}\right)\ln\frac{\sqrt{n+1}}{\gamma}}\right) \leqslant \gamma
\end{aligned} \tag{14}$$

Our sequence of random variables can i) either correspond to the samples $x_t$ each node $a \in \mathcal{A}$ generates at each discrete time instant $t$, which is i.i.d., with mean $\mu_a$ and $\sigma$-sub-Gaussian (by assumption), or ii) the truncated sequence up to $n_{a,a'}$ the node learns by querying its neighbors, possessing the same properties. Indeed, recall that for each *locally* available estimate $\bar{x}_{a,a'}$, each node keeps also the number of samples over which that estimate has been computed $n_{a,a'}^t$. For ease of notation we will drop the subscripts and superscripts, which for the sake of the lemma are superfluous. Being $\mu_{n'} = \mu$ and $\sigma_{n'} = \sigma$ constant in our case, it is immediate to write (considering just the first inequality for compactness):

$$\mathbb{P}\left(\exists n \in \mathbb{N}, \sum_{t=1}^{n}(x_t) - n\mu \geq \sqrt{2n\sigma^2\left(1+\frac{1}{n}\right)\ln\frac{\sqrt{n+1}}{\gamma}}\right) \leqslant \gamma$$

Dividing by $n$ both sides we obtain the sample mean $\bar{x} = \frac{1}{n}\sum_t^n x_t$ (in place of the summation) and the parametrization of confidence interval $\beta_\gamma(n)$, as we have introduced in Sec. 3 (which we restate here for completeness):

$$\beta_\gamma(n) := \sigma\sqrt{\frac{2}{n}\left(1+\frac{1}{n}\right)\ln\frac{\sqrt{n+1}}{\gamma}} \tag{15}$$

By a simple substitution, we get:

$$\mathbb{P}\left(\exists n \in \mathbb{N},\, \bar{x} - \mu \geq \beta_\gamma(n)\right) \leqslant \gamma \tag{16}$$

Lastly, recall we are interested in the probability of the true mean being in the bilateral interval bounded by the given parametrization $\beta_\gamma$. Hence we can bound this probability with $2\gamma$:

$$\mathbb{P}\left(\exists n \in \mathbb{N}, |\bar{x} - \mu| \geq \beta_\gamma(n)\right) \leqslant 2\gamma$$

This proves the first part of the lemma. Moreover, considering the complementary event and noting that $\mathbb{P}(e) \leqslant 2\gamma \iff \mathbb{P}(e^c) \geq 1-2\gamma$, where $e$ is a generic event and $e^c$ its complementary, it is immediate to obtain the lower bound (complementary to Eq. 13):

$$\mathbb{P}\left(\forall n \in \mathbb{N}, |\bar{x} - \mu| < \beta_\gamma(n)\right) \geq 1 - 2\gamma \tag{17}$$

Note also that the probabilistic confidence level $\gamma$ can be considered as a function of $\delta$. By choosing appropriately the function $\gamma = f(\delta)$ it is possible to provide the desired level $\delta$ for the PAC-convergence of a given algorithm.

**($x^t$ with the first 4 bounded polynomial moments)**. Now we release the assumption that $x_t$ are extracted from sub-Gaussian distributions. We only assume that $\mathbb{E}[(x_a^t - \mu_a)^4] \leqslant \mu_4$ for any $a \in \mathcal{A}$.

We start recalling the class of concentration inequalities which generalize the classical Chebyshev inequality:

**Proposition 11.** *Given a random variable $X$ with average $\mu < \infty$ and finite $2i$-central moment $\mathbb{E}[(X-\mu)^{2i}] = \mu_{(2i)}(X)$ for any $b > 0$ we have:*

$$\mathbb{P}(|X - \mu| > b) < \frac{\mu_{(2i)}(X)}{(b)^{2i}}$$

*Moreover:*

$$\mathbb{P}(X > b) < \frac{\mathbb{E}[Y^{2i}]}{(b)^{2i}}.$$

Applying previous inequality to our estimate $\bar{x} = \sum_{t=1}^n x^t$ for the case $i = 2$ we get:

$$\mathbb{P}(|\bar{x} - \mu| > \beta(n)) < \frac{\mu_{(4)}(\bar{x})}{(\beta_\gamma(n))^4}.$$

Now observe that

$$\begin{aligned}
\mu_{(4)}(\bar{x}) :=& \mathbb{E}\left[\frac{1}{n^4}\sum_{t=1}^{n}(x^t-\mu)\sum_{\tau=1}^{n}(x^\tau-\mu)\sum_{\theta=1}^{n}(x^\theta-\mu)\sum_{\phi=1}^{n}(x^\phi-\mu)\right] \\
=& \frac{1}{n^4}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\sum_{\theta=1}^{n}\sum_{\phi=1}^{n}\mathbb{E}[(x^t-\mu)(x^\tau-\mu)(x^\theta-\mu)(x^\phi-\mu)] \\
\leq& \frac{1}{n^4}[n\kappa\sigma^4 + 3n(n-1)\sigma^4]
\end{aligned} \tag{18}$$

observe, indeed, that from the independence of samples descends that whenever $t \notin \{\tau, \theta, \phi\}$

$$\mathbb{E}[(x^t-\mu)(x^\tau-\mu)(x^\theta-\mu)(x^\phi-\mu)] = \mathbb{E}[(x^t-\mu)]\mathbb{E}[(x^\tau-\mu)(x^\theta-\mu)(x^\phi-\mu)] = 0$$

while whenever $t = \tau \neq \theta = \phi$

$$\mathbb{E}[(x^t-\mu)(x^\tau-\mu)(x^\theta-\mu)(x^\phi-\mu)] = \mathbb{E}[(x^t-\mu)^2]\mathbb{E}[(x^\theta-\mu)^2] \leqslant \sigma^4.$$

Therefore

$$\mathbb{P}(|\bar{x}-\mu|>\beta_\gamma(n)) < \frac{\kappa\sigma^4}{(n^3\beta(n))^4} + \frac{3\sigma^4}{n^2(\beta_\gamma(n))^4}.$$

Now by sub-additivity of probability we get:

$$\mathbb{P}(\exists n\in\mathbb{N}, |\bar{x}-\mu|>\beta_\gamma(n)) = \mathbb{P}(\cup_n\{|\bar{x}-\mu|>\beta_\gamma(n)\}) < \sum_n \frac{\kappa\sigma^4}{n^3(\beta_\gamma(n))^4} + \sum_n \frac{3\sigma^4}{n^2(\beta_\gamma(n))^4}.$$

Now observe that $\{\beta_\gamma(n)\}_{n\in\mathbb{N}}$ on the one hand should be chosen as small as possible and in particular we should enforce $\beta_\gamma(n)\to 0$ as $n$ grows large; on the other hand, however the choice of $\{\beta_\gamma(n)\}_{n\in\mathbb{N}}$ must guarantee that:

$$\sum_n \frac{\kappa\sigma^4}{n^3(\beta_\gamma(n))^4} + \sum_n \frac{\sigma^4}{n^2(\beta_\gamma(n))^4} \leqslant 2\gamma$$

This is possible if by choosing $\beta(n) = Hn^{-\alpha}$ for an $\alpha < \frac{1}{4}$ arbitrarily close to $1/4$ and a properly chosen $H$. Indeed with this choice of $\beta(n)$ we have:

$$\begin{aligned}\sum_n \frac{\kappa\sigma^4}{n^3(\beta_\gamma(n))^4} + \sum_n \frac{3\sigma^4}{n^2(\beta_\gamma(n))^4} &= \sum_n \frac{\kappa\sigma^4}{H^4n^{3-4\alpha}} + \sum_n \frac{3\sigma^4}{H^4n^{2-4\alpha}} = \frac{(\kappa+3)\sigma^4}{H^4}\sum_n \frac{1}{n^{2-4\alpha}}\left(1+\frac{1}{n}\right)\\ &\leqslant \frac{2(\kappa+3)\sigma^4}{H^4}\sum_n \frac{1}{n^{2-4\alpha}} = \frac{2(\kappa+3)\sigma^4}{H^4}\zeta(2-4\alpha)\end{aligned}$$

where $\zeta(z) := \sum_n \frac{1}{n^z}$, with $z\in\mathbb{C}$, denotes the $\zeta$-Riemann function. We recall that $\zeta(x)<\infty$ for any real $x>1$. Therefore by selecting

$$\beta_\gamma(n) = Hn^{-\alpha} \qquad \text{with} \quad \alpha < \frac{1}{4} \quad \text{and} \quad H = \sqrt[4]{\frac{(\kappa+3)\sigma^4\zeta(2-4\alpha)}{\gamma}}$$

we guarantee that

$$\mathbb{P}(\exists n\in\mathbb{N}, |\bar{x}-\mu|>\beta_\gamma(n)) < 2\gamma.$$

A tighter expression can be obtained as follows. Set $\beta(n) = \left(\frac{H(1+\ln^2 n)}{n}\right)^{\frac{1}{4}}$:

$$\begin{aligned}\sum_{n=1}^\infty \frac{\kappa\sigma^4}{n^3(\beta_\gamma(n))^4} + \sum_{n=1}^\infty \frac{3\sigma^4}{n^2(\beta(n))^4} &\leqslant \sum_{n=1}^\infty \frac{(\kappa+3)\sigma^4}{n^2(\beta(n))^4} = \sum_{n=1}^\infty \frac{(\kappa+3)\sigma^4}{n(1+\ln^2 n)}\\ &\leqslant \frac{(\kappa+3)\sigma^4}{H}\left(1+\sum_{n=2}^\infty \frac{1}{n(1+\ln^2 n)}\right)\\ &\leqslant \frac{(\kappa+3)\sigma^4}{H}\left(1+\sum_{n=2}^\infty \frac{1}{n\ln^2 n}\right)\\ &\leqslant \frac{(\kappa+3)\sigma^4}{H}\left(1+\frac{1}{2\ln^2 2}+\int_2^\infty \frac{1}{x\ln^2 x}\mathrm{d}x\right)\\ &= \frac{(\kappa+3)\sigma^4}{H}\left(1+\frac{1}{2\ln^2 2}+\frac{1}{\ln 2}\right)\\ &\leqslant 4\frac{(\kappa+3)\sigma^4}{H}.\end{aligned}$$

Imposing that this is smaller than $2\gamma$, we can concluded that by selecting

$$\beta_\gamma(n) = \left(2\frac{(\kappa+3)\sigma^4}{\gamma}\right)^{\frac{1}{4}}\left(\frac{1+\ln^2 n}{n}\right)^{\frac{1}{4}},$$

we guarantee that

$$\mathbb{P}(\exists n\in\mathbb{N}, |\bar{x}-\mu|>\beta_\gamma(n)) < 2\gamma.$$

**Remark 1.** *When $x^t$ exhibits a larger number of finite moments we can refine our approach by employing Proposition* (11) *for a different (larger) choice of $i$. So doing we obtain a more favorable behavior for $\beta_\gamma(n)$. In particular we will get that*

$$\beta_\gamma(n) = O(n^{-\alpha}) \qquad \textit{with} \qquad \alpha < \frac{1}{2}\left(1 - \frac{1}{i}\right)$$

*At last we wish to emphasize that $\beta_\gamma(n)$ can not be properly defined when distribution $D_a$ exhibit less than four bounded polynomial moments. The application of Chebyshev inequality ($i = 1$), indeed, would lead to a too the following weak upper bound:*

$$\mathbb{P}(\exists n \in \mathbb{N}, |\bar{x} - \mu| > \beta - \gamma(n)) < \sum_n \frac{\sigma^2}{n(\beta_\gamma(n))^2}.$$

*Observe, indeed, that since $\zeta(1) = \sum_n \frac{1}{n}$ diverges, it is impossible the find of suitable expression for $\{\beta(n)\}$ which jointly satisfy: $\lim_{n\to\infty} \beta_\gamma(n) = 0$ and $\sum_n \frac{\sigma^2}{n(\beta_\gamma(n))^2} < \infty$.*

□

This result is the first fundamental building block to define a notion of distance (which uses the estimates $\bar{x}$ and the parametrization $\beta_\gamma$) for which it is possible to provide guarantees about the class membership.

We have bounded the probability of not having the *true* mean within the $\beta_\gamma$ confidence interval given a certain estimate $\bar{x}$ and the corresponding number of samples $n$. We now have to take a global perspective, so we consider the event $E := \left\{\forall a' \in \mathcal{CC}_a, \forall t \in \mathbb{N}, \forall a'' \in \mathcal{N}_{a'}, \left|\bar{x}^t_{a',a''} - \mu_{a''}\right| < \beta_\gamma(n)\right\}$, which is equivalent to say that, for every node $a' \in \mathcal{CC}_a$ in the connected component of node $a \in \mathcal{A}$ and for every instant $t \in \mathbb{N}$, the *true* mean value of each of the neighbors $a''$ of node $a'$, given the info $a'$ is able to collect (neighbor's sample mean and the number of samples), is within the confidence interval $I_{a',a''}$. We show that this holds with high probability with an appropriate choice of $\gamma$:

**Lemma 12** (Confidence of $\beta_\gamma$ interval)**.** *Considering the interval parametrization $\beta_\gamma(n)$ (Eq. (10) or (5)), setting $\gamma(\delta) = \frac{\delta}{4r|\mathcal{CC}_a|}$, it holds:*

$$\mathbb{P}\left(\forall a' \in \mathcal{CC}_a, \forall t \in \mathbb{N}, \forall a'' \in \mathcal{N}_{a'}, \left|\bar{x}^t_{a',a''} - \mu_{a''}\right| < \beta_\gamma(n^t_{a',a''})\right) \geq 1 - \frac{\delta}{2} \tag{19}$$

*Proof.* We have introduced the event $E = \left\{\forall a' \in \mathcal{CC}_a, \forall t \in \mathbb{N}, \forall a'' \in \mathcal{N}_{a'}, \left|\bar{x}^t_{a',a''} - \mu_{a''}\right| < \beta_\gamma(n^t_{a',a''})\right\}$, it is more convenient to work with the complementary event:

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^c) = 1 - \mathbb{P}\left(\exists a' \in \mathcal{CC}_a, \exists t \in \mathbb{N}, \exists a'' \in \mathcal{N}_{a'} \,:\, \left|\bar{x}^t_{a',a''} - \mu_{a''}\right| > \beta_\gamma(n^t_{a',a''})\right)$$

Applying a union bound with respect to the nodes $a' \in \mathcal{CC}_a$ and neighbors $a'' \in \mathcal{N}_{a'}$, and using Lemma 10 ($\mathbb{P}(\exists n \in \mathbb{N}, |\bar{x} - \mu| \leqslant \beta_\gamma(n)) \geq 2\gamma$), we can immediately obtain a lower bound on the probability of the event $E$:

$$\begin{aligned}\mathbb{P}(E) &\geq 1 - \sum_{a' \in \mathcal{CC}_a} \sum_{a'' \in \mathcal{N}_{a'}} \mathbb{P}\left(\exists t \in \mathbb{N} : \left|\bar{x}^t_{a',a''} - \mu_{a''}\right| \geq \beta_\gamma(n_{a',a''})\right) \\ &\geq 1 - r|\mathcal{CC}_a|(2\gamma) = 1 - 2r|\mathcal{CC}_a|\gamma\end{aligned}$$

Now, we set $\gamma = \frac{\delta}{4r|\mathcal{CC}_a|}$ and thus we immediately obtain $\mathbb{P}(E) \geq 1 - \frac{\delta}{2}$. This explains the value of the constant $\gamma$ in the theorem. The above bound would then be used in the $(\varepsilon - \delta)$-convergence of the B-ColME and C-ColME algorithm. □

This result provides a probabilistic bound for the situation in which the true value is not within the confidence interval and for which we cannot provide theoretical guarantees.

At this point, assuming that event $E$ holds (with high probability due to Lemma 12), we need to show that the *optimistic* distance $d^t_\gamma(a, a')$ allows an agent to discriminate whether one of its neighbors belongs to the same similarity class $\mathcal{C}_a$.

**Lemma 13** (Class membership rule)**.** *Conditionally over the event $E$, we have*

$$d^t_\gamma(a, a') > 0 \iff a' \notin \mathcal{N}_a \cap \mathcal{C}_a \tag{20}$$

*Proof.* Defined the *optimistic* distance[3] as $d^t_\gamma(a, a') := \left|\bar{x}^t_{a,a} - \bar{x}^t_{a,a'}\right| - \beta_\gamma(n^t_{a,a}) - \beta_\gamma(n^t_{a,a'})$ and denoted with $\Delta_{a,a'} = |\mu_a - \mu_{a'}|$ the gaps between the *true* mean of agents belonging to different similarity classes, by summing and subtracting $(\mu_a - \mu_{a'})$ inside the absolute value, it is immediate to obtain:

$$d^t_\gamma(a, a') = \left|(\mu_a - \mu_{a'}) + (\bar{x}^t_{a,a} - \mu_a) - (\bar{x}^t_{a,a'} - \mu_{a'})\right| - \beta_\gamma(n^t_{a,a}) - \beta_\gamma(n^t_{a,a'}) \tag{21}$$

[3]For ease of notation we will use $a, a'$, instead of $a', a''$ as we did in the previous Lemma.

Now, we show that conditionally over the event $E$, $d^t_\gamma(a,a')$ satisfies two inequalities which allow us to determine whether two nodes belong to the same similarity class by looking at the sign of $d^t_\gamma(a,a')$.

