Scalable Decentralized Algorithms for Online Personalized Mean Estimation

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

[1]*Politecnico di Torino*
[2]*INRIA*

Abstract

In numerous settings, agents lack sufficient data to directly learn a model. 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 simplified version 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 introduce two collaborative mean estimation algorithms: one uses a message-passing scheme, while the other employs a consensus-based approach, with complexity $O(r \cdot \log |\mathcal{A}|)$ and $O(r)$, respectively. We establish conditions under which both algorithms yield asymptotically optimal estimates and offer a theoretical characterization of their performance.

1 Introduction

Users' devices have become increasingly sophisticated and generate vast amounts of data. This wealth of data has paved the way for the development of accurate and complex models. However, it has also introduced challenges related to security, privacy, real-time processing, and resource constraints. In response, Federated Learning (FL) has emerged as a pivotal privacy-preserving methodology for collaborative model training [Kairouz et al., 2021; Li et al., 2020]. Traditional FL methods focus on learning a unified model applicable to all clients. Yet, the statistical diversity observed across clients' datasets has led to the pursuit of *personalized* models. These models aim to better align with the unique data distributions of individual clients [e.g., Ding and Wang, 2022; Fallah et al., 2020; Ghosh et al., 2020; Li et al., 2021; Marfoq et al., 2021].

Many personalized FL strategies involve grouping clients into homogeneous clusters to then tailor a model for each cluster [e.g., Ding and Wang, 2022; Ghosh et al., 2020; Sattler et al., 2021]. The optimal clustering would ideally group clients whose local optimal models are similar. However, the inherent challenge lies in the fact that these optimal models are unknown a priori. This dilemma results in the two tasks—model learning for the specified task and cluster identification—being deeply interconnected. Various studies have suggested empirical measures of similarity as a workaround [e.g., Ghosh et al., 2020; Sattler et al., 2021], while others rely on presumed knowledge of distances across data distributions [e.g. Ding and Wang, 2022; Even et al., 2022]. Nonetheless, accurately estimating these distances, especially within a FL framework where clients may possess limited data, proves to be particularly challenging. Such estimation difficulties are well discussed in the literature (e.g., in Even et al. [2022] [Sec. 6]), underscoring the problem of identifying similar clients for collaborative model learning as a significant, yet unresolved issue.

In this paper, we narrow our focus to a fundamental aspect of the broader challenge, specifically aiming to estimate the mean of a real-valued distribution, considered as the archetypal federated learning problem in many papers [Dorner et al., 2024; Grimberg et al., 2021; Tsoy et al., 2024]. This task 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]. Importantly, the task of personalized mean estimation serves as a foundational problem that can illuminate key aspects of personalized federated learning.

We are interested in an online decentralized version of personalized mean estimation, where at each time slot, a client collects new samples and can exchange information with a limited number of other clients. To the best of our knowledge, the state-of-the-art method in this setting is the Collaborative Mean Estimation algorithm, CollME, by Asadi et al. [2022]. Unfortunately, CollME struggles with scalability in large systems, as its per-agent space and time complexities are both linear in the number of clients $|\mathcal{A}|$.

To address these scalability challenges, we propose that clients self-organize into a network where each client communicates with at most a fixed set of r neighbors. Over time, this set of neighbors is pruned, as clients progressively exclude those that are more dissimilar. Within this context, we introduce two collaborative mean estimation algorithms: one uses a message-passing scheme and the other is based on consensus. The complexities of these algorithms are $O(r \cdot \log |\mathcal{A}|)$ and $O(r)$, respectively. We demonstrate that, even though each client exchanges information with only $r \ll |\mathcal{A}|$ clients, it is possible to achieve with high probability a speedup in the convergence time of mean estimates by a factor of $\Omega(|\mathcal{A}|^{1/2-\phi})$, with ϕ arbitrarily close to 0.

2 Related Work

For an overview of the field of personalized federated learning, we refer the reader to the recent survey Tan et al. [2023]. Here, we limit to mention the most relevant approaches for this paper.

Ghosh et al. [2020] and Sattler et al. [2021] were the first to propose clustered FL algorithms, which split the clients on the basis of the similarity of their data distributions; the similarity is evaluated empirically by the Euclidean distance of the 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 et al. [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 et al. [2022] prove that collaboration can at most speed up the convergence time linearly in the number of similar agents and provide algorithms, which, under apriori 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 on CoLME

We consider a set $\mathcal{A}$ of agents (computational units). Each agent $a \in \mathcal{A} = \{1, 2, \cdots, |\mathcal{A}|\}$ generates at each time instant t a new sample $x_a^t \in \mathbb{R}$ drawn i.i.d. from distribution D_a with expected value $\mu_a = \mathbb{E}[x_a^t]$. Expected values are not necessarily distinct across agents. Given two agents a and a', we denote by $\Delta_{a,a'} := |\mu_a - \mu_{a'}|$ the gap between the agents' *true* means, and with C_a, the group of agents with the same *true* mean of a, (i.e. such that $\Delta_{a,a'} = 0$). In the following we will refer to C_a as the 'similarity class of a.'

The goal of each agent $a \in \mathcal{A}$ is to estimate μ_a. To this purpose, the client can compute its *local* average estimate $\bar{x}_{a,a}^t = \frac{1}{t}\sum_1^t x_a^{t'}$, but can also obtain a more accurate estimates taking advantage of information from other agents in its similarity class.

Asadi et al. [2022] proposed CoLME as a collaborative algorithm for mean estimation. CoLME relies on two key steps: i) trying to identify the similarity classes; ii) improving local mean estimates by sharing information with agents believed to belong to the same class.