**(Forward implication)** First, let us apply the triangular inequality on the absolute value in Eq. (21) and bound it with the sum of the absolute values of the addends:

$$d^t_\gamma(a,a') \leqslant \Delta_{a,a'} + \left|\bar{x}^t_{a,a} - \mu_a\right| + \left|\bar{x}^t_{a,a'} - \mu_{a'}\right| - \beta_\gamma(n^t_{a,a}) - \beta_\gamma(n^t_{a,a'})$$

Now, conditionally over $E$, we have that $\left|\bar{x}^t_{a,a} - \mu_a\right| \leq \beta_\gamma(n^t_{a,a})$ and $\left|\bar{x}^t_{a,a'} - \mu_{a'}\right| \leq \beta_\gamma(n^t_{a,a'})$. Therefore whenever $a' \in \mathcal{C}_a \cup \mathcal{N}_a$, i.e., $\Delta_{a,a'} = 0$, we have:

$$d^t_\gamma(a,a') \leqslant \left|\bar{x}^t_{a,a} - \mu_a\right| - \beta_\gamma(n^t_{a,a}) + \left|\bar{x}^t_{a,a'} - \mu_{a'}\right| - \beta_\gamma(n^t_{a,a'}) \leqslant 0$$

So, conditionally over $E$, and if $a' \in \mathcal{C}_a \cup \mathcal{N}_a$ the optimistic distance $d^t_\gamma(a,a')$ is smaller or equal than 0, i.e., $a' \in \mathcal{C}_a \cup \mathcal{N}_a \implies d^t_\gamma(a,a') \leqslant 0$. Considering the contrapositive statement, we immediately prove the **forward implication** of the lemma, namely:

$$d^t_\gamma(a,a') > 0 \implies a' \notin \mathcal{C}_a \cup \mathcal{N}_a.$$

**(Backward implication)** At this point, we need to show that conditionally over the event $E$, whenever two nodes do not belong to the same similarity class, then the optimistic distance is positive (i.e., $a' \notin \mathcal{C}_a \cup \mathcal{N}_a \implies d^t_\gamma(a,a') > 0$). To do so, we start from Eq. (21), aiming at deriving a lower bound for $d^t_\gamma(a,a')$. By applying the reverse triangular inequality ($|a-b| > ||a|-|b|| \implies |a|-|b| \leqslant |a-b|$):

$$d^t_\gamma(a,a') \geq \left|\Delta_{a,a'} + \left(\bar{x}^t_{a,a} - \mu_a\right)\right| - \left|\bar{x}^t_{a,a'} - \mu_{a'}\right| - \beta_\gamma(n^t_{a,a}) - \beta_\gamma(n^t_{a,a'})$$

And then recalling that $|a+b| \geq |a| - |b|$, we get:

$$d^t_\gamma(a,a') \geq \Delta_{a,a'} - \left|\left(\bar{x}^t_{a,a} - \mu_a\right)\right| - \left|\bar{x}^t_{a,a'} - \mu_{a'}\right| - \beta_\gamma(n^t_{a,a}) - \beta_\gamma(n^t_{a,a'})$$

Again, conditionally over $E$, and we have $\left|\bar{x}^t_{a,a} - \mu_a\right| \leqslant \beta_\gamma(n^t_{a,a})$ and $\left|\bar{x}^t_{a,a'} - \mu_{a'}\right| \leq \beta_\gamma(n^t_{a,a'})$. Therefore we can write:

$$\begin{aligned} d^t_\gamma(a,a') &\geq \Delta_{a,a'} - \left|\left(\bar{x}^t_{a,a} - \mu_a\right)\right| - \left|\bar{x}^t_{a,a'} - \mu_{a'}\right| - \beta_\gamma(n^t_{a,a}) - \beta_\gamma(n^t_{a,a'}) \\ &\geq \Delta_{a,a'} - 2\beta_\gamma(n_{a,a}) - 2\beta_\gamma(n^t_{a,a'}) \end{aligned}$$

Moreover, consider that by definition[4] we have $n^t_{a,a} \geq n^t_{a,a'}$, thus $\beta_\gamma(n^t_{a,a}) \leqslant \beta_\gamma(n^t_{a,a'})$, so we can write:

$$d^t_\gamma(a,a') \geq \Delta_{a,a'} - 4\beta_\gamma(n^t_{a,a'})$$

Now we need to observe that, whereas conditionally over $E$ in the previous case ($a' \in \mathcal{C}_a \cup \mathcal{N}_a$) the optimistic distance always keeps negative (simply take in mind that by definition $\beta_\gamma(0) = +\infty$). When neighbor $a'$ belongs to a different similarity class of $a$ (i.e., $a' \notin \mathcal{C}_a \cup \mathcal{N}_a$), the optimistic distance $d^t_\gamma(a,a')$ will become positive (thus signaling $a$ and $a'$ belong to different similarity classes), as soon as the collected number of samples $n^t_{a,a'}$ becomes sufficiently large to guarantee $\beta_\gamma(n^t_{a,a'}) < \frac{\Delta_{a,a'}}{4}$.

Now denoted with $\beta^{-1}_\gamma(x)$ the inverse function of $\beta_\gamma(n)$, and defined:

$$n^\star_{a,a'} = \left\lceil \beta^{-1}_\gamma\left(\frac{\Delta_{a,a'}}{4}\right)\right\rceil, \tag{22}$$

conditionally over $E$, $\forall\, n_{a,a'} \geq n^\star_{a,a'}$, we have $a' \notin \mathcal{C}_a \cup \mathcal{N}_a \implies d^t_\gamma(a,a') > 0$. And this proves the **backward** implication, which concludes the proof of the lemma. □

To conclude our proof, observe that according to our scalable algorithms, at each time instant $t \in \mathbb{N}$ each node $a \in \mathcal{A}$ queries all the nodes that were in its *estimated* similarity class at the previous step $\mathcal{C}^{t-1}_a$ (for $t = 0$ all the neighbors $a' \in \mathcal{N}_a$ are contacted). Therefore, all received estimates $\bar{x}^t_{a,a'}$ suffer for a delay of 1 time instant, i.e., $n_{a,a'} = t-1$. Now, Lemma 12 ensures that, by choosing $\gamma = \frac{\delta}{4r|\mathcal{CC}_a|}$, we have $\mathbb{P}(E) \geq 1 - \frac{\delta}{2}$. Moreover, considering:

$$n^\star_\gamma(a) = \left\lceil \beta^{-1}_\gamma\left(\frac{\Delta_a}{4}\right)\right\rceil \tag{23}$$

where recall that $\Delta_a = \min_{a' \in \mathcal{A}\setminus\mathcal{C}_a} \Delta_{a,a'}$.

By Lemma 13 we have that, conditionally over $E$, as soon as $t-1 \geq n^\star_{a,a'}$ we have $d^t_\gamma(a,a') > 0$ for all pairs of neighboring nodes $(a,a')$ belonging to different similarity classes, while $d^t_\gamma(a,a') \leqslant 0$ for all pairs of neighboring nodes $(a,a')$ belonging to the same similarity class. Therefore, whenever $t-1 \geq n^\star_\gamma(a)$ this holds for all the pairs in the connected component $\mathcal{CC}_a$, as $\Delta_a \geq \min_{a' \in \mathcal{CC}_a \setminus \mathcal{C}_a} \Delta_{a,a'}$.

□

[4] As a matter of fact, for our scalable algorithms, the inequality is always strict as $n^t_{a,a} = t$ and $n^t_{a,a'} = t-1$.

# C Appendix C - Extension to the Multidimensional Case

In this Appendix, we show briefly as the previous approach can be generalized to the multidimensional case. We assume that at each instant $t$ every agent $a$ generates a new sample $\boldsymbol{x}_a^t \in \mathbb{R}^K$ drawn i.i.d. from distribution $D_a$ with expected value $\boldsymbol{\mu}_a = \mathbb{E}[\boldsymbol{x}_a^t]$. In particular, we show how the definition of confidence intervals $\beta_\gamma(n)$ can be extended to the multidimensional case. Let $\Delta_{a,a'} := ||\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'}||$.

## C.1 Sub-Gaussian case

In this case, we assume distributions $D_a$ to be multidimensional sub-Gaussians. Recalling the definition of multidimensional sub-Gaussian random variables, we have:

**Definition 1.** *A random vector $\boldsymbol{x} \in \mathbb{R}^K$ is said to be sub-Gaussian with variance proxy $\sigma^2$ if it is centered and for any $\boldsymbol{u} \in \mathbb{R}^K$ such that $||u|| = 1$, the real random variable $\boldsymbol{u}^T\boldsymbol{x}$ is sub-Gaussian with variance proxy $\sigma^2$.*

Therefore, we can choose an orthonormal basis $\boldsymbol{u}_i$ $\forall 1 \leq i \leq K$ in $\mathbb{R}^K$ and consider the uni-dimensional projections of samples along directions induced by $\boldsymbol{u}_i$, $\forall i$. Since projected samples are uni-dimensional sub-Gaussian, the theory developed in the previous sections applies to every projection $i$. In particular, we can compute the *optimistic* distance along axes $i$ as:

$$d^t_{i,\gamma/K}(a,a') := \left|\boldsymbol{u}_i^T(\bar{\boldsymbol{x}}^t_{a,a} - \bar{\boldsymbol{x}}^t_{a,a'})\right| - \beta_{\gamma/K}(n^t_{a,a}) - \beta_{\gamma/K}(n^t_{a,a'})$$

and decide that $a'$ is maintained in $\mathcal{C}_a^t$, as long as $d^t_{i,\gamma/K}(a,a') \leq 0 \; \forall \; 1 \leq i \leq d$. At last observe that $\Delta^{(i)}_{a,a'} = \boldsymbol{u}_i^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})$ satisfies:

$$\max_i \Delta^{(i)}_{a,a'} \geq \frac{1}{\sqrt{K}}\Delta_{a,a'}$$

as result of simple geometrical arguments.

## C.2 Fourth bounded moment

We start assuming that that $\mathbb{E}[||(\boldsymbol{x}_a^t - \boldsymbol{\mu}_a)||^4] < \infty$ for any $a \in \mathcal{A}$. Let $\bar{\boldsymbol{x}}_a = \frac{1}{n}\sum_{t=1}^n \boldsymbol{x}^t$ and $Y_a := ||\bar{\boldsymbol{x}}_a - \boldsymbol{\mu}_a|| \geq 0$, by applying Proposition 11 to $Y$ we have:

$$\mathbb{P}(|Y_a| > \beta_\gamma(n)) \leq \frac{\mathbb{E}[Y_a^4]}{\beta_\gamma(n)^4}$$

Now defined with $\mu_{i,a}$ the expectation of the $i$-th component of samples $\boldsymbol{x}_a^t$, i.e:

$$\mu_{i,a} := \mathbb{E}[x^t_{i,a}]$$

considering the Euclidean norm, we have:

$$\mathbb{E}[Y_a^4] = \mathbb{E}\left[\left(\sum_{i=1}^K (\bar{x}_{i,a} - \mu_{i,a})^2\right)^2\right] = \sum_{i=1}^K\sum_{h=1}^K \mathbb{E}[(\bar{x}_{i,a} - \mu_{i,a})^2(\bar{x}_{h,a} - \mu_{h,a})^2]$$
$$\leq \sum_{i=1}^K\sum_{h=1}^K \sqrt{\mathbb{E}[(\bar{x}_{i,a} - \mu_{i,a})^4]\mathbb{E}[(\bar{x}_{h,a} - \mu_{h,a})^4]} \leq K^2 \max_i \mathbb{E}[(\bar{x}_{i,a} - \mu_{i,a})^4]$$

where the second to last inequality follows by the applcation of the Cauchy-Schwarz inequality.

Now proceeding exactly as in (18) we obtain that for every $i$

$$\mathbb{E}[(\bar{x}_{i,a} - \mu_{i,a})^4] \leq \frac{1}{n^4}[n\kappa_i\sigma_i^4 + 3n(n-1)\sigma_i^4]$$

Now by sub-additivity of probability we get, recalling that $Y_a = ||\bar{\boldsymbol{x}}_a - \boldsymbol{\mu}_a||$:

$$\mathbb{P}(\exists n \in \mathbb{N}, Y_a > \beta_\gamma(n)) = \mathbb{P}(\cup_n\{Y_a > \beta_\gamma(n)\}) < K^2 \max_i \left(\sum_n \frac{\kappa_i\sigma_i^4}{n^3(\beta_\gamma(n))^4} + \sum_n \frac{3\sigma_i^4}{n^2(\beta_\gamma(n))^4}\right).$$

Then proceeding as in the proof of Theorem 1, (i.e. forcing the r.h.s to be less than $2\gamma$) we obtain the following expression for $\beta_\gamma(n)$:

$$\beta_\gamma(n) = \max_i \sqrt{K}\left(2\frac{(\kappa_i + 3)\sigma_i^4}{\gamma}\right)^{\frac{1}{4}}\left(\frac{1 + \ln^2 n}{n}\right)^{\frac{1}{4}}.$$

At last we have to generalize the definition of *optimistic* distance as follows:

$$d^t_\gamma(a,a') := \left|\left|\bar{\boldsymbol{x}}^t_{a,a} - \bar{\boldsymbol{x}}^t_{a,a'}\right|\right| - \beta_\gamma(n^t_{a,a}) - \beta_\gamma(n^t_{a,a'}).$$

Then proceeeding exactly as in the previous section, we can extend Lemma 13 statement to the multidimensional case (we recall that by triangular inequality $\left|\left|\bar{\boldsymbol{x}}^t_{a,a} - \bar{\boldsymbol{x}}^t_{a,a'}\right|\right| = \left|\left|(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'}) + (\bar{\boldsymbol{x}}^t_{a,a} - \boldsymbol{\mu}_a) - (\bar{\boldsymbol{x}}^t_{a,a'} - \boldsymbol{\mu}_{a'})\right|\right| \leq \Delta_{a,a'} + \left|\left|\bar{\boldsymbol{x}}^t_{a,a} - \boldsymbol{\mu}_a\right|\right| + \left|\left|\bar{\boldsymbol{x}}^t_{a,a'} - \boldsymbol{\mu}_{a'}\right|\right|$, as well as by reverse triangular inequality $\left|\left|\bar{\boldsymbol{x}}^t_{a,a} - \bar{\boldsymbol{x}}^t_{a,a'}\right|\right| \geq \Delta_{a,a'} - \left|\left|\bar{\boldsymbol{x}}^t_{a,a} - \boldsymbol{\mu}_a\right|\right| - \left|\left|\bar{\boldsymbol{x}}^t_{a,a'} - \boldsymbol{\mu}_{a'}\right|\right|$).

## C.3 Numerical Experiments

We experimentally assess the 'dicovery' phase, where nodes attempt to find peers in their neighborhood $\mathcal{N}_a$ that belong to the same similarity class (note that the guarantees we established for the estimate process, whether trough the message-passing scheme or consensus, hold for each component).

We consider a setup similar to Section 6, with $N = 10000$ agents, belonging to one of two classes: the first characterized by an all-zero vector mean $\boldsymbol{\mu_1} = \mathbf{0}$ and the second by $\boldsymbol{\mu_2} = \frac{1}{\sqrt{K}}\mathbf{1}$. Samples are drawn from a multivariate Gaussian distribution, where the common variance (the diagonal element in the covariance matrix) is $\sigma^2 = 4.0$. We consider the canonical basis , i.e., $\{\mathbf{e}_i\}_{i\in\{1,\dots,K\}}$ [5], as the vectors for projecting our vectorial samples. The covariance matrix can be arbitrary, provided that the diagonal elements (the variances) are $\sigma^2$. We consider $\varepsilon = 0.1$ and $\delta = 0.1$, averaging over 10 different realizations of the process.

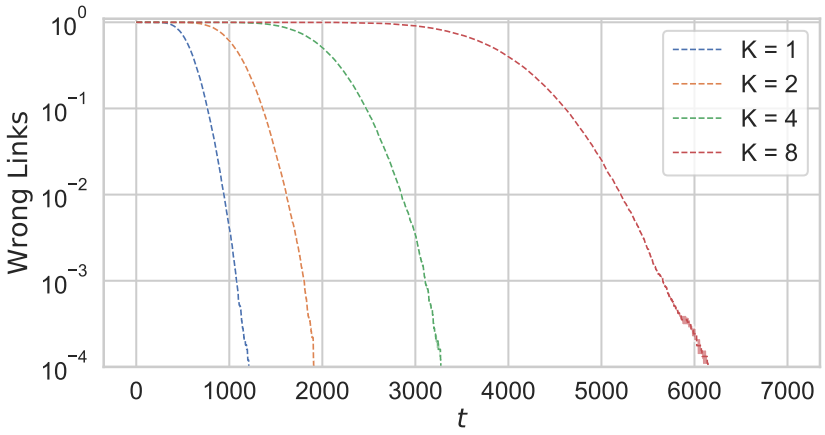



Figure 4: Fraction of the *wrong* links (log scale) over time, as a function of the dimension $K$ of the mean values, $\boldsymbol{x} \in \mathbb{R}^K$.

As the number of dimensions $K$ increases, the time required to discover peers in the same class also increases. However, doubling the number of dimensions results in less than a doubling of the discovery time. This slowdown occurs because the bounds on each component need to be more stringent (note the factor $\frac{\gamma}{K}$ in Sec. C.1) to guarantee the same bound on the probability of incorrectly hindering a connection to a same-class peer. However, it's also important to note that if any of the $K$ components of the distance $d^t_{\gamma/K}(a, a')$ becomes positive, it is a condition sufficient to remove the incorrect link. This mechanism contributes to the sublinear increase in discovery time withrespect to $K$.

[5] The vector $\mathbf{e}_i = (0, ..., 1, ..., 0)$ is the all-zero vector with a a 1 in its i-th component.

# D Appendix D - Schematic Representation for B-ColME

We provide a sketch for the functioning of the B-ColME which, together with Algorithm 2 (in the main text) provides a detailed explanation of the proposed algorithm.

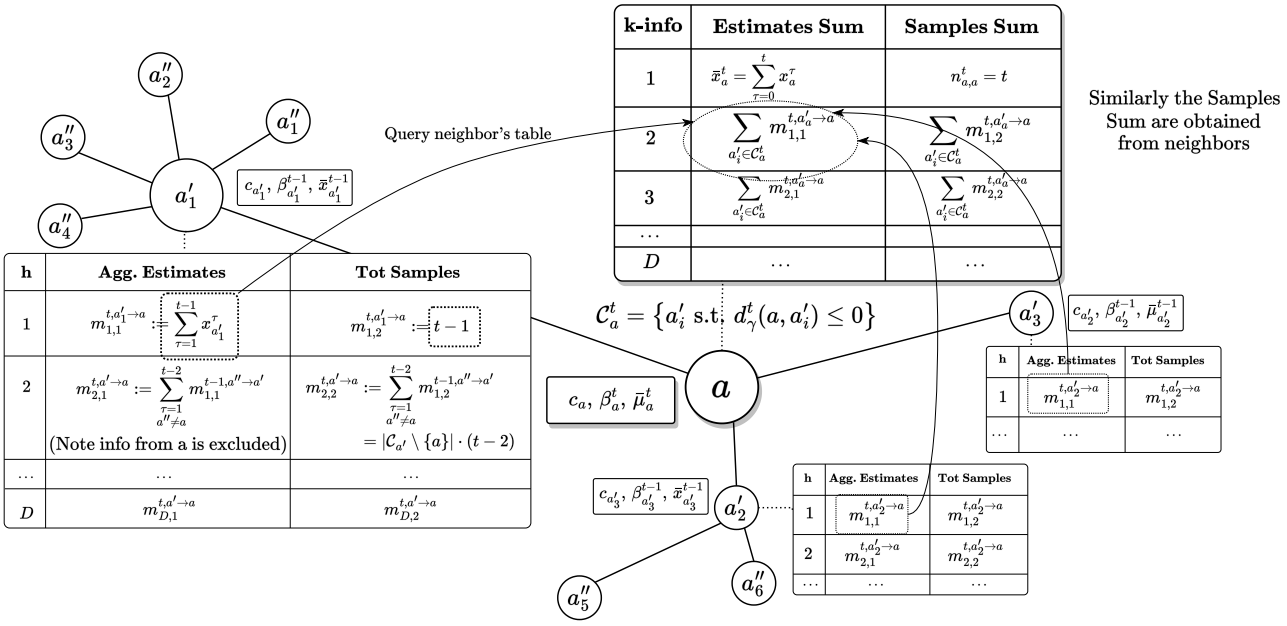



Figure 5: Simplified illustration of the functioning of the B-ColME algorithm from the point of view of node $a$ (all the quantities are already aggregated in the messages $m_{h,i}^{t,a' \to a}$, so that to exclude the "self" info sent by $a$).

# E Appendix E - Proof for B-ColME - Theorem 4

First, we recall the notion of $(\varepsilon, \delta)$-convergence (also referred to as PAC-convergence), that we use to assess theoretically the performance of the estimation algorithms. The definition is as follows:

**Definition 2** (PAC-convergence)**.** *An estimation procedure for agent $a$ is called $(\varepsilon, \delta)$-convergent if there exist $\tau_a \in \mathbb{N}$ such that:*

$$\mathbb{P}\left(\forall t > \tau_a, \left|\hat{\mu}_a^t - \mu_a\right| \leq \varepsilon\right) > 1 - \delta$$

Here we provide a convergence result in the sense of the above definition for B-ColME, we will derive a similar result for C-ColME in Appendix F.

**Theorem 14.** *Provided that $\mathcal{CC}_a^d$ has a tree structure, for any $\delta \in (0,1)$, employing B-ColME(d), we have:*

$$\mathbb{P}\left(\forall t > \tau_a^B, |\hat{\mu}_a^t - \mu_a| < \varepsilon\right) \geq 1 - \delta$$
$$\textit{with} \quad \tau_a^B = \max\left[\zeta_D + d, \frac{\widetilde{n}_{\frac{\delta}{2}}}{|\mathcal{CC}_a^d|} + d\right]$$

*where*

$$\widetilde{n}_{\frac{\delta}{2}}(\varepsilon) := \begin{cases} \min_n \left\{ n : \sum_{n+1}^{\infty} 2Q\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) < \frac{\delta}{2} \right\} & \textit{Gaussian distribitions} \\ \min_n \left\{ n : \sum_{n+1}^{\infty} 2\exp\left(\frac{n\varepsilon^2}{2\sigma^2}\right) < \frac{\delta}{2} \right\} & \textit{sub-Gaussian distributions} \\ \min_n \left\{ n : \sum_{n+1}^{\infty} \frac{\mu_4 + 2(\sigma^2)^2}{(\varepsilon n)^2} < \frac{\delta}{2} \right\} & \textit{distrib. with bounded fourth moment} \end{cases}$$

*with*

$$\widetilde{n}_{\frac{\delta}{2}}(\varepsilon) \leq \begin{cases} \left\lceil -\frac{\sigma^2}{\varepsilon^2} \log\left(\frac{\delta}{2}\left(1 - \mathrm{e}^{-\frac{\varepsilon^2}{\sigma^2}}\right)\right) \right\rceil & \textit{Gaussian distributions} \\ \left\lceil -\frac{\sigma^2}{\varepsilon^2} \log\left(\frac{\sqrt{2\pi}\varepsilon\delta}{2\sigma}\left(1 - \mathrm{e}^{-\frac{\varepsilon^2}{\sigma^2}}\right)\right) \right\rceil & \textit{sub-Gaussian distributions} \\ \left\lceil \frac{2(\kappa+3)\sigma^4}{\delta\varepsilon^4} \right\rceil & \textit{distrib. with bounded fourth moment} \end{cases}$$

where we recall that Gaussian distributions belong to the class of sub-Gaussians, and therefore the bound derived for sub-Gaussian distributions, can be applied also to Gaussian.

*Proof.* We start considering the case in which the distribution of samples for every agent $a$ is Gaussian (normal) with the same standard deviation $\sigma$, i.e. $D_a = \mathcal{N}(\mu_a, \sigma), \forall a$. In such a case, if we consider an empirical average $\bar{x}(n) = \frac{1}{n}\sum_1^n x_t$ where $x_t$ are i.i.d extracted from $D_a$, for some $a \in \mathcal{A}$, we have

$$\mathbb{P}(|\bar{x}(n) - \mu_a| > \varepsilon) = 2Q\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)$$

indeed observe that $\bar{x}(n)$, as immediate consequence of the elementary properties of normal random variables, is distributed as a Gaussian, with zero mean and standard deviation equal to $\frac{\sigma}{\sqrt{n}}$.

Then given an arbitrary $n_0 \in \mathbb{N}$,

$$\mathbb{P}(\exists n > n_0 \,:\, |\bar{x}(n) - \mu_a| > \varepsilon) \leq \sum_{n_0+1}^{\infty} \mathbb{P}(|\bar{x}(n) - \mu_a| > \varepsilon) = \sum_{n_0+1}^{\infty} 2Q\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)$$

Observe that $\sum_1^{\infty} 2Q\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)$ converges, therefore we can safely define

$$\widetilde{n}_{\frac{\delta}{2}}(\varepsilon) := \min_{n_0}\left\{ n_0 : \sum_{n_0+1}^{\infty} 2Q\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) < \frac{\delta}{2} \right\}.$$

Now, to conclude the proof, observe that conditionally over the fact that $\mathcal{C}_a^t = \mathcal{C}_a \cap \mathcal{N}_a$ for every time $t \geq t_0 - d$ , agent $a$ computes its average $\hat{\mu}_a$ as an average of samples collected by all agents in $\mathcal{CC}_a^d$. Such samples are i.i.d. and follows $D_a$; moreover its number easily lower bounded by $|\mathcal{CC}_a^d|(t_0 - d)$. Therefore whenever $|\mathcal{CC}_a^d|(t_0 - d) \geq \widetilde{n}_{\frac{\delta}{2}}$ by the definition of $\widetilde{n}_{\frac{\delta}{2}}$ we have

$$\mathbb{P}\left(\exists t > t_0 \,:\, |\hat{\mu}_a^t - \mu_a| > \varepsilon \;\middle|\; |\mathcal{CC}_a^d|(t_0 - d) \geq \widetilde{n}_{\frac{\delta}{2}}, \mathcal{C}_a^{t_0 - d} = \mathcal{C}_a \cap \mathcal{N}_a \,\forall a \in \mathcal{A}\right) < \frac{\delta}{2}.$$

The claim descends from the definition of $\zeta_D$ in Theorem 1.

Now consider the more general case in which the distribution of samples at nodes are Gaussian with possibly different standard deviations, uniformly bounded by $\sigma$, i.e., $D_a = \mathcal{N}(\mu_a, \sigma_a)$ with $\sigma_a \leq \sigma$, $\forall a \in \mathcal{A}$. In such a case if we consider an

empirical average $\bar{x}(n) = \frac{1}{n}\sum_1^n x_t$ of independent samples $x_t$ extracted from Gaussian distributions with the same average $\mu_a$, but with possibly different standard deviations $\sigma_a \leq \sigma$, we have that $\bar{x}(n)$ is distributed as a Gaussian with zero mean and standard deviation smaller or equal than $\frac{\sigma}{\sqrt{n}}$. Therefore:

$$\mathbb{P}(|\bar{x}(n) - \mu_a| > \varepsilon) \leq 2Q\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)$$

and we can proceed exactly as in the previous case.

Now previous approach rather immediately extends to the case in which $D_a$ are sub-Gaussian with parameter $\sigma$, since in this case $\bar{x}(n) = \frac{1}{n}\sum_1^n x_t$ of independent samples $x_t$ extracted from sub-Gaussian distributions with parameter $\sigma$ having the same average $\mu_a$ is sub-Gaussian with parameter $\sigma/\sqrt{n}$ and zero mean, and therefore by definition of sub-Gaussian (see (Vershynin 2018) for a discussion about sub-Gaussian distributions and their properties):

$$\mathbb{P}(|\bar{x}(n) - \mu_a| > \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

Now since $\sum_1^\infty \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$ is a converging geometrical series, we can safely define

$$\widetilde{n}_{\frac{\delta}{2}}(\varepsilon) := \min_{n_0}\left\{n_0 : \sum_{n_0+1}^{\infty} 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) < \frac{\delta}{2}\right\}.$$

and proceed as in the previous case.

At last consider the case in which $\mu_a$ have a fourth central moment uniformly bounded bu $\mu_4$, and a variance uniformly bounded by $\sigma^2$. In this case considering the empirical $\bar{x}(n) = \frac{1}{n}\sum_1^n x_t$ of independent samples $x_t$ extracted from arbitrary distributions with the same average $\mu_a$, following the same approach as in the proof of Proposition 11 we can bound its fourth moment as:

$$\mathbb{E}[(\bar{x}(n) - \mu_a)^4] \leq \frac{1}{n^4}[n\kappa + 3n(n-1)]\sigma^4$$

Therefore applying Chebyshev inequality (see Proposition 11) we have:

$$\mathbb{P}(|\bar{x}(n) - \mu_a| \geq \varepsilon) < \frac{\mathbb{E}[(\bar{x}(n) - \mu_a)^4]}{\varepsilon^4} \leq \frac{1}{(\varepsilon n)^4}[n\kappa + 3n(n-1)]\sigma^4$$

Then given an arbitrary $n_0 \in \mathbb{N}$

$$\mathbb{P}(\exists n > n_0,\, |\bar{x}(n) - \mu_a| > \varepsilon) \leq \sum_{n_0+1}^{\infty}\mathbb{P}(|\bar{x}(n) - \mu_a| > \varepsilon) = \sum_{n_0+1}^{\infty}\frac{1}{(\varepsilon n)^4}[[n\kappa + 3n(n-1)]\sigma^4] \leq \frac{(\kappa+3)\sigma^4}{\varepsilon^4 n^2}$$

again $\sum_1^\infty \frac{(\kappa+3)\sigma^4}{\varepsilon^4 n^2}$ converges, therefore we can define

$$\widetilde{n}_{\frac{\delta}{2}}(\varepsilon) = \min_{n_0}\left\{n_0 : \sum_{n_0+1}^{\infty}\frac{(\kappa+3)\sigma^4}{\varepsilon^4 n^2} < \frac{\delta}{2}\right\}$$

Then we can proceed exactly as in previous cases.

As last step, now we derive easy upper bounds for $\widetilde{n}_{\frac{\delta}{2}}$. We start from the last case. We have

$$\sum_{n_0+1}^{\infty}\frac{1}{n^2} \leq \int_{n_0+1}^{\infty}\frac{1}{(x-1)^2}\,\mathrm{d}x = \frac{1}{n_0}$$

and

$$\sum_{n_0+1}^{\infty}\frac{(\kappa+3)\sigma^4)}{\varepsilon^4 n^2} \leq \frac{(\kappa+3)\sigma^4}{\varepsilon^4 n_0} \qquad n_0 > 1$$

From which we obtain that:

$$\widetilde{n}_{\frac{\delta}{2}} \leq \left\lceil\frac{2(\kappa+3)\sigma^4}{\delta\varepsilon^4}\right\rceil$$

Now considering the Gaussian case, in this case we can exploit the following well known bounds for $Q(x)$:

$$\frac{x}{1+x^2}\theta(x) < Q(x) < \frac{1}{x}\theta(x) \qquad \forall x \geq 0$$

with $\theta(x) = \frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$. Therefore

$$\sum_{n_0+1}^{\infty} Q\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) \leq \sum_{n_0+1}^{\infty} \frac{\sigma \mathrm{e}^{-\frac{n\varepsilon^2}{2\sigma^2}}}{\sqrt{2\pi n}\varepsilon} \leq \sum_{n_0+1}^{\infty} \frac{\sigma \mathrm{e}^{-\frac{n\varepsilon^2}{2\sigma^2}}}{\sqrt{2\pi}\varepsilon} = \frac{\sigma}{\sqrt{2\pi}\varepsilon} \mathrm{e}^{-\frac{n_0\varepsilon^2}{2\sigma^2}} \sum_{n=1}^{\infty} \mathrm{e}^{-\frac{n\varepsilon^2}{2\sigma^2}}$$

with

$$\sum_{n=1}^{\infty} \mathrm{e}^{-\frac{n\varepsilon^2}{2\sigma^2}} = \frac{1}{1-\mathrm{e}^{-\frac{\varepsilon^2}{2\sigma^2}}}$$

Therefore imposing

$$2\frac{\sigma}{\sqrt{2\pi}\varepsilon} \mathrm{e}^{-\frac{\widetilde{n}_{\frac{\delta}{2}}\varepsilon^2}{2\sigma^2}} \frac{1}{1-\mathrm{e}^{-\frac{\varepsilon^2}{2\sigma^2}}} \leq \frac{\delta}{2}$$

we obtain the following upper bound on $\widetilde{n}_{\frac{\delta}{2}}$

$$\widetilde{n}_{\frac{\delta}{2}} \leq \left\lceil -\frac{2\sigma^2}{\varepsilon^2} \ln\left(\frac{\sqrt{2\pi}\varepsilon\delta}{4\sigma}\left(1-\mathrm{e}^{-\frac{\varepsilon^2}{2\sigma^2}}\right)\right) \right\rceil$$

At last consider the sub-Gaussian case:

$$\sum_{n_0+1}^{\infty} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) = \exp\left(-\frac{n_0\varepsilon^2}{2\sigma^2}\right) \sum_{n=1}^{\infty} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) = \frac{\exp\left(-\frac{n_0\varepsilon^2}{2\sigma^2}\right)}{1-\exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)}$$

from which we obtain:

$$\widetilde{n}_{\frac{\delta}{2}} \leq \left\lceil -\frac{2\sigma^2}{\varepsilon^2} \ln\left(\frac{\delta}{4}\left(1-\mathrm{e}^{-\frac{\varepsilon^2}{2\sigma^2}}\right)\right) \right\rceil$$

□

We observe that $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)$ is smaller than the corresponding term $\beta_{\frac{\delta}{2}}^{-1}(\varepsilon)$ appearing in Theorem 8 for ColME. Our proofs can readily be adapted to ColME, enabling to substitute $n_{\frac{\delta}{2}}^{\star}(\varepsilon)$ in (Asadi et al. 2022) [Theorem 2] with the smaller term $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)$.

# F Appendix F - Proofs for C-ColME - Theorem 2 and Theorem 3

It is convenient to consider an auxiliary system for which the consensus matrix can be arbitrary until time $\zeta_d$ and then switches to a situation where agents only communicate with their neighbours belonging to the same class, i.e., $W_t = W$ for any $t \geqslant \zeta_d$ and $W_{a,a'} > 0$ if and only if $a' \in \mathcal{C}_a \cap \mathcal{N}_a$.

We will derive some bounds for the auxiliary system for any choice of the matrices $W_1, W_2, \dots W_{\gamma_D}$. With probability $1 - \frac{\delta}{2}$ these bounds apply also to the original system under study, because with such probability each agent correctly detects the neighbours in the same class for any $t > \zeta_D$. From now on, we refer then to the auxiliary system.

## F.1 Preliminaries

After time $\zeta_D$ the original graph has been then split in $C$ connected components, where component $c$ includes $n_c$ agents. By an opportune permutation of the agents, we can write the matrix $W$ as follows

$$W = \begin{pmatrix} {}_1W & 0_{n_1 \times n_2} & \cdots & 0_{n_1 \times n_C} \\ 0_{n_2 \times n_1} & {}_2W & \cdots & 0_{n_2 \times n_C} \\ \cdots & \cdots & \cdots & \cdots \\ 0_{n_C \times n_1} & 0_{n_C \times n_2} & \cdots & {}_CW \end{pmatrix},$$

where $0_{n \times m}$ denotes an $n \times m$ matrix with 0 elements.