1) Identifying Similarity Classes. We denote by C_a^t the set of agents that agent a estimates, at time t, to be part of its own similarity class C_a. Agent a decides that agent a' belongs to the same class if their local average estimates are sufficiently close, specifically, if two appropriately defined confidence intervals centered on these estimates overlap. In general, at time t, agent a will not have access to the most recent local average estimates of agent a', but to a stale value, corresponding to the last time the two clients have exchanged information. We denote such value as $\bar{x}_{a,a'}^t$ and the corresponding number of samples as $n_{a,a'}^t$[1]. Agent a can then estimate its true mean and the true mean of agent a' to belong to the confidence intervals $I_{a,a} = \left[\bar{x}_{a,a}^t - \beta_\gamma(t), \bar{x}_{a,a}^t + \beta_\gamma(t)\right]$, $I_{a,a'} = \left[\bar{x}_{a,a'}^t - \beta_\gamma(n_{a,a'}^t), \bar{x}_{a,a'}^t + \beta_\gamma(n_{a,a'}^t)\right]$, 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 assume that a' belongs to its similarity class C_a if the two intervals overlap, i.e., $I_{a,a} \cap I_{a,a'} \neq \emptyset$, or equivalently if:

$$d_\gamma^t(a, a') := |\bar{x}_{a,a} - \bar{x}_{a,a'}| - \beta_\gamma(t) - \beta_\gamma(n_{a,a'}^t) < 0. \tag{1}$$

In order to estimate its similarity class in the most accurate way, agent a could retrieve at eath time t the most recent estimate $\bar{x}_{a',a'}^t$ from each other agent a'. This, however, would lead to a quadratic communication burden $\Theta(|\mathcal{A}|^2)$. To limit such a burden, node a cyclically queries a single node in C_a^{t-1} at each time instant t, according to a *Round-Robin* scheme. This leads to the following update rule for C_a^t:

$$C_a^t = \{a' \in C_a^{t-1} : d_\gamma^t(a, a') \leqslant 0\}, \tag{2}$$

and we observe that by construction $C_a^t \subseteq C_a^{t-1}$. Initially, the agent sets $C_a^0 = \mathcal{A}$ and progressively takes irreversible decisions to remove agents.

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

$$\hat{\mu}_a^t = \sum_{a' \in C_a^t} \frac{n_{a,a'}^t}{\sum_{a' \in C_a^t} n_{a,a'}^t} \bar{x}_{a,a'}^t$$

We observe that both in terms of memory and per-time computations, the complexity of CoLME is linear, i.e., $\Theta(|\mathcal{A}|)$, since every agent a must keep in memory all local estimates $\bar{x}_{a,a'}$ and number of samples $n_{a,a'}^t$. And, at every t, for every a', it has to recompute $d_\gamma^t(a, a')$, using the last available value for $\bar{x}_{a,a}$ (and $n_{a,a}^t$). Indeed, note that the local value of $\bar{x}_{a,a}$ is updated at every t and this triggers a change in the optimistic distance for every t.

3.1 CoLME's Guarantees

We provide the the two main theoretical results from Asadi et al. [2022] which were proved only when the distributions

[1] Note that $n_{a,a'}^t \leqslant t$ coincides with the last time-slot at which the two agents have communicated.

$\{D_a, a \in \mathcal{A}\}$ are sub-Gaussians[2].

Select the amplitude of the confidence intervals as:

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

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 $\Delta_a := \min_{a' \in \mathcal{A}\setminus C_a} \Delta_{a,a'}$; 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 1) 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, $C^t_a = C_a$, with high probability.[3]

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

$$\mathbb{P}\left(\exists t > \zeta_a : C^t_a \neq C_a\right) \leqslant \frac{\delta}{2}, \tag{4}$$

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

Theorem 1 shows that the time ζ_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 Δ_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 $\zeta_a \in O(|\mathcal{A}|)$ 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 2. *[Asadi et al., 2022, Theorem 2] Assume distributions $\{D_a, \forall a \in \mathcal{A}\}$ are sub-Gaussians with parameter σ^2. For any $\delta \in (0,1)$, and with $\gamma = \frac{\delta}{4|\mathcal{A}|}$, employing CoLME, we have:*

$$\mathbb{P}\left(\forall t > \tau_a, |\hat{\mu}^t_a - \mu_a| < \varepsilon\right) \geqslant 1-\delta, \tag{5}$$

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

Theorem 2 admits a straightforward explanation. Provided that agent a has successfully estimated its similarity class at time t (i.e., $C^t_a = C_a$), the error in the mean estimation will depend only on the available number of samples of agents in 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 $t \geqslant t^*_a = \frac{n^\star_{\frac{\delta}{2}}(\varepsilon)}{|C_a|} + \frac{|C_a|-1}{2}$ 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 $C^t_a = C_a$ and $|\hat{\mu}^t_a - \mu_a| \leqslant \varepsilon$ hold.

[2] A distribution D is sub-Gaussian with parameter σ^2 if $\forall \lambda \in \mathbb{R}$, $\log \mathbb{E}_{x\sim D} \exp(\lambda(x-\mu)) \leqslant \frac{1}{2}\lambda^2\sigma^2$.

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

3.2 Extensions

We extend Theorem 1 and 2 to distributions that are not sub-Gaussians as far as the width of the confidence intervals $\beta_\gamma(\cdot)$ satisfies the following assumption:

Assumption 1. *There exists a function $\beta_\gamma(\cdot) \in o(1)$ such that the true mean belongs to all (nested) intervals of amplitude $\pm\beta_\gamma(t)$ for $t \in \mathbb{N}$ with confidence $1-2\gamma$, i.e.,*

$$\mathbb{P}\left(\forall t \in \mathbb{N}, \left|\bar{x}^t_{a,a} - \mu\right| < \beta_\gamma(t)\right) \geq 1-2\gamma, \forall a \in \mathcal{A}. \tag{6}$$

For generic distributions $\{D_a, a \in \mathcal{A}\}$, it is difficult to find such function $\beta_\gamma(\cdot)$, but for sub-Gaussian distributions Eq. (3) satisfies indeed Assumption 1 Maillard [2019].

In Appendix A, we show that Assumption 1 is satisfied when the distributions $\{D_a, \forall a \in \mathcal{A}\}$ have variance upper-bounded by σ^2, and fourth (central) moment upper-bounded by $\kappa\sigma^4$ (when all the variables are identically distributed κ can be interpreted as the kurtosis and σ as the standard deviation). In particular, in this case β_γ, can be selected as follows:

$$\beta_\gamma(n) = \left(2\frac{\kappa+3\sigma^4}{\gamma}\right)^{\frac{1}{4}}\left(\frac{1+\ln^2 n}{n}\right)^{\frac{1}{4}}, \tag{7}$$

If the distributions $\{D_a, \forall a \in \mathcal{A}\}$ exhibit a larger number of bounded polynomial moments, a more favorable expression for $\beta_\gamma(n)$ can be obtained (see Remark 1 in Appendix A).