We focus on a given component $c$ with $n_c$ agents. All agents in the same component share the same expected value, which we denote by $\mu(c)$. Moreover, let ${}_c\boldsymbol{\mu} = \mu(c)\mathbf{1}_c$. We denote by ${}_c\mathbf{x}^t$ and ${}_c\hat{\boldsymbol{\mu}}^t$ the $n_c$-dimensional vectors containing the samples' empirical averages and the estimates for the agents in component $c$.

For $t > \zeta_D$, the estimates in component $c$ evolves independently from the other components and we can write:

$${}_c\hat{\boldsymbol{\mu}}^{t+1} = (1 - \alpha_t)\, {}_c\bar{\mathbf{x}}^{t+1} + \alpha_t\, {}_cW\hat{\boldsymbol{\mu}}^t(c).$$

It is then easy to prove by recurrence that

$${}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu} = \alpha_{\zeta_D,t}\, {}_cW^{t+1-\zeta_D}(\hat{\boldsymbol{\mu}}^{\zeta_D}(c) - {}_c\boldsymbol{\mu}) + \sum_{\tau=\zeta_D}^{t} (1-\alpha_\tau)\alpha_{\tau+1,t}{}_cW^{t-\tau}({}_c\bar{\mathbf{x}}^{\tau+1} - {}_c\boldsymbol{\mu}), \tag{24}$$

where $\alpha_{i,j} \triangleq \prod_{\ell=i}^{j} \alpha_\ell$, with the usual convention that $\alpha_{i,j} = 1$ if $j < i$.

Let ${}_cP \triangleq \frac{1}{n_c}\mathbf{1}\mathbf{1}^\top$. It is easy to check that the doubly-stochasticity of $W$ implies that $({}_cW - {}_cP)^t = {}_cW^t - P$. From which it follows

$$\begin{aligned}
{}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu} &= \alpha_{\zeta_D,t}\, {}_cW^{t+1-\zeta_D}(\hat{\boldsymbol{\mu}}^{\zeta_D}(c) - {}_c\boldsymbol{\mu}) + \sum_{\tau=\zeta_D}^{t} (1-\alpha_\tau)\alpha_{\tau+1,t}{}_cW^{t-\tau}({}_c\bar{\mathbf{x}}^{\tau+1} - {}_c\boldsymbol{\mu}) && (25)\\
&= \alpha_{\zeta_D,t}\, {}_cW^{t+1-\zeta_D}(\hat{\boldsymbol{\mu}}^{\zeta_D}(c) - {}_c\boldsymbol{\mu}) \\
&\quad + \sum_{\tau=\zeta_D}^{t} (1-\alpha_\tau)\alpha_{\tau+1,t}{}_cW^{t-\tau}({}_c\bar{\mathbf{x}}^{\tau+1} - {}_cP\,{}_c\bar{\mathbf{x}}^{\tau+1}) \\
&\quad + \sum_{\tau=\zeta_D}^{t} (1-\alpha_\tau)\alpha_{\tau+1,t}({}_cP\,{}_c\bar{\mathbf{x}}^{\tau+1} - {}_c\boldsymbol{\mu}) && (26)\\
&= \alpha_{\zeta_D,t}\, {}_cW^{t+1-\zeta_D}(\hat{\boldsymbol{\mu}}^{\zeta_D}(c) - {}_c\boldsymbol{\mu}) \\
&\quad + \sum_{\tau=\zeta_D}^{t} (1-\alpha_\tau)\alpha_{\tau+1,t}{}_cW^{t-\tau}({}_c\bar{\mathbf{x}}^{\tau+1} - {}_cP\,{}_c\bar{\mathbf{x}}^{\tau+1}) \\
&\quad + \sum_{\tau=\zeta_D}^{t} (1-\alpha_\tau)\alpha_{\tau+1,t}({}_cP\,{}_c\bar{\mathbf{x}}^{\tau+1} - {}_c\boldsymbol{\mu}) && (27)\\
&= \alpha_{\zeta_D,t}\, {}_cW^{t+1-\zeta_D}(\hat{\boldsymbol{\mu}}^{\zeta_D}(c) - {}_c\boldsymbol{\mu}) \\
&\quad + \sum_{\tau=\zeta_D}^{t} (1-\alpha_\tau)\alpha_{\tau+1,t}({}_cW - {}_cP)^{t-\tau}({}_c\bar{\mathbf{x}}^{\tau+1} - {}_cP\,{}_c\bar{\mathbf{x}}^{\tau+1}) \\
&\quad + \sum_{\tau=\zeta_D}^{t} (1-\alpha_\tau)\alpha_{\tau+1,t}({}_cP\,{}_c\bar{\mathbf{x}}^{\tau+1} - {}_c\boldsymbol{\mu}). && (28)
\end{aligned}$$

## F.2 Technical Results

**Lemma 15.** *For* $0 < \beta < 1$

$$\sum_{\tau=t_0}^{t} \frac{\beta^{t-\tau}}{\tau+1} \in \mathcal{O}\left(\left(1+\frac{1}{\ln\frac{1}{\beta}}\right)\frac{1}{t+1}\right) \tag{29}$$

*Proof.*

$$\sum_{\tau=t_0}^{t} \frac{\beta^{t-\tau}}{\tau+1} = \beta^{t+1} \sum_{\tau=t_0+1}^{t+1} \frac{\beta^{-\tau}}{\tau} \tag{30}$$

Let $t' = \max\left\{\left\lceil \frac{1}{\ln\frac{1}{\beta}} \right\rceil, t_0+1\right\}$. For $\tau_2 \geqslant \tau_1 \geqslant t'$, $\frac{\beta^{-\tau_1}}{\tau_1} \leqslant \frac{\beta^{-\tau_2}}{\tau_2}$.

$$\begin{aligned}
\sum_{\tau=t_0+1}^{t+1} \frac{\beta^{-\tau}}{\tau} &= \underbrace{\sum_{\tau=t_0}^{t'-1} \frac{\beta^{-\tau}}{\tau}}_{C} + \sum_{\tau=t'}^{t} \frac{\beta^{-\tau}}{\tau} + \frac{\beta^{-t-1}}{t+1} && (31)\\
&\leqslant C + \frac{\beta^{-t-1}}{t+1} + \int_{t'}^{t+1} \frac{\beta^{-\tau}}{\tau}\,\mathrm{d}\tau && (32)\\
&= C + \frac{\beta^{-t-1}}{t+1} + \int_{t'}^{t+1} \frac{e^{\ln\frac{1}{\beta}\tau}}{\tau}\,\mathrm{d}\tau && (33)\\
&= C + \frac{\beta^{-t-1}}{t+1} + \int_{t'\ln\frac{1}{\beta}}^{(t+1)\ln\frac{1}{\beta}} \frac{e^{x}}{x}\,\mathrm{d}x && (34)\\
&\leqslant C + \frac{\beta^{-t-1}}{t+1} + \mathrm{Ei}\left((t+1)\ln\frac{1}{\beta}\right) - \mathrm{Ei}\left(t'\ln\frac{1}{\beta}\right) && (35)\\
&\leqslant C + \frac{\beta^{-t-1}}{t+1} + \mathrm{Ei}\left((t+1)\ln\frac{1}{\beta}\right) && (36)\\
&\leqslant C + \frac{\beta^{-t-1}}{t+1} + \frac{e^{(t+1)\ln\frac{1}{\beta}}}{(t+1)\ln\frac{1}{\beta}}\left(1+\frac{3}{(t+1)\ln\frac{1}{\beta}}\right) && (37)\\
&= C + \left(1+\frac{1}{\ln\frac{1}{\beta}} + \frac{3}{(t+1)\ln\frac{1}{\beta}}\right)\frac{\beta^{-t-1}}{t+1}, && (38)
\end{aligned}$$

where $\mathrm{Ei}(t) \triangleq \int_{-\infty}^{t} \frac{e^x}{x}\,\mathrm{d}x$ is the exponential integral. The third inequality follows from the series representation

$$\mathrm{Ei}(t) = \frac{e^t}{t}\left(\sum_{k=0}^{n} \frac{k!}{t^k} + e_n(t)\right),$$

where $e_n(t) \triangleq (n+1)!te^{-t}\int_{-\infty}^{t} \frac{e^x}{x^{n+2}}\,\mathrm{d}x$ for $n = 0$. The remainder $e_n(t)$ can be bounded by the $n+1$-th term times the factor $1 + \sqrt{\pi}\frac{\Gamma(n/2+3/2)}{\Gamma(n/2+1)}$ (of Mathematical Functions 2023). This factor is smaller than 3 for $n = 0$.

Finally, from (30) and (38), we obtain

$$\sum_{\tau=t_0}^{t} \frac{\beta^{t-\tau}}{\tau+1} \leqslant \beta^{t+1}C + \left(1+\frac{1}{\ln\frac{1}{\beta}} + \frac{3}{(t+1)\ln\frac{1}{\beta}}\right)\frac{1}{t+1} \in \mathcal{O}\left(\left(1+\frac{1}{\ln\frac{1}{\beta}}\right)\frac{1}{t+1}\right). \tag{39}$$

□

**Lemma 16.** *Let* $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t, \ldots)$ *be a sequence of i.i.d. vectorial random variables in* $\mathbb{R}^n$ *with expected value* $\mathbf{0}$ *and finite* 4*-th moment* $\mathbb{E}[\|z_i\|^4]$*, and* $\{A(t_1,t_2), (t_1,t_2) \in \mathbb{N}^2\}$ *a set of* $n \times n$ *matrices with bounded norms. Let* $\mathbf{b}_\tau \triangleq \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbf{z}_t$*. It holds:*

$$\mathbb{E}\left[\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\mathbf{b}_{\tau_1}^\top A(\tau_1,\tau_2)\mathbf{b}_{\tau_2}\mathbf{b}_{\tau_3}^\top A(\tau_3,\tau_4)\mathbf{b}_{\tau_4}\right] \leqslant 2\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\left(\frac{\|A(\tau_1,\tau_2)\|\cdot\|A(\tau_3,\tau_4)\|}{\tau_2\tau_3} + \frac{\|A(\tau_1,\tau_2)\|\cdot\|A(\tau_3,\tau_4)\|}{\tau_2\tau_4}\right) \tag{40}$$

*Proof.* We will omit the indices when they run from 1 to $t$ and denote by $\mathbf{z}$ and $\mathbf{z}'$ two generic indipendent random variables distributed as $\mathbf{z}_\tau$.

$$\mathbb{E}\left[\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\mathbf{b}_{\tau_1}^\top A(\tau_1,\tau_2)\mathbf{b}_{\tau_2}\mathbf{b}_{\tau_3}^\top A(\tau_3,\tau_4)\mathbf{b}_{\tau_4}\right] \tag{41}$$

$$= \sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1=1}^{\tau_1}\sum_{t_2=1}^{\tau_2}\sum_{t_3=1}^{\tau_3}\sum_{t_4=1}^{\tau_4}\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_2}\mathbf{z}_{t_3}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_4}\right] \tag{42}$$

$$= \sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1,t_2,t_3,t_4}\mathbf{1}_{t_1\leqslant\tau_1}\mathbf{1}_{t_2\leqslant\tau_2}\mathbf{1}_{t_3\leqslant\tau_3}\mathbf{1}_{t_4\leqslant\tau_4}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_2}\mathbf{z}_{t_3}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_4}\right] \tag{43}$$

$$\begin{aligned}= \sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1,t_2,t_3,t_4}\mathbf{1}_{t_1\leqslant\tau_1,t_2\leqslant\tau_2,t_3\leqslant\tau_3,t_4\leqslant\tau_4}\times\Big(&\mathbf{1}_{t_1=t_2=t_3=t_4}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_2}\mathbf{z}_{t_3}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_4}\right]\\ &+\mathbf{1}_{t_1=t_2,t_3=t_4,t_1\neq t_3}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_2}\mathbf{z}_{t_3}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_4}\right]\\ &+\mathbf{1}_{t_1=t_3,t_2=t_4,t_1\neq t_2}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_2}\mathbf{z}_{t_3}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_4}\right]\\ &+\mathbf{1}_{t_1=t_4,t_2=t_3,t_1\neq t_2}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_2}\mathbf{z}_{t_3}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_4}\right]\Big)\end{aligned} \tag{44}$$

$$\begin{aligned}=&\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1}\mathbf{1}_{t_1\leqslant\tau_1,t_1\leqslant\tau_2,t_1\leqslant\tau_3,t_1\leqslant\tau_4}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_1}\mathbf{z}_{t_1}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_1}\right]\\ &+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1,t_3}\mathbf{1}_{t_1\leqslant\tau_1,t_1\leqslant\tau_2,t_3\leqslant\tau_3,t_3\leqslant\tau_4}\mathbf{1}_{t_1\neq t_3}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_1}\mathbf{z}_{t_3}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_3}\right]\\ &+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1,t_2}\mathbf{1}_{t_1\leqslant\tau_1,t_2\leqslant\tau_2,t_1\leqslant\tau_3,t_2\leqslant\tau_4}\mathbf{1}_{t_1\neq t_2}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_2}\mathbf{z}_{t_1}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_2}\right]\\ &+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1,t_2}\mathbf{1}_{t_1\leqslant\tau_1,t_2\leqslant\tau_2,t_2\leqslant\tau_3,t_1\leqslant\tau_4}\mathbf{1}_{t_1\neq t_2}\,\mathbb{E}\left[\mathbf{z}_{t_1}^\top A(\tau_1,\tau_2)\mathbf{z}_{t_2}\mathbf{z}_{t_2}^\top A(\tau_3,\tau_4)\mathbf{z}_{t_1}\right]\end{aligned} \tag{45}$$

$$\begin{aligned}=&\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1}\mathbf{1}_{t_1\leqslant\tau_1,t_1\leqslant\tau_2,t_1\leqslant\tau_3,t_1\leqslant\tau_4}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z}\mathbf{z}^\top A(\tau_3,\tau_4)\mathbf{z}\right]\\ &+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1,t_3}\mathbf{1}_{t_1\leqslant\tau_1,t_1\leqslant\tau_2,t_3\leqslant\tau_3,t_3\leqslant\tau_4}\mathbf{1}_{t_1\neq t_3}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z}\right]\mathbb{E}\left[\mathbf{z}'^\top A(\tau_3,\tau_4)\mathbf{z}'\right]\\ &+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1,t_2}\mathbf{1}_{t_1\leqslant\tau_1,t_2\leqslant\tau_2,t_1\leqslant\tau_3,t_2\leqslant\tau_4}\mathbf{1}_{t_1\neq t_2}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z}'\mathbf{z}^\top A(\tau_3,\tau_4)\mathbf{z}'\right]\\ &+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_1\tau_2\tau_3\tau_4}\sum_{t_1,t_2}\mathbf{1}_{t_1\leqslant\tau_1,t_2\leqslant\tau_2,t_2\leqslant\tau_3,t_1\leqslant\tau_4}\mathbf{1}_{t_1\neq t_2}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z}'\mathbf{z}'^\top A(\tau_3,\tau_4)\mathbf{z}\right]\end{aligned} \tag{46}$$

$$= \sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{\min\{\tau_1,\tau_2,\tau_3,\tau_4\}}{\tau_1\tau_2\tau_3\tau_4}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z}\mathbf{z}^\top A(\tau_3,\tau_4)\mathbf{z}\right]$$

$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{\min\{\tau_1,\tau_2\}(\min\{\tau_3,\tau_4\}-1)}{\tau_1\tau_2\tau_3\tau_4}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z}\right]\mathbb{E}\left[\mathbf{z'}^\top A(\tau_3,\tau_4)\mathbf{z'}\right]$$
$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{\min\{\tau_1,\tau_3\}(\min\{\tau_2,\tau_4\}-1)}{\tau_1\tau_2\tau_3\tau_4}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z'}\mathbf{z}^\top A(\tau_3,\tau_4)\mathbf{z'}\right]$$
$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{\min\{\tau_1,\tau_4\}(\min\{\tau_2,\tau_3\}-1)}{\tau_1\tau_2\tau_3\tau_4}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z'}(\mathbf{z'})^\top A(\tau_3,\tau_4)\mathbf{z}\right], \tag{47}$$

where (44) follows from the independence of the variables $\{\mathbf{z}_t\}_{t\in\mathbb{N}}$ and the fact that $\mathbb{E}[\mathbf{z}_t]=\mathbf{0}$ for any $t$.

Now we can upperbound the three terms in (47).

$$\mathbb{E}\left[\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\mathbf{b}_{\tau_1}^\top A(\tau_1,\tau_2)\mathbf{b}_{\tau_2}\mathbf{b}_{\tau_3}^\top A(\tau_3,\tau_4)\mathbf{b}_{\tau_4}\right] \tag{48}$$
$$\leqslant\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_2\tau_3\tau_4}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z}\mathbf{z}^\top A(\tau_3,\tau_4)\mathbf{z}\right]$$
$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_2\tau_3}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z}\right]\mathbb{E}\left[\mathbf{z'}^\top A(\tau_3,\tau_4)\mathbf{z'}\right]$$
$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_2\tau_3}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z'}\mathbf{z}^\top A(\tau_3,\tau_4)\mathbf{z'}\right]$$
$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_2\tau_4}\,\mathbb{E}\left[\mathbf{z}^\top A(\tau_1,\tau_2)\mathbf{z'}\mathbf{z'}^\top A(\tau_3,\tau_4)\mathbf{z}\right] \tag{49}$$
$$\leqslant\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_2\tau_3\tau_4}\|A(\tau_1,\tau_2)\|\cdot\|A(\tau_3,\tau_4)\|\cdot\mathbb{E}\left[\|\mathbf{z}\|^4\right]$$
$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_2\tau_3}\|A(\tau_1,\tau_2)\|\cdot\|A(\tau_3,\tau_4)\|\cdot\mathbb{E}\left[\|\mathbf{z}\|^2\right]^2$$
$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_2\tau_3}\|A(\tau_1,\tau_2)\|\cdot\|A(\tau_3,\tau_4)\|\cdot\mathbb{E}\left[\|\mathbf{z}\|^2\cdot\|\mathbf{z'}\|^2\right]$$
$$+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\frac{1}{\tau_2\tau_4}\|A(\tau_1,\tau_3)\|\cdot\|A(\tau_3,\tau_4)\|\cdot\mathbb{E}\left[\|\mathbf{z}\|^2\cdot\|\mathbf{z'}\|^2\right] \tag{50}$$
$$\leqslant 2\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\left(\frac{\|A(\tau_1,\tau_2)\|\cdot\|A(\tau_3,\tau_4)\|}{\tau_2\tau_3}+\frac{\|A(\tau_1,\tau_2)\|\cdot\|A(\tau_3,\tau_4)\|}{\tau_2\tau_4}\right).$$

□

We now particularize the result of Lemma 16 to the two cases of interest for what follows.