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

The main limitations of CoLME are acknowledged by Asadi et al. [2022]: each agent needs to perform a number of operations and maintain a memory proportional to $|\mathcal{A}|$, leading to total time and space costs which are $O(|\mathcal{A}|^2)$ per timeslot. This is impractical in large-scale systems. Moreover, while the *Round-Robin* query scheme reduces the communication burden, it entails a significant delay in the estimation ($\zeta_a \in O(|\mathcal{A}|)$). Total space and time computational complexity can be made $\tilde{O}(1)$ with our *scalable* approaches: B-CoLME (see Section 4.1) and C-CoLME (see Section 4.2). Both approaches consider agents $\mathcal{A}$ over a fixed graph $\mathcal{G}(\mathcal{A}, \mathcal{E})$ and restrict the communication to pairs of agents adjacent in $\mathcal{G}$. We denote by CC_a the maximal connected component consisting of nodes in C_a to which a belongs, and by CC^d_a the agents in C_a which are at most d hops away from a. Similarly, 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 by a, respectively.

Each agent a aims to determine which nodes in its neighborhood $\mathcal{N}_a$ belong to its similarity class C_a. To this purpose, agent a receives at time t from each neighbor $a' \in \mathcal{N}_a$ an updated version of its local mean estimate. We denote with $C^t_a \subseteq \mathcal{N}_a$ the

set of neighbors that agent a deems to belong to its own similarity class at time t. At the beginning, $C_a^0 = \mathcal{N}_a$. Similarly to CoLME, at each time t, agent a first computes the distance $d_\gamma^t(a, a')$ for every $a' \in C_a^{t-1}$ according to (1) and then updates C_a^t according to (2). As for CoLME, therefore, $C_a^t \subseteq C_a^{t-1}$, as soon as a removes a' from C_a^t, it stops communicating with a'. Then, over time, communications occur over a pruned graph $\mathcal{G}^t = (\mathcal{A}, \mathcal{E}^t)$, where $\mathcal{E}^t = \{(a, a') \in \mathcal{E} : a' \in C_a^t\}$.

We aim to determine the time needed for all agents in the connected component CC_a to identify their neighbors, i.e., $C_{a'}^t = C_{a'} \cap \mathcal{N}_{a'}, \forall a' \in CC_a$. Following the same approach as in [Asadi et al., 2022, Theorem 1] (reported above as Theorem 1), we can prove[4] that:

Theorem 3. *For any $\delta \in (0, 1)$, employing either B-CoLME or C-CoLME we have:*

$$\mathbb{P}\left(\exists t > \zeta_a^D, \exists a' \in CC_a : C_{a'}^t \neq C_{a'} \cap \mathcal{N}_{a'}\right) \leqslant \frac{\delta}{2}, \tag{8}$$

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

Proof. With respect to the sample mean $\bar{x}$, choosing $\beta_\gamma(n)$ as in Eq. (3) or (7) guarantees that $\mathbb{P}(\exists n \in \mathbb{N}, |\bar{x} - \mu| \leq 2\gamma)$, i.e., the probability of the true mean being outside the confidence interval is bounded above by 2γ (Lemma 10). Applying a union bound over $a' \in CC_a$ and hence taking $\gamma = \frac{\delta}{4r|CC_a|}$, we ensure (Lemma 12) $\mathbb{P}\left(\exists a' \in CC_a, \exists t \in \mathbb{N}, \exists a'' \in \mathcal{N}_{a'} : \left|\bar{x}_{a',a''}^t - \mu_{a''}\right| > \beta_\gamma(n_{a',a''}^t)\right) < \frac{\delta}{2}$. Conditioned on the complementary event, it is rather immediate to prove $d_\gamma^t(a, a') > 0 \iff a' \notin \mathcal{N}_a \cap C_a$ (see Appendix). After $n_\gamma^\star(a) := \lceil \beta_\gamma^{-1}(\Delta_a/4) \rceil$, it holds that $d_\gamma^t > 0$ for all pairs (a, a') with a' in a different similarity class, and $d_\gamma^t \leq 0$ otherwise. The similarity class to which a belongs has been correctly identified. □

When comparing Theorem 3 with Theorem 1, we observe that for large systems, if $r|CC_a| \in \Theta(|\mathcal{A}|)$, $\zeta \approx |\mathcal{A}| + \zeta_a^D$, showing that, as expected, agents can identify much faster similar agents in their neighborhood than in the whole population $\mathcal{A}$.[5] A more detailed comparison of CoLME, B-CoLME, and C-CoLME is in Sec. 5.

We now present two decentralized algorithms that allow agents to improve the estimates of their true means.

4.1 Message-passing B-CoLME

In B-CoLME algorithm, each node $a \in \mathcal{A}$, through a continuous exchange of messages with its direct neighbors C_a^t, can acquire not only their local estimates $\{\bar{x}_{a,a'}^t, a' \in C_a^t\}$, but also, in an aggregate form, estimates from further away nodes up to distance d in the graph $\mathcal{G}^t$ (d is a tunable parameter, the need to introduce it will be explained soon). Indeed, node $a' \in C_a^t$ serves as a *forwarder*, providing to its neighbor $a \in \mathcal{N}_{a'}$ access to the records it has collected from *other* neighbors $a'' \in C_a^t \setminus \{a\}$. As far as each agent has correctly identified all similar nodes in its neighborhood, the agent potentially has access to the local estimates of all agents in CC_a^d.

According to our *message-passing* scheme, agent $a \in \mathcal{A}$ retrieves at time t a message $M^{t,a'\to a}$ from every of its neighbors $a' \in C_a^t$ (for a visual representation see Fig. 4 in the Appendix). Message $M^{t,a'\to a}$ is a $d \times 2$ table whose element $m_{h,1}^{t,a'\to a}$ contains a sum of samples, while $m_{h,2}^{t,a'\to a}$ indicates the total number of samples contributing to the aforementioned sum. In particular, at time t, the first table entries are set as: $m_{1,1}^{t,a'\to a} = \sum_{\tau=1}^t x_{a'}^\tau$ and $m_{1,2}^{t,a'\to a} = t$. The other entries are built through the following recursion:

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

for $h \in \{2, \dots, d\}$ and $i \in \{1, 2\}$.