**Corollary 17.** *Let $(\mathbf{z}_1,\mathbf{z}_2,\dots,\mathbf{z}_t,\dots)$ be a sequence of i.i.d. vectorial random variables in $\mathbb{R}^n$ with expected value $\mathbf{0}$ and finite 4-th moment $\mathbb{E}[\|z_i\|^4]$ and $\{A(t_1,t_2),(t_1,t_2)\in\mathbb{N}^2\}$ a set of $n\times n$ symmetric stochastic matrices. Let $\mathbf{b}_\tau\triangleq\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbf{z}_t$. It holds:*

$$\mathbb{E}\left[\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t}\mathbf{b}_{\tau_1}^\top A(\tau_1,\tau_2)\mathbf{b}_{\tau_2}\mathbf{b}_{\tau_3}^\top A(\tau_3,\tau_4)\mathbf{b}_{\tau_4}\right]\leqslant 4\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]t^2(1+\ln t)^2.$$

*Proof.* It is sufficient to observe that $\|A(\tau_1,\tau_2)\|=1$ and that $\sum_{\tau=1}^{t}\frac{1}{\tau}\leqslant 1+\ln t$. □

**Corollary 18.** *Let $(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_t, \dots)$ be a sequence of i.i.d. vectorial random variables in $\mathbb{R}^n$ with expected value $\mathbf{0}$ and finite 4-th moment $\mathbb{E}[\|z_i\|^4]$ and $\{A(t_1,t_2), (t_1,t_2) \in \mathbb{N}^2\} = \beta^{2t-t_1-t_2} B^{2t-t_1-t_2}$, where $\beta \in [0,1]$, $B$ a symmetric matrix. Let $\rho(B)$ denote the spectral norm of $B$ and $\mathbf{b}_\tau \triangleq \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbf{z}_t$. If $\beta\rho(B) < 1$, then it holds:*

$$\mathbb{E}\left[\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t} \mathbf{b}_{\tau_1}^\top A(\tau_1,\tau_2)\mathbf{b}_{\tau_2}\mathbf{b}_{\tau_3}^\top A(\tau_3,\tau_4)\mathbf{b}_{\tau_4}\right] \in \mathcal{O}\left(\mathbb{E}\left[\|\mathbf{z}\|^4\right]\frac{1}{(1-\beta\rho(B))^2}\left(1+\frac{1}{\ln\frac{1}{\beta\rho(B)}}\right)^2\frac{1}{(t+1)^2}\right).$$

*Proof.* As $B$ is symmetric, then $\|B\| = \rho(B)$. From Lemma 16, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau_1,\tau_2,\tau_3,\tau_4=t_0}^{t} \mathbf{b}_{\tau_1}^\top A(\tau_1,\tau_2)\mathbf{b}_{\tau_2}\mathbf{b}_{\tau_3}^\top A(\tau_3,\tau_4)\mathbf{b}_{\tau_4}\right] \\
&\leqslant 2\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\left(\sum_{\tau_1,\tau_2,\tau_3,\tau_4=1}^{t}\frac{(\beta\lambda_2(B))^{4t-\tau_1-\tau_2-\tau_3-\tau_4}}{\tau_2\tau_3} + \sum_{\tau_1,\tau_2,\tau_3,\tau_4=1}^{t}\frac{(\beta\lambda_2(B))^{4t-\tau_1-\tau_2-\tau_3-\tau_4}}{\tau_2\tau_4}\right) &(51)\\
&= 4\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\sum_{\tau_1,\tau_2,\tau_3,\tau_4=1}^{t}\frac{(\beta\lambda_2(B))^{4t-\tau_1-\tau_2-\tau_3-\tau_4}}{\tau_2\tau_3} &(52)\\
&= 4\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\left(\sum_{\tau_1=1}^{t}(\beta\lambda_2(B))^{t-\tau_1}\right)^2\left(\sum_{\tau_2=1}^{t}\frac{(\beta\lambda_2(B))^{t-\tau_2}}{\tau_2}\right)^2 &(53)\\
&\leq 4\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\frac{1}{(1-\beta\lambda_2(B))^2}\left(\sum_{\tau_2=1}^{t}\frac{(\beta\lambda_2(B))^{t-\tau_2}}{\tau_2}\right)^2 &(54)\\
&\in \mathcal{O}\left(\mathbb{E}\left[\|\mathbf{z}\|^4\right]\frac{1}{(1-\beta\lambda_2(B))^2}\left(1+\frac{1}{\ln\frac{1}{\beta\lambda_2(B)}}\right)^2\frac{1}{(t+1)^2}\right), &(55)
\end{aligned}$$

where in the last step we used Lemma 15. □

**Lemma 19.** *Let $W$ be an $n \times n$ symmetric, stochastic, and irreducible matrix and $P = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, where $\mathbf{1}_n$ is an $n$-dimensional vector whose elements are all equal to 1, then $\rho(W-P) = \lambda_2(W) < 1$.*

*Proof.* As $W$ is symmetric and stochastic, then the module of its largest eigenvalue is equal to 1. The vector $\mathbf{1}_n$ is both a left and right eigenvector of $W$ relative to the simple eigenvalue 1. Then, $P$ is the projector onto the null space of $W - I$ along the range of $W - I$ (Meyer 2023)[p.518]. The spectral theorem leads us to conclude that the eigenvalues of $W - P$ (counted with their multiplicity) are then all eigenvalues of $W$ except 1 and with the addition of 0. We can then conclude that $\|W - P\| = \rho(W - P) = \lambda_2(W)$. $W$ is irreducible with non-negative elements on the diagonal, then it is primitive (Meyer 2023)[Example 8.3.3], i.e., 1 is the only eigenvalue on the unit circle.

□

**Lemma 20.** *For any $a > 0$, if $x \geq \max\{a\ln^2 a, 1\}$ then $x \geqslant \frac{a}{4}\ln^2 x$.*

*Proof.* We consider first $a \geq 1$.

$$\begin{aligned}
x \geqslant a\ln^2 a \implies \sqrt{x} \geqslant \sqrt{a}|\ln a| \underset{a\geq 1}{=} \sqrt{a}\ln a &(56)\\
\implies \sqrt{x} \geqslant \sqrt{a}\ln\sqrt{x} = \frac{\sqrt{a}}{2}\ln x &(57)\\
\implies x \geqslant \frac{a}{4}\ln^2 x, &(58)
\end{aligned}$$

where (57) follows from Lemma A.1 in (Shalev-Shwartz and Ben-David 2014). For $a < 1$, it is easy to check that for $x \geqslant 1$, $x \geqslant \frac{a}{4}\ln^2 x$ holds unconditionally. □

### F.3 Bounding the 4-th Moment of the Estimation Error.

For convenience, we omit from now on the dependence on the specific clustered component $c$.

**Theorem 21.** *Let $\lambda_2(W)$ denote the module of the second largest eigenvalue in module of $W$. It holds:*

$$\begin{aligned}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{t+1}-\boldsymbol{\mu}\|^4\right] \in \mathcal{O}&\left(\sup_{W_1,W_2,\cdots,W_{\zeta_D}} \mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]\alpha^{4t}\right)\\ &+\mathcal{O}\left(\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}\|^4\right]\frac{(1-\alpha)^4}{(1-\alpha\lambda_2(W))^2}\left(1+\frac{1}{\ln\frac{1}{\alpha\lambda_2(W)}}\right)^2\frac{1}{(t+1)^2}\right)\\ &+\mathcal{O}\left(\mathbb{E}\left[\|P\mathbf{x}-\boldsymbol{\mu}\|^4\right](1-\alpha)^2\left(1+\frac{1}{\ln\frac{1}{\alpha}}\right)^2\frac{1}{(t+1)^2}\right), \qquad \text{if } \alpha_t=\alpha, \end{aligned} \tag{59}$$

$$\begin{aligned}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{t+1}-\boldsymbol{\mu}\|^4\right] \in \mathcal{O}&\left(\sup_{W_1,W_2,\cdots,W_{\zeta_D}} \mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]\frac{1}{(t+1)^4}\right)\\ &+\mathcal{O}\left(\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}\|^4\right]\frac{1}{(1-\lambda_2(W))^2}\left(1+\frac{1}{\ln\frac{1}{\lambda_2(W)}}\right)^2\frac{1}{(t+1)^4}\right)\\ &+\mathcal{O}\left(\mathbb{E}\left[\|P\mathbf{x}-\boldsymbol{\mu}\|^4\right]\left(\frac{1+\ln t}{1+t}\right)^2\right), \qquad \text{if } \alpha_t=\frac{t}{t+1}. \end{aligned} \tag{60}$$

*Proof.* $W$ is irreducible (the graph component is connected and $W_{i,j}>0$ for each link), then by Lemma 19, $\lambda_2(W)<1$. For $\alpha_t=\alpha<1$, it would be sufficient to observe that $\lambda_2(W)\leqslant\rho(W)=1$, but for $\alpha_t=\frac{t}{t+1}$, we need the strict inequality.

Our starting point is (28), which we repeat here (omitting the dependence on the specific clustered component $c$):

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{t+1}-\boldsymbol{\mu} = &\alpha_{\zeta_D,t}W^{t+1-\zeta_D}(\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu})\\ &+\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(W-P)^{t-\tau}(\bar{\mathbf{x}}^{\tau+1}-P\bar{\mathbf{x}}^{\tau+1})\\ &+\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(P\bar{\mathbf{x}}^{\tau+1}-\boldsymbol{\mu}).\end{aligned} \tag{61}$$

Applying twice $(\sum_{i=1}^n a_i)^2\leq n\sum_{i=1}^n a_i^2$ with $n=3$, and applying the expectation we obtain:

$$\begin{aligned}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{t+1}-\boldsymbol{\mu}\|^4\right] &\leq 27\cdot\alpha_{\zeta_D,t}^4\cdot\mathbb{E}\left[\|W\|^{4(t+1-\zeta_D)}\cdot\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]\\ &\quad+27\,\mathbb{E}\left[\left\|\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(W-P)^{t-\tau}(\bar{\mathbf{x}}^{\tau+1}-P\bar{\mathbf{x}}^{\tau+1})\right\|^4\right]\\ &\quad+27\,\mathbb{E}\left[\left\|\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(P\bar{\mathbf{x}}^{\tau+1}-\boldsymbol{\mu})\right\|^4\right]\\ &\leq 27\cdot\underbrace{\alpha_{\zeta_D,t}^4\cdot\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]}_{C_1}\\ &\quad+27\,\underbrace{\mathbb{E}\left[\left\|\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(W-P)^{t-\tau}(\bar{\mathbf{x}}^{\tau+1}-P\bar{\mathbf{x}}^{\tau+1})\right\|^4\right]}_{C_2}\end{aligned} \tag{62}$$

$$+ 27\,\underbrace{\mathbb{E}\left[\left\|\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(P\bar{\mathbf{x}}^{\tau+1}-\boldsymbol{\mu})\right\|^4\right]}_{C_3}, \tag{63}$$

where in the last step we took advantage of the fact that $W$ is doubly stochastic and symmetric and then $\|W\| = 1$.

We now move to bound the three terms $C_1$, $C_2$, and $C_3$. We observe that

$$\alpha_{t_0+1,t} = \begin{cases} \alpha^{t-t_0}, & \text{if } \alpha_t = \alpha, \\ \frac{t_0+1}{t} & \text{if } \alpha_t = \frac{t}{t+1}, \end{cases} \tag{64}$$

and

$$(1-\alpha_{t_0})\alpha_{t_0+1,t} = \begin{cases} (1-\alpha)\alpha^{t-t_0}, & \text{if } \alpha_t = \alpha, \\ \frac{1}{t} & \text{if } \alpha_t = \frac{t}{t+1}, \end{cases} \tag{65}$$

$$C_1 \leqslant \alpha_{\zeta_D,t}^4 \sup_{W_1,W_2,\cdots,W_{\zeta_D}} \mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right] \in \begin{cases} \mathcal{O}\left(\sup_{W_1,W_2,\cdots,W_{\zeta_D}} \mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]\alpha^{4t}\right), & \text{if } \alpha_t = \alpha, \\ \mathcal{O}\left(\sup_{W_1,W_2,\cdots,W_{\zeta_D}} \mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]\frac{1}{t^4}\right) & \text{if } \alpha_t = \frac{t}{t+1}. \end{cases} \tag{66}$$

$$\begin{aligned} C_2 = \mathbb{E}\Bigg[&\left(\sum_{\tau_1=\zeta_D}^{t}(1-\alpha_{\tau_1})\alpha_{\tau_1+1,t}(W-P)^{t-\tau_1}(\bar{\mathbf{x}}^{\tau_1+1}-P\bar{\mathbf{x}}^{\tau_1+1})\right)^\top \\ &\left(\sum_{\tau_2=\zeta_D}^{t}(1-\alpha_{\tau_2})\alpha_{\tau_2+1,t}(W-P)^{t-\tau_2}(\bar{\mathbf{x}}^{\tau_2+1}-P\bar{\mathbf{x}}^{\tau_2+1})\right) \\ &\left(\sum_{\tau_3=\zeta_D}^{t}(1-\alpha_{\tau_3})\alpha_{\tau_3+1,t}(W-P)^{t-\tau_3}(\bar{\mathbf{x}}^{\tau_3+1}-P\bar{\mathbf{x}}^{\tau_3+1})\right)^\top \\ &\left(\sum_{\tau_4=\zeta_D}^{t}(1-\alpha_{\tau_4})\alpha_{\tau_4+1,t}(W-P)^{t-\tau_4}(\bar{\mathbf{x}}^{\tau_4+1}-P\bar{\mathbf{x}}^{\tau_4+1})\right)\Bigg] \end{aligned} \tag{67}$$

$$\begin{aligned} = \mathbb{E}\Bigg[&\sum_{\tau_1,\tau_2,\tau_3,\tau_4=\zeta_D}^{t}(1-\alpha_{\tau_1})\alpha_{\tau_1+1,t}(1-\alpha_{\tau_2})\alpha_{\tau_2+1,t}\,(1-\alpha_{\tau_3})\alpha_{\tau_3+1,t}(1-\alpha_{\tau_4})\alpha_{\tau_4+1,t} \\ &(\bar{\mathbf{x}}^{\tau_1+1}-P\bar{\mathbf{x}}^{\tau_1+1})^\top(W-P)^{2t-\tau_1-\tau_2}(\bar{\mathbf{x}}^{\tau_1+1}-P\bar{\mathbf{x}}^{\tau_1+1}) \\ &(\bar{\mathbf{x}}^{\tau_3+1}-P\bar{\mathbf{x}}^{\tau_3+1})^\top(W-P)^{2t-\tau_3-\tau_4}(\bar{\mathbf{x}}^{\tau_4+1}-P\bar{\mathbf{x}}^{\tau_4+1})\Bigg] \\ \in &\begin{cases} \mathcal{O}\left(\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}\|^4\right]\frac{(1-\alpha)^4}{(1-\alpha\lambda_2(W))^2}\left(1+\frac{1}{\ln\frac{1}{\alpha\lambda_2(W)}}\right)^2\frac{1}{(t+1)^2}\right), & \text{if } \alpha_t = \alpha, \\ \mathcal{O}\left(\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}\|^4\right]\frac{1}{(1-\lambda_2(W))^2}\left(1+\frac{1}{\ln\frac{1}{\lambda_2(W)}}\right)^2\frac{1}{(t+1)^4}\right), & \text{if } \alpha_t = \frac{t}{t+1}, \end{cases} \end{aligned} \tag{68}$$

where the last result follows from observing that $\rho(W-P) = \lambda_2(W) < 1$ and $\|\mathbf{x}-P\mathbf{x}\|^4 \leq \|\mathbf{x}-\boldsymbol{\mu}\|^4$ and then applying Corollary 18 with $\mathbf{z} = \mathbf{x} - P\mathbf{x}$, $B = W - P$ and 1) $\beta = \alpha$ for $\alpha_t = \alpha$, 2) $\beta = 1$ for $\alpha_t = \frac{t}{t+1}$.

The calculations to bound $C_3$ are similar:

$$C_3 = \mathbb{E}\left[\sum_{\tau_1,\tau_2,\tau_3,\tau_4=\zeta_D}^{t} (1-\alpha_{\tau_1})\alpha_{\tau_1+1,t}(1-\alpha_{\tau_2})\alpha_{\tau_2+1,t}\,(1-\alpha_{\tau_3})\alpha_{\tau_3+1,t}(1-\alpha_{\tau_4})\alpha_{\tau_4+1,t}\right.$$
$$\left.(P\bar{\mathbf{x}}^{\tau_1+1}-\boldsymbol{\mu})^\top(P\bar{\mathbf{x}}^{\tau_1+1}-\boldsymbol{\mu})\qquad(P\bar{\mathbf{x}}^{\tau_3+1}-\boldsymbol{\mu})^\top(P\bar{\mathbf{x}}^{\tau_4+1}-\boldsymbol{\mu})\right]$$
$$\in \begin{cases}\mathcal{O}\left(\mathbb{E}\left[\|P\mathbf{x}-\boldsymbol{\mu}\|^4\right](1-\alpha)^2\left(1+\frac{1}{\ln\frac{1}{\alpha}}\right)^2\frac{1}{(t+1)^2}\right), & \text{if } \alpha_t=\alpha,\\ \mathcal{O}\left(\mathbb{E}\left[\|P\mathbf{x}-\boldsymbol{\mu}\|^4\right]\left(\frac{1+\ln t}{1+t}\right)^2\right), & \text{if } \alpha_t=\frac{t}{t+1}.\end{cases} \tag{69}$$

In this case, we apply 1) Corollary 18 with $\mathbf{z}=\mathbf{x}-P\mathbf{x}$, $\beta=\alpha$, and $B=I$, for $\alpha_t=\alpha$, and 2) Corollary 17 with $\mathbf{z}=\mathbf{x}-P\mathbf{x}$ and $A(t_1,t_2)=I,\forall(t_1,t_2)\in\mathbb{N}^2$, for $\alpha_t=\frac{t}{t+1}$.

The result follows by simply aggregating the three bounds.

□

**Remark 2.** *We observe that*

$$\mathbb{E}\left[\|_c\mathbf{x}-{}_c\boldsymbol{\mu}\|^4\right] = n_c\kappa\sigma^4+n_c(n_c-1)\sigma^4, \tag{70}$$
$$\mathbb{E}\left[\|_c\mathbf{x}-{}_cP\,{}_c\mathbf{x}\|^4\right] = \left(n_c-2+\frac{1}{n_c}\right)\kappa\sigma^4+(n_c-1)\left(n_c-2+\frac{3}{n_c}\right)\sigma^4, \tag{71}$$
$$\mathbb{E}\left[\|_cP\,{}_c\mathbf{x}-{}_c\boldsymbol{\mu}\|^4\right] = \frac{1}{n_c}\kappa\sigma^4+3\frac{n_c-1}{n_c}\sigma^4,. \tag{72}$$

*where $\kappa$ is the kurtosis index (and then $\kappa\sigma^4$ is the fourth moment). Then, for $n_c\leqslant 2$*

$$\mathbb{E}\left[\|_cP\,{}_c\mathbf{x}-{}_c\boldsymbol{\mu}\|^4\right]\leqslant\frac{3}{n_c^2}\mathbb{E}\left[\|_c\mathbf{x}-{}_cP\,{}_c\mathbf{x}\|^4\right]\leqslant\frac{3}{n_c^2}\mathbb{E}\left[\|_c\mathbf{x}-{}_c\boldsymbol{\mu}\|^4\right], \tag{73}$$

*showing the advantage of averaging the estimates across all agents in the same connected components.*

### F.4 $(\varepsilon,\delta)$-Bounds: Proof of Theorem 3

We prove this theorem whose scope is larger.

**Theorem 22.** *Consider a graph component $c$ and pick uniformly at random an agent $a$ in $c$, then*

$$\mathbb{P}\left(\forall t>\tau_a^C,|\hat{\mu}_a^t-\mu_a|<\varepsilon\right)\geq 1-\delta$$

*where*

$$\tau_a^C=\max\left\{\zeta_D,C'\frac{\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}\|^4\right]}{n_c\varepsilon^4\delta}\left(\frac{(1-\alpha)^2}{(1-\alpha\lambda_2(W))^2}\left(1+\frac{1}{\ln\frac{1}{\alpha\lambda_2(W)}}\right)^2+\frac{1}{n_c^2}\left(1+\frac{1}{\ln\frac{1}{\alpha}}\right)^2\right)\right\}$$

*for $\alpha_t=\alpha$, and*

$$\tau_a^C\triangleq\max\left\{\zeta_D,C''\frac{\mathbb{E}\left[\|_cP\,{}_c\mathbf{x}-{}_c\boldsymbol{\mu}\|^4\right]}{n_c\varepsilon^4\delta}\ln^2\left(eC''\frac{\mathbb{E}\left[\|_cP\,{}_c\mathbf{x}-{}_c\boldsymbol{\mu}\|^4\right]}{n_c\varepsilon^4\delta}\right)\right\}.$$

*for $\alpha_t=t/(t+1)$.*

*Proof.* We start considering the auxiliary system studied in the previous sections: consensus matrices can be arbitrary until time $\zeta_D$ and then agents acquire perfect knowledge about which neighbours belong to the same class and simply rely on information

arriving through these links.

$$\mathbb{P}\left(|\hat{\mu}_a^{t+1} - \mu_a| \geq \varepsilon\right) = \mathbb{P}\left(\left(\hat{\mu}_a^{t+1} - \mu_a\right)^4 \geq \varepsilon^4\right) \tag{74}$$

$$\leqslant \frac{\mathbb{E}\left[\left(\hat{\mu}_a^{t+1} - \mu_a\right)^4\right]}{\varepsilon^4} \tag{75}$$

$$= \frac{\frac{1}{n_c}\sum_{a'=1}^{n_c}\mathbb{E}\left[\left(\hat{\mu}_a^{t+1} - \mu_a\right)^4\right]}{\varepsilon^4} \tag{76}$$

$$= \frac{\mathbb{E}\left[\left\|{}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu}\right\|^4\right]}{n_c\varepsilon^4} \tag{77}$$

Applying the union bound, we obtain:

$$\mathbb{P}\left(\exists t \geq t' | |\hat{\mu}_a^{t+1} - \mu_a| \geq \varepsilon\right) \leq \sum_{t=t'}^{\infty} \frac{\mathbb{E}\left[\left\|{}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu}\right\|^4\right]}{n_c\varepsilon^4}. \tag{78}$$

When $\alpha_t = \alpha$, considering the dominant term in (59) and (73) leads to

$$\mathbb{E}\left[\left\|{}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu}\right\|^4\right] \leqslant \frac{C'}{2}\frac{\mathbb{E}\left[\|{}_c\mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{(t+1)^2}\left(\frac{(1-\alpha)^2}{(1-\alpha\lambda_2(W))^2}\left(1 + \frac{1}{\ln\frac{1}{\alpha\lambda_2(W)}}\right)^2 + \frac{1}{n_c^2}\left(1 + \frac{1}{\ln\frac{1}{\alpha}}\right)^2\right) \tag{79}$$

We observe that

$$\sum_{t=t'}^{\infty}\frac{1}{(t+1)^2} \leqslant \int_{t'}^{\infty}\frac{1}{t^2}\,\mathrm{d}t = \frac{1}{t'}, \tag{80}$$

from which we conclude:

$$\mathbb{P}\left(\exists t \geq t' : |\hat{\mu}_a^{t+1} - \mu_a| \geq \varepsilon\right) \leq \frac{C'}{2}\frac{\mathbb{E}\left[\|\mathbf{x} - \boldsymbol{\mu}\|^4\right]}{n_c\varepsilon^4 t}\left(\frac{(1-\alpha)^2}{(1-\alpha\lambda_2(W))^2}\left(1 + \frac{1}{\ln\frac{1}{\alpha\lambda_2(W)}}\right)^2 + \frac{1}{n_c^2}\left(1 + \frac{1}{\ln\frac{1}{\alpha}}\right)^2\right). \tag{81}$$

This probability is then smaller than $\delta/2$ for

$$t' \geqslant \tau_a^C = \max\left\{\zeta_D, C'\frac{\mathbb{E}\left[\|\mathbf{x} - \boldsymbol{\mu}\|^4\right]}{n_c\varepsilon^4\delta}\left(\frac{(1-\alpha)^2}{(1-\alpha\lambda_2(W))^2}\left(1 + \frac{1}{\ln\frac{1}{\alpha\lambda_2(W)}}\right)^2 + \frac{1}{n_c^2}\left(1 + \frac{1}{\ln\frac{1}{\alpha}}\right)^2\right)\right\}. \tag{82}$$

Similarly, when $\alpha_t = \frac{t}{t+1}$, considering the dominant term in (60) leads to

$$\mathbb{E}\left[\left\|{}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu}\right\|^4\right] \leqslant \frac{C''}{16}\mathbb{E}\left[\|{}_cP\,{}_c\mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]\left(\frac{\ln(1+t)}{1+t}\right)^2. \tag{83}$$

We observe that

$$\sum_{t=t'}^{\infty}\left(\frac{\ln(1+t)}{t+1}\right)^2 \leqslant \int_{t'}^{\infty}\left(\frac{\ln t}{t}\right)^2\mathrm{d}t \tag{84}$$

$$= \int_{\ln t'}^{\infty} x^2 e^{-x}\,\mathrm{d}x \tag{85}$$

$$= \left(e^{-x}\left(x^2 + 2x + 2\right)\right)\Big|_{\infty}^{\ln t'} \tag{86}$$

$$= \frac{(\ln t')^2 + 2\ln t' + 2}{t'}, \tag{87}$$

$$= \frac{(\ln t' + 1)^2 + 1}{t'}, \tag{88}$$

$$\leqslant 2\frac{(\ln t' + 1)^2}{t'}, \tag{89}$$

$$= 2\frac{\ln^2(et')}{t'} \tag{90}$$

from which we conclude:

$$\mathbb{P}\left(\exists t \geq t' : |\hat{\mu}_a^{t+1} - \mu_a| \geq \varepsilon\right) \leqslant \frac{C''}{8} \frac{\mathbb{E}\left[\|_c P_c \mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{n_c \varepsilon^4} \frac{\ln^2(et')}{t'}. \tag{91}$$

This probability is then smaller than $\delta/2$ for

$$t' \geqslant \frac{C''}{4} \frac{\mathbb{E}\left[\|_c P_c \mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{n_c \varepsilon^4 \delta} \ln^2(et'), \tag{92}$$

and by applying Lemma 20 with $x = t'e$, we obtain that a sufficient condition is

$$t' \geqslant C'' \frac{\mathbb{E}\left[\|_c P_c \mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{n_c \varepsilon^4 \delta} \ln^2\left(eC'' \frac{\mathbb{E}\left[\|_c P_c \mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{n_c \varepsilon^4 \delta}\right). \tag{93}$$

Let then define

$$\tau_a^C \triangleq \max\left\{\zeta_D, C'' \frac{\mathbb{E}\left[\|_c P_c \mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{n_c \varepsilon^4 \delta} \ln^2\left(eC'' \frac{\mathbb{E}\left[\|_c P_c \mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{n_c \varepsilon^4 \delta}\right)\right\}. \tag{94}$$

Finally, let us consider the system of interest. With probability $1 - \delta/2$ the agents will have identified the correct links by time $\zeta_D$. The corresponding trajectories coincide with trajectories of the auxiliary system we studied. For the auxiliary system, the estimates have the required precision after time $\tau_a^C$ with probability $1 - \delta/2$. It follows that the probability that the estimates in the stochastic have not the required precision after time $\tau_a^C$ is at most $\delta$. □

# G Appendix G - On the Structure of $G(N,r)$ and its Impact on Performance of our Algorithms

## G.1 On the Local Tree Structure of $G(N,r)$

Let $\mathcal{N} = G(\mathcal{V},\mathcal{E})$ be a network sampled from the class of random regular graphs $G(N,r)$ with fixed degree $r$. [6] Here and in the following we fix $\mathcal{V} = \mathcal{A}$ and $n = |\mathcal{V}| = |\mathcal{A}|$. We are interested to investigate the structure of the $d$-deep neighborhood, $\mathcal{N}_d^{(v)}$, of a generic node $v$ (i.e. the sub-graph inter-connecting nodes at distance smaller or equal than $d$v from $v$). Observe that as $n$ grows large, for any $d \in \mathbb{N}$, we should expect that $\mathcal{N}_d^{(v)}$ is likely equal to $\mathcal{T}_d$, a perfectly balanced tree of depth $d$, in which the root has $r$ children, and all the other nodes have $r-1$. children.

To this end using an approach inspired by Lemma 5 in (Rossi, Como, and Fagnani 2019) (which applies to directed graphs), we can claim that:

**Theorem 23.** *Whenever $v_0$ is chosen uniformly at random, we have*

$$\mathbb{P}(\mathcal{N}_d^{(v_0)} \neq \mathcal{T}_d) \leq \frac{(H+1)H}{2N}$$

*where $H = 1 + \sum_{d'=1}^{d} r(r-1)^{d'-1}$.*

*Proof.* First observe that realizations of $G(N,r)$ graphs are typically obtained through the following standard procedure: every node is initially connected to $r \geq 2$ stubs; then stubs then are sequentially randomly matched/paired to form edges, as follows: at each stage, an arbitrarily selected free/unpaired stub is selected and paired with a different stub picked uniformly at random among the still unpaired stubs.

We identify every stub with different number in $[1, rN]$ (we assume $rN$ to be even). Now, let $\nu(i)$ for $i \in [1, rN]$ be the function that returns the identity of the node to which the stub is connected. See Fig. 6

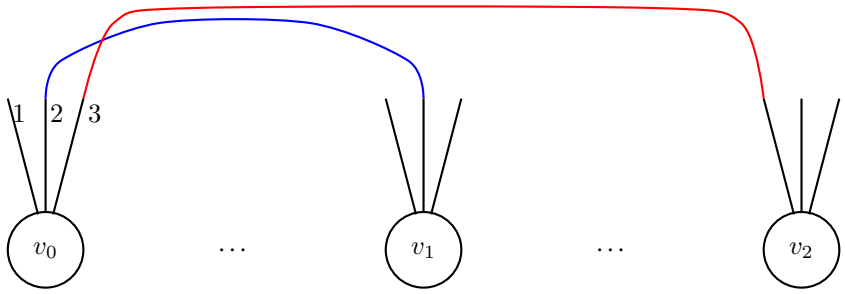



Figure 6: Random matching of stubs, $r = 3$

Our procedure explores the $d$-neighborhood of a node $v_0$, taken at random, by sequentially unveiling stub-pairings, according to a bread-first approach. At every step the procedure checks weather the already explored-portion of $\mathcal{N}_d^{(v_0)}$, $G(\mathcal{V},\mathcal{E})$, has a tree structure.

The procedure initializes $\mathcal{V}^{(0)} = v_0$ and $\mathcal{E}^{(0)} = \emptyset$.

At step 1, our procedure takes a stub $k_1$ connected to $v_0$ (i.e. such that $\nu(k_1) = v_0$) and matches it, uniformly at random with another stub $r(k_1) \neq k_1$. Let $v_1 := \nu(r(k_1))$. Then the procedure updates sets $\mathcal{V}$ and $\mathcal{E}$ according to the rule: $\mathcal{V}^{(1)} = \mathcal{V}^{(0)} \cup \{v_1\}$ and $\mathcal{E}^{(1)} = \mathcal{E}^{(0)} \cup \{(v_0, v_1)\}$. At this stage $G(\mathcal{V}^{(1)}, \mathcal{E}^{(1)})$ is a tree only provided that $v_1 \notin \mathcal{V}^{(0)}$. This happens with a probability:

$$p_1 := \mathbb{P}(v_1 := \nu(r(k_1)) \notin \{v_0\}) = 1 - \frac{r-1}{Nr-1}.$$

In case $v_1 \neq v_0$ the algorithm proceeds, otherwise it prematurely terminates providing $G(\mathcal{V}^{(1)}, \mathcal{E}^{(1)})$. At step 2, our procedure takes a new unmatched stub $k_2$ (connected again to $v_0$). $k_2$ is matched uniformly at random with another free (i.e. still unmatched) stub $r(k_2) \notin \{k_1, r(k_1), k_2\}$, let $v_2 := \nu(r(k_2))$. Then sets $\mathcal{V}$ and $\mathcal{E}$ are updated: $\mathcal{V}^{(2)} = \mathcal{V}^{(1)} \cup \{v_2\}$ and $\mathcal{E}^{(2)} = \mathcal{E}^{(2)} \cup \{(v_0, v_2)\}$. $G(\mathcal{V}^{(2)}, \mathcal{E}^{(2)})$ is a tree only under the condition that $v_2 := \nu(r(k_2)) \notin \mathcal{V}^{(1)}$. This happens with a probability:

$$p_2 := \mathbb{P}(v_2 := \nu(r(k_2)) \notin \{v_0, v_1\} \mid v_0 \neq v_1) = 1 - \frac{2r-2}{Nr-3}$$

again if $G(\mathcal{V}^{(2)}, \mathcal{E}^{(2)})$ is a tree the algorithm proceeds, otherwise it prematurely terminates, providing $G(\mathcal{V}^{(2)}, \mathcal{E}^{(2)})$. At a generic step $h$, our procedure takes a free stub $k_h$, connected to vertex $v_{\lfloor (h-1)/r \rfloor} \in \{v_0, v_1, \cdots, v_{h-1}\}$, (recall that we explore

[6] observe that graphs in $G(N,r)$ are not necessarily simple

nodes/edges according to a breadth-first approach), and matches it with a randomly chosen (still unmatched) stub $r(k_h)$. Let $v_h = r(k_h)$. Then Then sets $\mathcal{V}$ and $\mathcal{E}$ are updated as follows: $\mathcal{V}^{(h)} = \mathcal{V}^{(h-1)} \cup \{v_h\}$ and $\mathcal{E}^{(h)} = \mathcal{E}^{(h-1)} \cup \{(v_{\lfloor (h-1)/r \rfloor}, v_h)\}$. Again $G(\mathcal{V}^{(h)}, \mathcal{E}^{(g)})$ is a tree only if $v_h \notin \mathcal{V}^{(h-1)}$, and this happens with a probability

$$\begin{aligned} p_h :=& \mathbb{P}(\nu(r(k_h)) \notin \{v_0, v_1, \cdots, v_{h-1}\} \\ & \mid \{v_0 = v_1 =, \cdots, = v_{h-1}\}) \\ =& 1 - \frac{hr - (2h-1)}{Nr - 2(h-1)}, \end{aligned}$$

in such a case the algorithm proceeds, otherwise it prematurely terminates, providing $G(\mathcal{V}^{(h)}, \mathcal{E}^{(h)})$ in output.