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:

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

and, under the local tree structure assumption, this corresponds to performing an empirical average over all the samples generated by agent a up to t and all agents in $\mathcal{G}^t$ at distance $1 \leqslant h \leqslant d$ up to $t - h$.

If $\mathcal{G}^t \cap \mathcal{N}_a^d$ is not a tree, then the contribution of the samples collected by a given agent a'' may be included in messages arriving to a through different neighbors in C_a^t and then been erroneously counted multiple times in (9). The parameter d needs to be selected to prevent this eventuality (or make it very unlikely), as we discuss in Sec. 4.3.

As a result of previous arguments, we obtain the following results:

Theorem 4. *. Provided that CC_a^d is a tree, for any $\delta \in (0, 1)$, employing B-CoLME, we have:*

$$\mathbb{P}\left(\forall t > \tau_a',\ |\hat{\mu}_a^t - \mu_a| < \varepsilon\right) \geqslant 1 - \delta$$
$$\text{with} \quad \tau_a' = \max\left[\zeta_a^D + d,\ \frac{\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)}{|CC_a^d|} + d\right],$$

$$\widetilde{n}_{\frac{\delta}{2}}(\varepsilon) \leqslant \begin{cases} \left\lceil -\frac{2\sigma^2}{\varepsilon^2} \ln\left(\frac{\delta}{4}\left(1 - e^{-\frac{\varepsilon^2}{\sigma^2}}\right)\right) \right\rceil & \text{sub-Gaussian,} \\ \left\lceil \frac{2(\kappa+3)\sigma^4}{\delta\varepsilon^4} \right\rceil & \text{otherwise.} \end{cases}$$

Proof. The first step consists in evaluating $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)$, defined as the minimum number of i.i.d samples extracted from D_a, which guarantees that estimate error $|\bar{x}_{a,a}^t - \mu_a|$ is definitively smaller than ε with a probability larger por equal than $1 - \delta/2$. Specifically, we require $\sum_{t=\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)}^\infty \mathbb{P}(|\bar{x}_{a,a}^t - \mu_a| < \varepsilon) > 1 - \delta/2$. Once $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)$ has been computed, the claim can be easily proved. Indeed under the assumption that CC_a^d has been correctly identified a lower bound to the minimum time at which $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)$ i.i.d

[4] Full proofs of this and the following results are in the supplemental material.

[5] 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, $\zeta \approx |\mathcal{A}|/r + \zeta_a^D$ and the conclusion does not change.

Algorithm 1 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 $\mathbb{P}\left(|\hat{\mu}_a - \mu_a| < \varepsilon\right) \geqslant 1-\delta$
$C_a^0 \leftarrow \mathcal{N}_a, \forall a \in \mathcal{A}$
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. (7)
  for neighbor a' in 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**
      $C_a^t \leftarrow C_a^t \setminus \{a'\}$
    end if
  end for
  for neighbor a' in C_a^t **do**
    Compute $M^{t,a\to a'}$ and send it to a'
  end for
  Wait for messages $M^{t,a'\to a}$ $\forall a' \in C_a^t$
  $\hat{\mu}_a^t \leftarrow \frac{\sum_{\tau=1}^t x_a^\tau + \sum_{a'\in C_a^t}\sum_{h=1}^d m_{h,1}^{t-1,a'\to a}}{t + \sum_{a'\in C_a^t}\sum_{h=1}^d m_{h,2}^{t,a'\to a}}$
end for

samples extracted from D_a are available at agent a is given by $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)$ $\frac{}{|CC_a^d|} + d$. At last observe that for $t \geqslant \zeta_a^D + d$, CC_a^d has been correctly identified with probability larger than $1-\delta/2$, by Theorem 3. The claim follows easily by applying the union bound to the probability of event $\{\exists t \geqslant \tau_a' : |\mu_a^t - \mu_a| > \varepsilon\}$. □

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

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

Now Theorem 23 (in the Appendix) provides upper bounds to δ' when $\mathcal{G}$ is a random regular graph. In particular, as long we set d as in Proposition 8, δ' converges to 0 as the number of agents $|\mathcal{A}|$ increases. Our proofs can readily be adapted to CoLME, enabling to substitute $n^\star_{\frac{\delta}{2}}(\varepsilon)$ in Theorem 2 with the smaller term $\widetilde{n}_{\frac{\delta}{2}}(\varepsilon)$.

4.2 Consensus-based C-CoLME

In this section, we present a second possible collaborative approach for estimating the mean, see Algorithm 2. It takes inspiration from consensus algorithms in dynamic settings, as in [Franceschelli and Gasparri, 2019; Montijano et al., 2014]. The basic idea is that each agent maintains two metrics, namely the empirical average of its local samples $\bar{x}_{a,a}^t$, and the "*consensus*" estimate $\hat{\mu}_a^t$, which is updated at time t by computing a convex combination of the local empirical average and a weighted sum of the consensus estimates in its (close) neighborhood $\{\hat{\mu}_{a'}^{t-1}, a' \in C_a^{t-1} \cup \{a\}\}$.

The dynamics of all estimates are captured by:

$$\hat{\boldsymbol{\mu}}^{t+1} = (1-\alpha_t)\,\bar{\mathbf{x}}^{t+1} + \alpha_t W_t \hat{\boldsymbol{\mu}}^t, \tag{10}$$

where $(W_t)_{a,a'} = 0$ if $a' \notin C_a^t$ and $\alpha_t \in (0,1)$ is the memory parameter. Once the agents have discovered which neighbors belong to the same class at time τ, 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 $a' \in C_a \cap \mathcal{N}_a$. 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.[6] By time τ, 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

$$W = \begin{pmatrix} {}_1W & 0_{n_1\times n_2} & \cdots & 0_{n_1\times n_C} \\ 0_{n_2\times n_1} & {}_2W & \cdots & 0_{n_2\times n_C} \\ \cdots & \cdots & \cdots & \cdots \\ 0_{n_C\times n_1} & 0_{n_C\times n_2} & \cdots & {}_CW \end{pmatrix}, \tag{11}$$

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.