The algorithm naturally terminates (providing a tree) when all the nodes in $\mathcal{N}_d^{(v_0)}$ have been unveiled (i.e., placed in $\mathcal{V}$) and the corresponding unveiled graph $G(\mathcal{V}, \mathcal{E})$ is a tree. This happens at step $H$. The probability that the algorithm terminates providing a tree is given by:

$$\begin{aligned} \mathbb{P}(\mathcal{N}_d^{(V)} \neq \mathcal{T}_d) =& 1 - \prod_{h=1}^{H} p_h \leq \sum_h (1 - p_h) \\ =& \sum_{h=1}^{H} \frac{hr - (2h-1)}{Nr - 2(h-1)} \\ \leq& \sum_{h=1}^{H} \frac{h}{N} = \frac{1}{N} \sum_1^H h = \frac{(H+1)H}{2N} \end{aligned}$$

□

Now denoting with $M$ the number of vertices $v \in \mathcal{V}$ for which $\mathcal{N}_d^{(v)} \neq \mathcal{T}_d$, we have that

$$M = \sum_{v \in \mathcal{V}} \mathbf{1}_{\{\mathcal{N}_d^{(v)} \neq \mathcal{T}_d\}}$$

and therefore

$$\begin{aligned} \mathbb{E}[M] &= \sum_{v \in \mathcal{V}} \mathbb{E}[\mathbf{1}_{\{\mathcal{N}(v)_d \neq \mathcal{T}_d\}}] = \sum_{v \in \mathcal{V}} \mathbb{P}(\mathcal{N}_d^{(v)} \neq \mathcal{T}_d) \\ &= N \mathbb{P}(\mathcal{N}_d^{(v_0)} \neq \mathcal{T}_d) \leq \frac{(H+1)H}{2} \end{aligned}$$

where $v_0$ is uniformly taken at random. By assuming $(H+1)H = o(N)$, and applying Markov inequality we can claim that for any $\varepsilon > 0$ arbitrarily slowly:

$$\mathbb{P}\left(\frac{M}{N} > \varepsilon\right) \downarrow 0 \quad \forall \varepsilon > 0. \tag{95}$$

i.e. the fraction of nodes $v$ for which $\mathcal{N}_d^{(v_0)} \neq \mathcal{T}_d$ is negligible with a probability tending to 1.

At last we would like to highlight that (95) can be transferred to the class $G_0(N, r)$ of uniformly chosen *simple* regular graphs thanks to:

**Proposition 24** (Amini (2010))**.** *Any sequence of event $E_n$ occurring with a probability tending to 1 (0) in $G(N, r)$ occurs well in $G_0(N, r)$ with a probability tending to 1 (0).*

Moreover, recalling that by construction $\frac{M}{N} \leq 1$, from (95) we can immediately deduce that that $\mathbb{E}[M]/N \to 0$ on $G_0((N, r)$.

At last we would like to mention the following result, from which Proposition 24 rather immediately descends.

**Theorem 25** ((Janson 2009))**.** *t*

$$\liminf_{n \to \infty} \mathbb{P}(G((N, r) \textit{ is simple}) > 0$$

Observe Theorem 25 provides a theoretical foundation to the to design a simple algorithm for the generation of a graph in class $G_0((N, r)$, based on the superposition of an acceptance/rejection procedure to the generation of graphs in $G(N, r)$. More efficient algorithms are, however, well known in literature (McKay and Wormald 1990).

## G.2 The Structure of $\mathcal{CC}_a^d$ and $\mathcal{CC}_a$

Now we investigate on the structure of $\mathcal{CC}_a^d$. We assume that agents $a \in \mathcal{A}$ are partitioned into a finite number $K$ of similarity classes. Agents are assigned to similarity classes independently. We indicate with $p_k$ the probability according to which agent $a$ is assigned to class $k$. Note that by construction $\sum_{k=1}^{K} p_k = 1$. We denote with $k_a$ the similarity class to which $a$ is assigned. In this scenario the structure of $\mathcal{CC}_a^d$ can be rather easily analyzed, it turns out that:

**Theorem 26.** *Conditionally over the event $\{\mathcal{N}_a^d$ is a tree$\}$. $\mathcal{CC}_a^d$ has the structure of a Branching process originating from a unique ancestor, obeying to the following properties:*

- *the number of off-springs of different nodes are independent.*
- *the number of off-spring of the ancestor (generation 0 node) is distributed as a Bin$(r, p_{k_a})$;*
- *while the number of offs-springs of any generation $i$ node (with $1 \leq i < d$) is distributed as Bin$(r-1, p_{k_a})$.*
- *generation-$d$ nodes have no off-springs.*

*Proof.* The proof is rather immediate. Consider $a$ and explore $\mathcal{CC}_a^d$ according to a breath-first exploration process that stops at depth-$d$. The number of off-springs of $a$, $O_a$, by construction, equals the number of nodes in $\mathcal{N}_a$ that belongs to class $k_a$. This number $O_a$ is distributed as: $O_a \overset{L}{=} \text{Bin}(r, p_{k_a})$. Consider, now, any other explored node $a'$, at distance $i < d$ from the ancestor, this node, by construction, will have a unique parent node $p_{a'}$, off-springs of $a'$, $O_{a'}$ will be given by all nodes in $\mathcal{N}_{a'} \setminus \{p_{a'}\}$ that belong to class $k_a$. Their number, $O_{a'}$, is distributed as: $O_{a'} \overset{L}{=} \text{Bin}(r, p_{k_{a'}})$. See Fig. 7. When the exploration reaches a $d$-depth node, it stops, therefore the number of off-springs for every $d$-depth nodes is zero. A last, since the sets $\mathcal{N}_{a'} \setminus \{p_{a'}\}$ are disjoint (as immediate consequence of the fact that $\mathcal{N}_a^d$ is a tree), the variables in $\{O_{a'}\}_{a' \in \mathcal{CC}_a^d}$ are independent. □

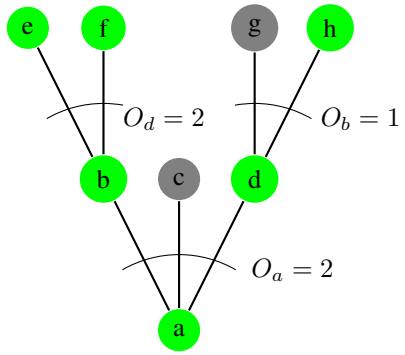



Figure 7: Structure $|\mathcal{CC}_a^d|$: an example for $r = 3$, $d = 2$; green nodes belong to $|\mathcal{CC}_a^d|$.

We can now recall and adapt a few standard asymptotic results on Galton-Watson (GW) Processes (in a standard GW process all nodes in the tree give origin to number of off-springs, which are identically distributed and independent). First observe that, in our case, as $|\mathcal{A}| \to \infty$ we can assume that $d \to \infty$ as well, for example choosing $d$ as in Proposition 6.

We say that a standard GW process is super-critical if the average number off-springs of every node $\mathbb{E}[O_a] := m > 1$. Now for a standard supercritical GW process, denoted with $Z_i$ the number of nodes belonging to generation $i$, we have:

**Theorem 27** (extinction-explosion principle)**.** *For every super-critical nontrivial [7] GW process $\{Z_i\}_i$ is bound to either extinction or explosion, i.e.,*

$$\mathbb{P}\left(Z_i = 0 \textit{ eventually}\right) + \mathbb{P}\left(\lim \frac{Z_i}{m^i} > 0\right) = 1.$$

Previous result can immediately extended to our two stages branching process as $d \to \infty$. Denoted with $m_0 = rp_{k_a}$ and $m := (r-1)p_{k_a}$, under the assumption that $m > 1$ we obtain that:

$$\mathbb{P}\left(Z_i = 0 \text{ eventually}\right) + \mathbb{P}\left(\lim \frac{Z_i}{m_0 m^{i-1}} > 0\right) = 1. \tag{96}$$

Indeed the distribution of off-springs at the root, has not impact on structural properties of the process.

Now we focus on the extinction probability $q_{2BP} := \mathbb{P}\left(Z_i = 0 \text{ eventually}\right)$ in a two-stages branching process, as our process. To compute it, we can adapt classical results on GW. It turns out that:

[7] a GW process is non trivial if $\mathbb{P}(O_a = 0) > 0$

| | $p_{k_a}=1/2$ | $p_{k_a}=1/4$ | $p_{k_a}=1/8$ |
|---|---|---|---|
| $r=4$ | $0.146$ | $1$ | $1$ |
| $r=8$ | $4.17\cdot 10^{-3}$ | $0.176$ | $1$ |
| $r=16$ | $1.52\cdot 10^{-5}$ | $1.08\cdot 10^{-2}$ | $0.190$ |
| $r=32$ | $2.32\cdot 10^{-10}$ | $1.01\cdot 10^{-4}$ | $1.51\cdot 10^{-2}$ |

Table 3: extinction probability, $q_{2BP}$, for different values of $r$ and $p_{k_a}$.

**Theorem 28.** *Consider a two-stages branching process, in which the number of offspring of the root is Bin$(p_k, r)$ while the off-spring of every other node Bin$(p_k, r-1)$. Its extinction probability $q_{2BP}$ is given by*

$$q_{2BP}=\sum_{h=0}^{r}\binom{r}{h}p_k^h(1-p_k)^{r-h}q_{GW}^h=[(1-p_k)+p_kq_{GW}]^r$$

*where $q_{GW}$ is the extinction probability of a standard GW, with distribution of off-springs given by Bin$(r-1,p_k)$. $q_{GW}$ can be easily computed as the only solution in $(0,1)$ of equation:*

$$t=[(1-p_k)+p_kt]^{r-1}$$

*Proof.* The proof is immediate, considering that: i) every sub-tree originated by an offspring of the ancestor has the same structure of a standard GW. The event $\{Z_i=0\}$ is equivalent to event {every sub-tree originated by every offspring of the ancestor is extincted within $i-1$ generations}. Then conditioning on the number of off-springs of the ancestor we get the claim. Of course previously computed asymptotic extinction probability $q_{2BP}$ provides an upper-bound to the probability of early extinction of a $d$-depth truncated two-stages Branching process. □

At last observe that, choosing $d$ as in Proposition 6, we have $m_0m^{d-1}=\widetilde{\Theta}(|\mathcal{A}|^{\frac{1}{2}(1+\log_{r-1}p_{k_a})})$, therefore, recalling that by construction $|\mathcal{CC}_a^d|>Z_d$, by (96), we can always select $f(|\mathcal{A}|)$ that satisfies jointly $f(|\mathcal{A}|)=o(m_0m^{d-1})$ and $f(|\mathcal{A}|)=\widetilde{\Theta}(|\mathcal{A}|^{\frac{1}{2}(1+\log_{r-1}p_{k_a})})$, such that:

$$\lim_{|\mathcal{A}|\to\infty}\mathbb{P}(|\mathcal{CC}_a^d|>f(n))=1-q_{2BP}.$$

Moreover, note that the extinction probability $q_{2BP}$ is actually a function of its two parameters $(r,p_{k_a})$, its is rather immediate to show that

$$\lim_{r\to\infty}q_{2BP}(r,p_{k_a})=0\qquad\forall p_{k_a}>0$$

Therefore choosing $r$ sufficiently large we can make $q_{2BP}(r,p_{k_a})$ arbitrarily small and at the same time guarantee $|\mathcal{A}|^{\frac{1}{2}(1+\log_{r-1}p_{k_a})}>|\mathcal{A}|^{\frac{1}{2}-\phi}$ for an arbitrarily small $\phi>0$.

At last observe that if we turn our attention to $\mathcal{CC}_a$, since by construction we have $\mathcal{CC}_a^d\subseteq\mathcal{CC}_a\ \forall d,a$, rather immediately we have:

$$\lim_{|\mathcal{A}|\to\infty}\mathbb{P}(|\mathcal{CC}_a|>g(n))=1-q_{2BP}.$$

for any $g(n)=o(|\mathcal{A}|^{\frac{1}{2}(1+\log_r p_{k_a})})$.

# H Appendix H - Proof of the results in Table 2

The results in Table 2 follow immediately, once we derive the asymptotics for $n^\star_\gamma(x)$ in the sub-Gaussian setting and in bounded 4-th moment setting. The asymptotics for $\tilde{n}_\gamma(x)$ are immediate to derive from the bounds in Theorem 4.

## H.1 $n^\star_\gamma(x)$, sub-Gaussian setting

Remember that $n^\star_\gamma(x) \coloneqq \lfloor \beta_\gamma^{-1}(x) \rfloor$, i.e., $n^\star_\gamma(x)$ is the smallest integer $n$ such that

$$\beta_\gamma(n) := \sigma\sqrt{\frac{2}{n}\left(1+\frac{1}{n}\right)\ln(\sqrt{(n+1)}/\gamma)} \leq x.$$

As we are interested in upper bounds for $n^\star_\gamma(x)$, we can start by upperbounding the left-hand side expression.

$$\sigma\sqrt{\frac{2}{n}\left(1+\frac{1}{n}\right)\ln\left(\frac{\sqrt{(n+1)}}{\gamma}\right)} \leq \sigma\sqrt{\frac{4}{n}\ln\left(\frac{\sqrt{2n}}{\gamma}\right)} = \sigma\sqrt{\frac{2}{n}\ln\left(\frac{2n}{\gamma^2}\right)}, \tag{97}$$

and imposing that the right-hand side is smaller than $x$, we obtain:

$$\sigma^2\frac{2}{n}\ln\left(\frac{2n}{\gamma^2}\right) \leq x^2 \tag{98}$$

$$\frac{4\sigma^2}{\gamma^2x^2}\ln\left(\frac{2n}{\gamma^2}\right) \leq \frac{2n}{\gamma^2}. \tag{99}$$

From (Shalev-Shwartz and Ben-David 2014, Lemma A.1) a sufficient condition for this inequality to hold is

$$\frac{8\sigma^2}{\gamma^2x^2}\ln\left(\frac{4\sigma^2}{\gamma^2x^2}\right) \leq \frac{2n}{\gamma^2}, \tag{100}$$

and then

$$\frac{4\sigma^2}{x^2}\ln\left(\frac{4\sigma^2}{\gamma^2x^2}\right) \leq n. \tag{101}$$

We can conclude that

$$n^\star_\gamma(x) \in \mathcal{O}\left(\frac{\sigma^2}{x^2}\ln\left(\frac{\sigma}{\gamma x}\right)\right), \tag{102}$$

from which the asymptotics for $\zeta_a$ and $\tau_a$ can be derived opportunely replacing $x$ and $\gamma$.

## H.2 $n^\star_\gamma(x)$, Bounded 4-th Moment Setting

The reasoning is analogous, but we start from

$$\beta_\gamma(n) = \left(2\frac{\kappa+3\sigma^4}{\gamma}\right)^{\frac{1}{4}}\left(\frac{1+\ln^2 n}{n}\right)^{\frac{1}{4}}. \tag{103}$$

For $n \geq 3$

$$\left(2\frac{\kappa+3\sigma^4}{\gamma}\right)^{\frac{1}{4}}\left(\frac{1+\ln^2 n}{n}\right)^{\frac{1}{4}} \leq \left(2\frac{\kappa+3\sigma^4}{\gamma}\right)^{\frac{1}{4}}\left(\frac{2\ln^2 n}{n}\right)^{\frac{1}{4}}, \tag{104}$$

and imposing the RHS to be smaller than $x$:

$$4\frac{\kappa+3\sigma^4}{\gamma}\frac{\ln^2 n}{n} \leq x^4 \tag{105}$$

$$4\frac{\kappa+3\sigma^4}{\gamma x^4}\ln^2 n \leq n, \tag{106}$$

and from Lemma 20 a sufficient condition for this inequality to hold is

$$\max\left\{16\frac{\kappa+3\sigma^4}{\gamma x^4}\ln^2\left(16\frac{\kappa+3\sigma^4}{\gamma x^4}\right), 1\right\} \leq n. \tag{107}$$

We can conclude that

$$n^\star_\gamma(x) \in \tilde{\mathcal{O}}\left(\frac{\kappa+3\sigma^4}{\gamma x^4}\right). \tag{108}$$

# I Appendix I - Additional Performance Evaluation of the B-ColME and C-ColME

In this Appendix, we report further results on the performance of B-ColME and C-ColME, as a function of the underlying graph structure and characterizing parameters.

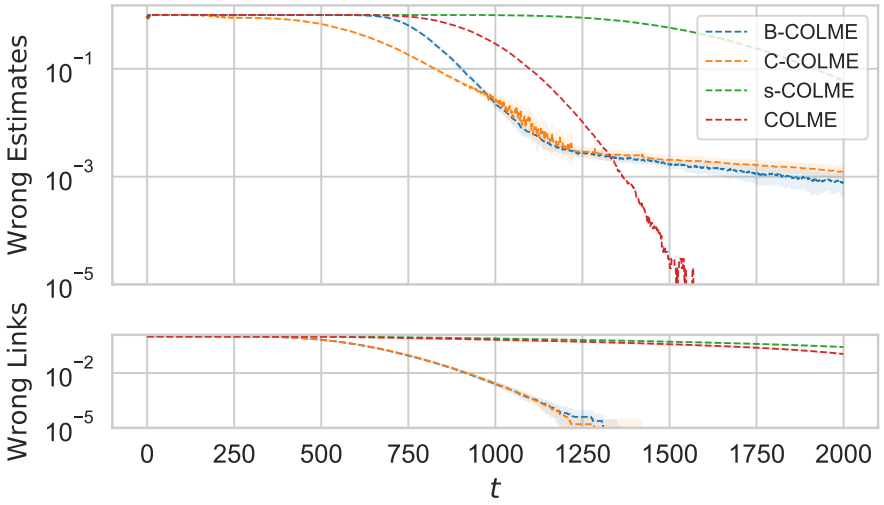



Figure 8: All approaches compared considering a smaller $r$ compared to Figure 2 ($r = 5$), this shows that $r$ needs to be chosen appropriately (large enough).

## I.1 Over a $G_0(N, r)$ Varying $r$

Here, we explore the performance of the two proposed *scalable* algorithms as a function of the number of neighbors they are allowed to contact during the dynamics, i.e., the parameter $r$ of the $G_0(N, r)$ graph.

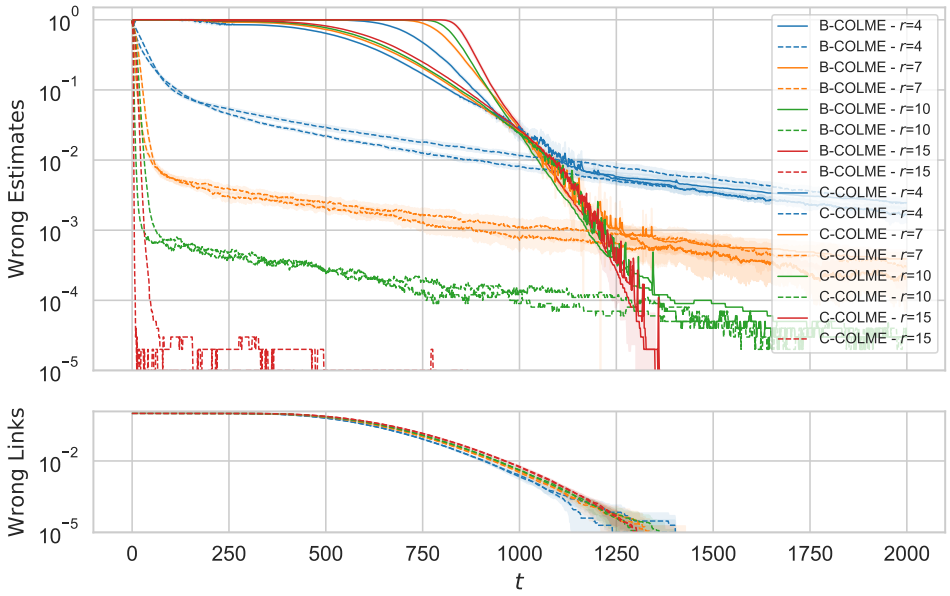



Figure 9: Performance comparison as a function of the parameter $r$ of $G_0(N, r)$ for B-ColME and C-ColME. Dashed line is used for the *oracle* while solid line for the algorithm.

## I.2 B-ColME as a Function of the Depth of Information Kept

In the B-ColME algorithm, many cycles in the graph can degrade the performance of the algorithm (hence the need for a tree-like local structure). Here we study the performance of the algorithm as a function of the depth $d$ of the neighborhood that

receives the estimate of a given node, and we find that a high value of this parameter eventually degrades the performance of the algorithm.

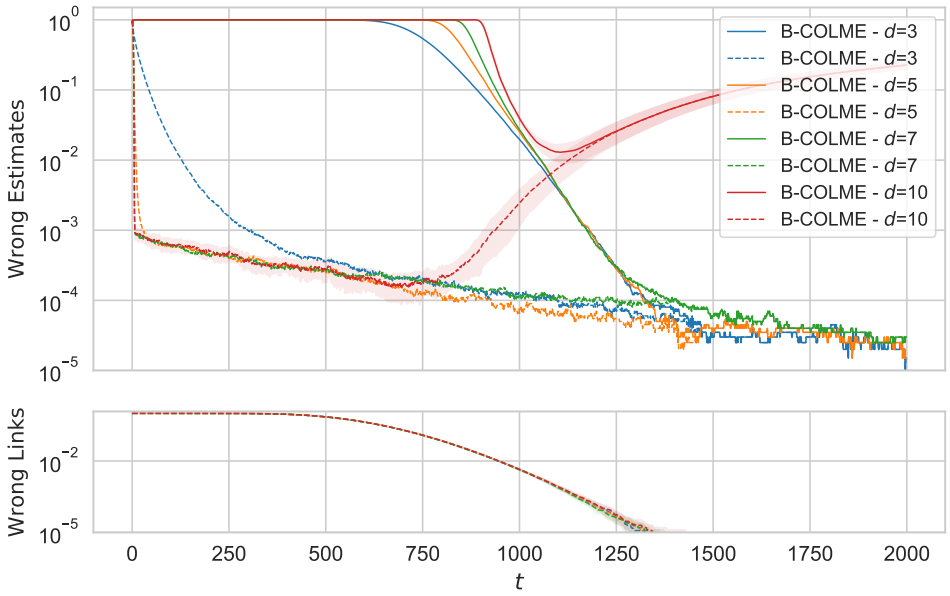



Figure 10: Performance comparison of B-ColME as a function of the *depth* $\kappa$ of the info kept. Dashed line is used for the *oracle* while solid line for the algorithm.

### I.3 C-ColME (Constant $\alpha$) as a Function of the Weight $\alpha$

We report some experiments considering $\alpha$ constant, as opposed as $\alpha = \frac{t}{t+1}$ (refreshed at each topology modification), and explore the impact of the parameter on the probability of error.

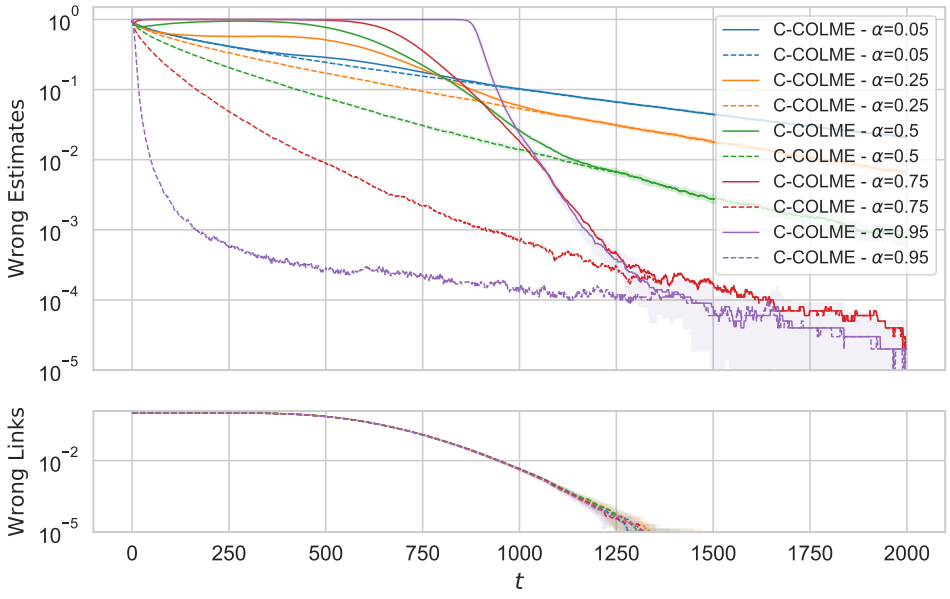



Figure 11: Performance comparison of C-ColME as a function of the weight $\alpha$ when it is considered as constant. Dashed line is used for the *oracle* while solid line for the algorithm.

# J Appendix J - Additional Details on the Decentralized Federated Learning Approach FL-DG

B-ColME and C-ColME are useful in contexts where nodes want to learn the preferences of users by using information from other peers in addition to local data, or when sensors jointly try to estimate certain quantities in a very heterogeneous environment (e.g., smart farming). Nevertheless, it is of great interest to apply these techniques in the context of decentralized personalized federated learning, where the goal of the nodes is to learn a machine learning model on a given local dataset $D_a$ while having the possibility to collaborate with other agents (assuming that they can be classified into one of $C$ possible classes, each characterized by a distribution $D_c$). We propose FL-DG, a decentralized FL algorithm inspired by B-ColME and C-ColME, where agents decide which peers to collaborate with based on the cosine similarity between the weights updates of their models. We compare it to a classical decentralized FL algorithm over a static graph (FL-SG).

Algorithm 3: FL-SG Training

**Input:** $\mathcal{G} = (\mathcal{A}, \mathcal{E})$, $D_a \in \{D_c\}_{c=1}^{C}\ \forall a \in \mathcal{A}$, $\theta_0$
**Output:** collaborative FL-SG model $\theta_a$, $\forall a \in \mathcal{A}$
$\mathcal{C}_a^0 \leftarrow \mathcal{N}_a^t\ \forall a \in \mathcal{A}, \quad \theta_a^0 \leftarrow \theta_0, \quad \theta_l^0 \leftarrow \theta_0\ \forall l \in \mathcal{E}$
**while** new sample $s_a^t$ arrives at time $t$ **do**
*// Training Phase*
**for** node $a$ in $\mathcal{A}$ **do**
**for** epoch $e$ in $\{1, .., E\}$ **do**
**for** minibatch $M_a^t$ in $\{M_{a,0}^t, M_{a,1}^t\}$ **do**
$\theta_a^{t+1} \leftarrow \theta_a^t + \text{SGD}(\theta_a^t, D_a^t)$
**for** neighbor $a'$ in $\mathcal{N}_a \cap \mathcal{C}_a^t$ **do**
$\theta_{a,a'}^{t+1} \leftarrow \theta_{\{a,a'\}}^t + \text{SDG}(\theta_{\{a,a'\}}^t, M_a^t)$
**end for**
**end for**
**end for**
**end for**
*// Discovery Phase*
**for** undirected link $\{a, a'\}$ in $\mathcal{E}$ **do**
$\Delta\theta_{a,a'}^{t+1} \leftarrow \theta_{a,a'}^{t+1} - \theta_{\{a,a'\}}^t$
$\Delta\theta_{a',a}^{t+1} \leftarrow \theta_{a',a}^{t+1} - \theta_{\{a,a'\}}^t$
$\omega_{\{a,a'\}}^{t+1} \leftarrow \frac{1}{t+1} \frac{\langle \Delta\theta_{a,a'}^{t+1}, \Delta\theta_{a',a}^{t+1} \rangle}{||\Delta\theta_{a,a'}^{t+1}|| \cdot ||\Delta\theta_{a',a}^{t+1}||} + \frac{t}{t+1}\omega_{\{a,a'\}}^t$
**if** $\omega_{a,a'}^{t+1} < \varepsilon_1$ **then**
$\mathcal{C}_a^{t+1} \leftarrow \mathcal{C}_a^t \setminus a'$, and $\mathcal{C}_{a'}^{t+1} \leftarrow \mathcal{C}_{a'}^t \setminus a$
**end if**
**end for**
*// Model Updating Phase*
**for** node $a$ in $\mathcal{A}$ **do**
$\theta_a^{t+1} \leftarrow \frac{1}{|\mathcal{C}_a^{t+1}|+1}\theta_a^{t+1}$
**for** opt neighbor $a'$ in $\mathcal{C}_a^{t+1}$ **do**
$\theta_a^{t+1} \leftarrow \theta_a^{t+1} + \frac{1}{|\mathcal{C}_a^{t+1}|+1}\theta_{a'}^{t+1}$
**end for**
**end for**
**for** undirected link $\{a, a'\}$ in $\mathcal{E}$ **do**
$\theta_{\{a,a'\}}^{t+1} \leftarrow \frac{\theta_{a,a'}^{t+1} + \theta_{a',a}^{t+1}}{2}$ *// Update Link Model*
**end for**
$t \leftarrow t + 1$
**end while**

We focus on the case $C = 2$ and use the MNIST dataset (Deng 2012). To obtain two different distributions from the MNIST dataset, we simply swap two labels ("3" and "5" as well as "1" and "7"). Each node has the task of recognizing handwritten digits using its local data and collaborating with neighboring nodes over $\mathcal{G}$, which is a complete graph in our scenario. We use a very simple feedforward neural network model for all nodes. It consists of the input layer, a hidden layer with 100 nodes, and the output layer. We indicate the parameters of the NN as $\theta_a^t$, for agent $a$ at time $t$.

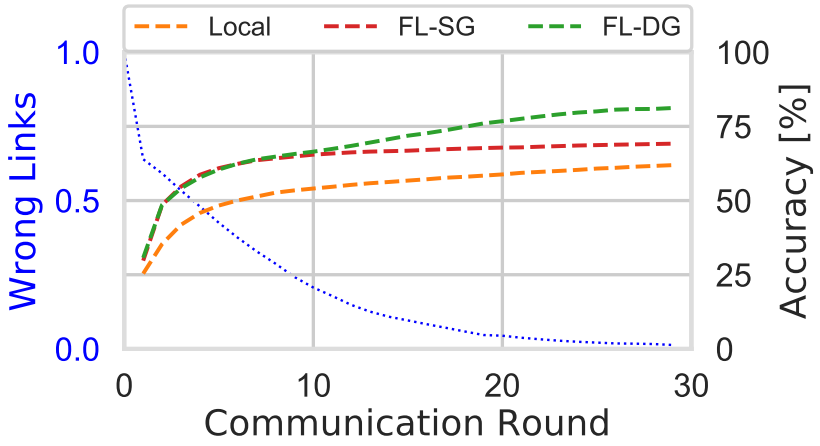



Figure 12: Accuracy of a local model (Local), a decentralized FL over a static graph (FL-SG), and our approach over a dynamic graph (FL-DG). We show also the fraction of links between communities over time for FL-DG.

Again, we consider an online setting in which agents receive new samples over time. In particular, each agent $a \in \mathcal{A}$ receives a new sample $s_a^t$ at every time instant $t$. Agents are initially assigned a local database of $M_{a,0}^0$ samples (in our example $|M_{a,0}^0| = 30$), and with the new samples they construct two overlapping minibatch $M_{a,0}^t$ and $M_{a,1}^t$. We consider a time horizon leading to two non-overlapping minibatch, i.e., $t = |M_{a,0}^0|$. The agents train their model for $E$ epochs ($E = 15$) over the two minibatch at each time instant.

Differently from the B-ColME and C-ColME mean estimation algorithms we have to modify the *discovery* phase to a large extent since the task of mean estimation and model training are structurally different. In (Sattler, Müller, and Samek 2021) it was shown that it is possible to partition the agents in a federated learning framework by using the cosine similarity of the gradient updates (or the *parameters* updates) of the considered agents. Note that cosine similarity values close to 1 indicate similar models/agents, while lower values indicate increasingly different agents. This can be intuitively understood by observing that two nodes with different data distributions are optimizing different loss functions, and, if we constrain the starting point of the optimization to be the same for both agents, we will observe an increase in the angle between the vectors corresponding to the gradient updates (see Figure 2 in (Sattler, Müller, and Samek 2021) for an illustrative example). Subject to some regularity assumptions, it is indeed possible to use the cosine similarity of the parameter updates instead of gradients. Let us denote the updates as $\Delta\theta^t = \theta^{t+1} - \theta^t$.

To allow nodes to discover their *similar* neighbors, we define a *link* model $\theta_{\{a,a'\}}^t$ (for each unordered pair $(a, a')$, "shared" between the nodes) and a node-link model $\theta_{a,a'}^t$ associated with a certain (ordered) neighbors pair $(a, a')$. Thus, every node $a \in \mathcal{A}$ keeps a model for each of its neighbors $a' \in \mathcal{N}_a$, i.e., $\theta_{a,a'}^t$. Then, at each training round, node $a$ retrieves the shared model $\theta_{\{a,a'\}}^t$ and, starting from those parameters, trains the node-link model $\theta_{a,a'}^t$ on its local data.

After all nodes have performed the *training* phase, they compute the *similarity* metric between the models, i.e., the cosine similarity $\omega_{a,a'}^t$, which allows them to determine whether to collaborate with a neighbor or not. We can compute $\omega_{a,a'}^t$ as:

$$\omega_{\{a,a'\}}^t = \frac{\langle \Delta\theta_{a,a'}^t, \Delta\theta_{a',a}^t \rangle}{||\Delta\theta_{a,a'}^t|| \cdot ||\Delta\theta_{a',a}^t||} \tag{109}$$

This metric is updated at each iteration by making an average with the previous value (see Algorithm 3). Whenever $\omega_{a,a'}^t$ goes below a certain threshold $\varepsilon_1$, link $\{a, a'\}$ is deemed to be connecting nodes of different classes and is removed from $\mathcal{E}$.

Lastly, agents update their collaborative models $\theta_a^t$, averaging the parameters of the agents $a'$ in their estimated similarity class $\mathcal{C}_a^t$. Moreover, all the link models $\theta_{\{a,a'\}}^t$ are updated averaging the two node-link models of the nodes at the ends of the link, i.e, $\theta_{\{a,a'\}}^{t+1} = \theta_{\{a,a'\}}^t + \frac{\Delta\theta_{a,a'}^t + \Delta\theta_{a',a}^t}{2}$. A detailed explanation of the model is provided in Algorithm 3.

We report the results (Fig 3 in the main article) obtained over 30 communication rounds comparing the Local model, which uses only the local dataset of each node, FL-SG a decentralized FL approach that averages the model parameters of all the neighbors over a static graph, and our FL-DG approach, again an averaging model where the nodes dynamically remove connections on the basis of the cosine similarity.