To obtain deeper insights into C-CoLME and to support the proof of our (ϵ,δ) convergence results, it is beneficial to consider an auxiliary system. In this system, the consensus matrix may be arbitrary up until a given time $\tilde{\tau}$, after which it transitions to W.

Algorithm 2 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 $\mathbb{P}\left(|\hat{\mu}_a - \mu_a| < \epsilon\right) \geqslant 1-\delta$
$C_a^0 \leftarrow \mathcal{N}_a, \forall a \in \mathcal{A}$
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. (7)
  for neighbor a' in $\mathcal{N}_a \cap C_a^{t-1}$ **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
  $C_a^t \leftarrow \left\{a' \in \mathcal{N}_a \cap C_a^{t-1} \text{ s.t. } d_\gamma^t(a,a') \leqslant 0\right\}$
  $\hat{\mu}_a^t \leftarrow (1-\alpha_t)\bar{x}_a^t + \alpha_t \sum_{a'\in C_a\cup\{a\}}(W_t)_{a,a'}\hat{\mu}_{a'}^{t-1}$
end for

Theorem 6. *Consider a system which evolves according to* (10) *with $W_t = W$ in* (11)*, for $t \geqslant \tilde{\tau}$. With probability at least $1-\delta/2$, for $\alpha_t = \alpha$, it holds:*

$$\begin{aligned}\mathbb{E}\left[\|{}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu}\|^4\right] &\in O\left(\sup_{W_1,\cdots,W_{\zeta_D}} \mathbb{E}\left[\|{}_c\hat{\boldsymbol{\mu}}^{\zeta_D} - {}_c\boldsymbol{\mu}\|^4\right]\alpha^{4t}\right) \\ &+ O\left(\frac{(1-\alpha)^4\cdot\left(1-1/\ln(\alpha\lambda_{2,c})\right)^2}{(1-\alpha\lambda_{2,c})^2}\,\frac{\mathbb{E}\left[\|{}_c\mathbf{x} - {}_cP_{\,c}\mathbf{x}\|^4\right]}{(t+1)^2}\right) \\ &+ O\left((1-\alpha)^2\left(1-\frac{1}{\ln\alpha}\right)^2\frac{\mathbb{E}\left[\|{}_cP_{\,c}\mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]}{(t+1)^2}\right),\end{aligned} \tag{12}$$

[6] This can be obtained in a distributed way, e.g., setting $\forall a' \in C_a^t$ $(W_t)_{a,a'} = (\max\{|C_a^t|, |C_{a'}^t|\}+1)^{-1}$; $(W_t)_{a,a} = 1 - \sum_{a'\in C_a^t}(W_t)_{a,a'}$.

and, for $\alpha_t = \frac{t}{t+1}$, it holds:

$$\mathbb{E}\left[\|{}_c\hat{\boldsymbol{\mu}}^{t+1} - {}_c\boldsymbol{\mu}\|^4\right] \in O\left(\sup_{W_1,\cdots,W_{\zeta_D}} \frac{\mathbb{E}\left[\|{}_c\hat{\boldsymbol{\mu}}^{\zeta_D} - {}_c\boldsymbol{\mu}\|^4\right]}{(t+1)^4}\right)$$
$$+ O\left(\frac{\left(1-1/\ln\lambda_{2,c}\right)^2}{(1-\lambda_{2,c})^2}\frac{\mathbb{E}\left[\|{}_c\mathbf{x} - {}_cP\,{}_c\mathbf{x}\|^4\right]}{(t+1)^4}\right)$$
$$+ O\left(\mathbb{E}\left[\|{}_cP\,{}_c\mathbf{x} - {}_c\boldsymbol{\mu}\|^4\right]\left(\frac{1+\ln t}{1+t}\right)^2\right). \tag{13}$$

Proof. The presence of the memory term and the requirement to consider convergence of the fourth moment (rather than the second) prevent the direct application of existing results for the convergence of the consensus algorithm. The proof begins by decomposing the fourth moment into three terms. The first term is dependent on on ${}_c\hat{\boldsymbol{\mu}}^{\zeta_D} - {}_c\boldsymbol{\mu}$, that is the estimate errors at time ζ_D. The second term involves the differences ${}_cW^{t-\tau}({}_c\bar{\mathbf{x}}^{\tau+1} - {}_cP_c\bar{\mathbf{x}}^{\tau+1})$, reflecting the effectiveness of consensus averaging—specifically, how well consensus reduces the gap between local estimates ${}_c\bar{\mathbf{x}}$ and their empirical averages ${}_cP_c\bar{\mathbf{x}}$. This terms is minimized when ${}_cW = {}_cP$, which is, as expected, the best consensus matrix. Finally, the third term depends on ${}_cP_c\bar{\mathbf{x}}^{\tau+1} - {}_c\boldsymbol{\mu}$ and represents the minimum error achievable by averaging the estimates of all agents in the component. The remainder of the proof focuses on bounding these three terms using the stochasticity and symmetry of the matrix ${}_cW$. A recurrent key step in the proof is bounding sums of the kind $\sum_{\tau=t_0}^{t}\frac{\beta^{t-\tau}}{\tau+1}$, which we demonstrate to be $O\left((1-1/\ln\beta)\frac{1}{t+1}\right)$ in Lemma 15. It is then shown that the first two terms are $O\left(1/t^4\right)$, while the final term determines the asymptotic behavior. □

Theorem 6 also reveals that the choice of the memory parameter α_t affects differently the three terms. In particular, the speedup proportional to the size of the connected component $|CC_a| = n_c$, as observed for B-CoLME in Theorem 4, is not achieved with a constant $\alpha_t = \alpha$. However, it becomes attainable when $\alpha_t = \frac{t}{t+1}$, as it is detailed in the subsequent theorem:

Theorem 7. *Consider a graph component c and pick uniformly at random an agent a in c. Let $g(x) := x\ln^2(ex)$ and $\alpha_t = \frac{t}{t+1}$. It holds:*

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

where $\tau_a^{''} = \max\left\{\zeta_a^D, g\left(C\frac{\mathbb{E}\left[\|{}_cP\,{}_c\mathbf{x}-{}_c\boldsymbol{\mu}\|^4\right]}{|CC_a|\epsilon^4\delta}\right) \in \tilde{O}\left(\frac{\tilde{n}_{\frac{\delta}{2}}(\varepsilon)}{|CC_a|}\right)\right\}$.

Proof. Due to Theorem 3, for $t \geq \zeta_a^D$, the fourth moment of the estimate error within component c can be bounded as in Theorem 6 w.p. at least $1-\delta/2$. Subsequently, we apply Markov inequality to establish a probabilistic bound at a specific time t. By employing a union bound over t and suitably majorizing the resultant series, we achieve the final result. □

The supplementary material (Sec. D.3) presents convergence results for the case where $\alpha_t = \alpha$, a setting that does not enjoy the same speedup factor.

4.3 Choice of the Graph and other Parameters

The selection of the graph $G(\mathcal{A},\mathcal{E})$ plays a pivotal role in the efficacy of our algorithms. We discuss key desirable properties in $G(\mathcal{A},\mathcal{E})$. First, Theorems 4 and 7 show that learning timescales, τ_a' and τ_a'', decrease as the size of collaborating agent groups, CC_a^d and CC_a, increase. Thus, a highly connected graph is preferred to foster the formation of large class-homogeneous agent groups after the disconnection of inter-class edges. Second, the complexities—both spatial and temporal—of B-CoLME and C-CoLME are directly proportional to the agent's degree within the graph. We would like then the degree to be small and possibly uniform across the agents to balance computation across agents. We observe that these two desiderata can be at odds, as a higher degree generally leads to larger groups CC_a^d and CC_a. 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.

Taking these factors into account, 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. For instance, for any $r \geqslant 3$, the probability that the sampled graph is connected approaches one as N increases. Additionally, the sample graph demonstrates a local tree-like structure with high probability, as discussed in Appendix E. The selection of r illustrates the trade-off mentioned above: a lower r reduces the computational and memory load on each agent, but may also lead to smaller the expected sizes of CC_a^d and CC_a, affecting the accuracy of estimates.

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 CC_a^d, for a randomly chosen $a \in \mathcal{A}$, has a tree structure with high probability. However, choosing a d that is too small could unnecessarily restrict the size of CC_a^d, thereby undermining the effectiveness of the estimation process (Theorem 4). For a comprehensive analysis on how the parameters r and d influence both the structure of $\mathcal{N}_a^d$ and the size of CC_a^d, we direct the reader to Appendix E. In this section, we informally summarize the main result:

Proposition 8. *By selecting $d = \left\lfloor \frac{1}{2}\log_{r-1}\frac{|\mathcal{A}|}{\log_{r-1}||A|} \right\rfloor$ 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 $r \in \Theta(\log(1/\delta))$, $|CC_a^d|$ (with d chosen as above) is in $\Omega(|\mathcal{A}|^{\frac{1}{2}-\phi})$ for any arbitrarilyy 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 6. 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: $(W_t)_{a,a'} = (\max\{|C_a^t|,|C_a^t|\}+1)^{-1}$ for all $a' \in C_a^t$ and $(W_t)_{a,a} = 1 - \sum_{a'\in C_a^t}(W_t)_{a,a'}$.

	Per-agent space/time complexity	Convergence time	
		sub-Gaussian	bounded 4-th moment
ColME (r comms)	$\lvert\mathcal{A}\rvert$	$\frac{1}{\Delta_a^2}\log\frac{\lvert\mathcal{A}\rvert}{\Delta_a\delta}+\frac{\lvert\mathcal{A}\rvert}{r}+\frac{1}{\lvert C_a\rvert}\frac{1}{\varepsilon^2}\log\frac{1}{\delta\varepsilon^2}$	$\frac{1}{\Delta_a^4}\frac{\lvert\mathcal{A}\rvert}{\delta}+\frac{\lvert\mathcal{A}\rvert}{r}+\frac{1}{\lvert C_a\rvert}\frac{1}{\delta\varepsilon^4}$
B-ColME	rd	$\frac{1}{\Delta_a^2}\log\frac{\lvert CC_a\rvert r}{\Delta_a\delta}+d+\frac{1}{\lvert CC_a^d\rvert}\frac{1}{\varepsilon^2}\log\frac{1}{\delta\varepsilon^2}$	$\frac{1}{\Delta_a^4}\frac{\lvert CC_a\rvert r}{\delta}+d+\frac{1}{\lvert CC_a^d\rvert}\frac{1}{\delta\varepsilon^4}$
C-ColME ($\alpha_t=\frac{t}{t+1}$)	r	–	$\frac{1}{\Delta_a^4}\frac{\lvert CC_a\rvert r}{\delta}+\frac{1}{\lvert CC_a\rvert}\frac{1}{\delta\varepsilon^4}$

Table 1: Comparison of collaborative estimation algorithms. The convergence time is provided in order sense.

5 Algorithms' Comparison

Table 1 presents a comparative analysis of the three algorithms: ColME, B-ColME, and C-ColME. For a fair comparison, we consider a variant of ColME, where each agent is allowed to communicate with r agents at each time t. This adjustment guarantees that all three algorithms incur the same communication overhead.

The second column in Table 1 details the space and time complexities for each algorithm. Remarkably, even when r and d are allowed to increase logarithmically with the number of agents $|\mathcal{A}|$, B-ColME maintains its efficiency advantage over ColME. Moreover, C-ColME demonstrates further improvements, reducing the per-agent burden beyond the savings achieved by B-ColME.

The third and fourth columns detail the characteristic times required to achieve (ϵ,δ) convergence for the estimates generated by the three algorithms when local data distributions are sub-Gaussian and when they simply exhibit bounded fourth moments. The characteristic times correspond to τ_a, τ'_a, and τ''_a in Theorems 2, 4, and 7, respectively. The table reports their asymptotic behavior as the number of agents $|\mathcal{A}|$ increases ignoring logarithmic factors. The detailed derivations and analyses that underpin these results are provided in Appendix F.

Three factors contribute to the characteristic times. The first factor corresponds to the time required to correctly identify the potential collaborators. In the case of ColME, this involves each agent classifying the other $|\mathcal{A}|-1$ agents, leading to a term that scales as $\log|\mathcal{A}|$ or $|\mathcal{A}|$, contingent upon the assumed properties of the local distribution. For B-ColME and C-ColME, $|\mathcal{A}|$ is replaced by $|CC_a|r$, representing an upper bound on the number of connections agents in CC_a may have initially established with agents from different classes. This substitution is not immediately intuitive, as one might initially anticipate the relevant scale to be simply r. However, this adjustment accounts for the potential ripple effect of classification errors: a mistake by any agent a can adversely affect the estimates of all agents within the same connected component CC_a.

The second factor contributing to the characteristic times is the time each agent needs to collect all relevant information. In the case of 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. We observe that the corresponding term does not appear for C-ColME, because it is dominated by the final term.

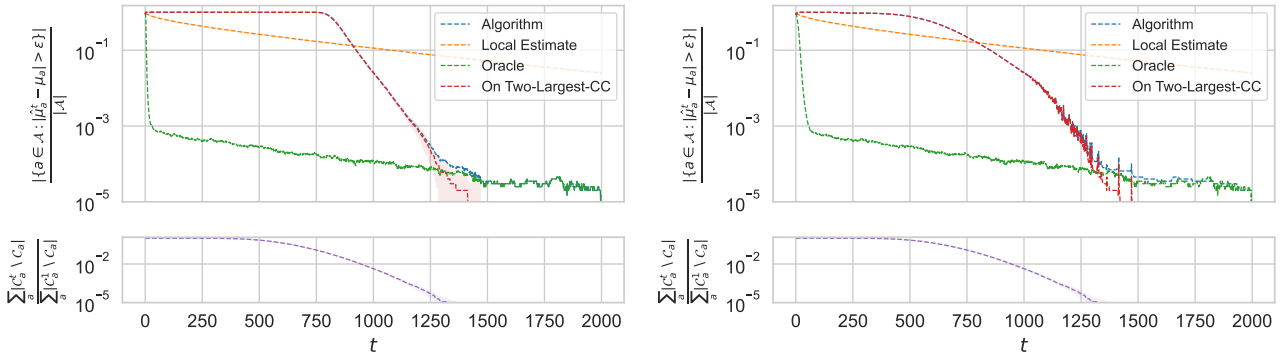
The final term pertains to the period necessary for achieving accurate mean estimation once the collaborators are identified and showcases the benefits of collaboration. In ColME, the collaboration's benefit is particularly striking, as all agents within the same class unite to improve their estimates. This collective effort effectively reduces the convergence time by a factor proportional to the size of the collaborating group, $|C_a|$. For B-ColME and C-ColME, although the speed-up remains proportional to the number of collaborating agents, the actual numbers of collaborators, $|CC_a^d|$ for B-ColME and $|CC_a|$ for C-ColME, are in general smaller.

In conclusion, while ColME potentially offers the most accurate estimates, it does so with longer convergence times and greater demands on memory and computation for each agent. In contrast, B-ColME and C-ColME present more efficient alternatives, achieving faster convergence with reduced per-agent resource requirements. However, this efficiency may come at the expense of the maximum attainable accuracy. In the next section, we quantify this trade-off experimentally.

6 Numerical Experiments

We evaluate the performance of the proposed algorithms over the class $\mathcal{G}_0(N,r)$ of simple regular graph, which models a situation where an agent is connected to r other agents selected uniformly at random and provides the tree-like structure needed for the B-ColME as discussed in Sec. 5. The agents can belong to one of two *similarity* classes, associated with Gaussian distributions ${}_1D\sim\mathcal{N}({}_1\mu=0,\sigma^2=4)$ and ${}_2D\sim\mathcal{N}({}_2\mu,\sigma^2=4)$. At initialization, each node is assigned one of the two classes with equal probability. Unless otherwise mentioned, in the experiments $|\mathcal{A}|=N=10000$, $r=10$, $d=4$, $\varepsilon=0.1$, $\delta=0.1$, and $\beta_\gamma(n)$ as per Eq. (3). In Appendix G we provide additional experiments varying the system's parameters.

Figure 1 showcases the performance of B-ColME and C-ColME by illustrating two key metrics: the fraction of agents with wrong estimates (more than ε away from the true mean), and the fraction of wrong edges still used (a wrong edge is an edge connecting two agents belonging to a different class). We compare our algorithms against two benchmarks. In the first scenario, each agent independently relies on its *local* estimate, denoted as $\tilde{x}^t_{a,a}$. In the second scenario, every agent a is endowed with an *oracle*. This oracle precisely identifies which neighbors are in the same class as the agent, signified by $C^t_a=C_a,\forall t$. The figure reveals that B-ColME has a longer transient phase but then exhibits a slightly steeper convergence than C-ColME. Notably, there is no apparent improvement in B-ColME's estimates until about 90% of the incorrect edges 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 esti-



(a) B-CoLME (b) C-CoLME

Figure 1: Fraction of agents whose estimates deviate from the true value by more than ϵ (top) and a fraction of neighbors that have not yet been identified as belonging to a different class (bottom) for B-CoLME (left) and C-CoLME (right). Averages and 95% confidence intervals were computed over 20 realizations.

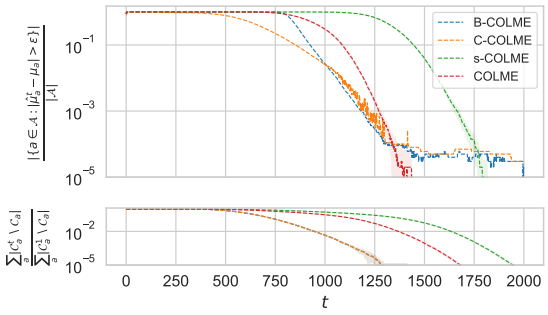


Figure 2: Comparison of our algorithms and the two versions of CoLME, on a $G_0(N = 10000, r = 10)$ over 10 realizations.

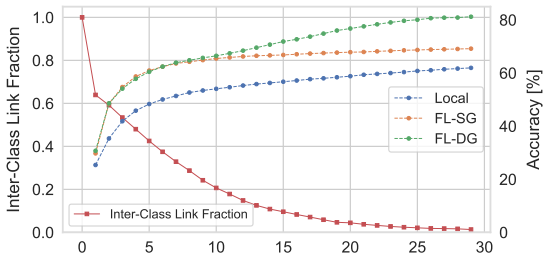


Figure 3: Accuracy of a local model (Local), a decentralized FL over a static graph (FL-SG), and our approach over a dynamic graph (FL-DG). We show also the fraction of links between communities over time for FL-DG.

mates at agent a are not influenced by the removal of some incorrect edges as long as its d-hop neighborhood $\mathcal{N}_a^d$ remains unchanged. For instance, a given node $a' \notin C_a$ 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 consensus coefficients along the path. As paths are progressively removed, the negative impact of a' on agent a's estimate is gradually reduced. However, once all incorrect edges are removed, B-CoLME benefits from its estimates being computed solely on agents belonging to the same class, while in C-CoLME, some time is required for the effect of past estimates to fade away.

We also compare the proposed algorithms with CoLME and a simplified version (s-CoLME) in which the optimistic distance $d_\gamma^t(a, a')$ is recomputed only for the r agents that are queried at time t, achieving an $O(r)$ per-agent computational cost (note that the memory cost remains $O(|\mathcal{A}|)$). 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 CC_a^d in the case of B-CoLME, and CC_a^d in the case of C-CoLME. CoLME pays for this asymptotic improvement with a $O(|\mathcal{A}|)$ space-time complexity per agent, which makes the method impractical for large-scale systems. We observe that s-CoLME improves CoLME's complexity at the cost of a much slower discovery of same-class neighbors, which significantly affects the estimation process.

While we focused on the fundamental problem of online mean estimation, our approach could be adapted to decentralized federated learning, for example by replacing the optimistic distance $d_\gamma^t(a, a')$ with some estimator of the distance of the two local distributions. To illustrate this possibility, we consider a consensus-based decentralized federated learning algorithm Koloskova et al. [2020], where pairwise distances are computed on the basis of the cosine similarity of agents' updates, similarly to what done in CLUSTEREDFL Chen et al.; Ghosh et al. [2020] (details in Appendix H).

Figure 3 shows the performance of this algorithm considering $|\mathcal{A}| = 100$ agents organized over a complete graph, and whose data is drawn from one of two different distributions. In our decentralized FL over a dynamic graph (FL-DG), agents exclude neighbors over time that they deem to belong to a different class, and the agent's model is built only over similar neighbors, achieving higher accuracy.

7 Conclusions

In this paper, we introduced B-CoLME and C-CoLME, two scalable and fully distributed algorithms designed for collaborative estimation of the local mean. We have comprehensively evaluated the performance of these algorithms through both the-

oretical analysis and experimental validation. Our algorithms can be adapted to cooperatively learning personalized models within the context of federated learning.

References

E. Adi, A. Anwar, Z. Baig, and S. Zeadally. Machine learning and data analytics for the iot. *Neural Computing and Applications*, 32(20):16205–16233, May 2020. ISSN 1433-3058. doi: 10.1007/s00521-020-04874-y.

H. Amini. Bootstrap percolation and diffusion in random graphs with given vertex degrees. *The Electronic Journal of Combinatorics*, 17(1), Feb. 2010. doi: 10.37236/297.

M. Asadi, A. Bellet, O.-A. Maillard, and M. Tommasi. Collaborative Algorithms for Online Personalized Mean Estimation. 2022. ISSN 2835-8856.

M. Beaussart, F. Grimberg, M.-A. Hartley, and M. Jaggi. Waffle: Weighted averaging for personalized federated learning, 2021.

E. M. Chayti, S. P. Karimireddy, S. U. Stich, N. Flammarion, and M. Jaggi. Linear speedup in personalized collaborative learning, 2022.

M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui. A Joint Learning and Communications Framework for Federated Learning Over Wireless Networks. 20(1):269–283. ISSN 1558-2248. doi: 10.1109/TWC.2020.3024629.

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

Digital Library of Mathematical Functions. Asymptotic expansions: Exponential and logarithmic integral, 2023. [Online; accessed 23-January-2024].

S. Ding and W. Wang. Collaborative Learning by Detecting Collaboration Partners. 35:15629–15641, 2022.

K. Donahue and J. Kleinberg. Model-sharing Games: Analyzing Federated Learning Under Voluntary Participation. volume 35, pages 5303–5311, 2021.

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

M. Even, L. Massoulié, and K. Scaman. On Sample Optimality in Personalized Collaborative and Federated Learning. volume 35, pages 212–225, 2022.

A. Fallah, A. Mokhtari, and A. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3557–3568. Curran Associates, Inc., 2020.

M. Franceschelli and A. Gasparri. Multi-stage discrete time and randomized dynamic average consensus. *Automatica*, 99:69–81, Jan. 2019. ISSN 0005-1098. doi: 10.1016/j.automatica.2018.10.009.

A. Ghosh, J. Chung, D. Yin, and K. Ramchandran. An Efficient Framework for Clustered Federated Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 19586–19597. Curran Associates, Inc., 2020.

F. Grimberg, M.-A. Hartley, S. P. Karimireddy, and M. Jaggi. Optimal model averaging: Towards personalized collaborative learning, 2021.

S. Janson. The probability that a random multigraph is simple. *Combinatorics, Probability and Computing*, 18(1-2):205–225, 2009. doi: 10.1017/S0963548308009644.

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. Nitin Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and Open Problems in Federated Learning. 14(1–2):1–210, 2021. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000083.

A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 5381–5393. JMLR.org, 2020.

T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated Learning: Challenges, Methods, and Future Directions, 2020. ISSN 1053-5888, 1558-0792.

T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6357–6368. PMLR, 18–24 Jul 2021.

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

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

O. Marfoq, G. Neglia, A. Bellet, L. Kameni, and R. Vidal. Federated multi-task learning under a mixture of distributions. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

B. D. McKay and N. C. Wormald. Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, 11 (1):52–67, 1990. ISSN 0196-6774. doi: https://doi.org/10.1016/0196-6774(90)90029-E.

C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

E. Montijano, J. I. Montijano, C. Sagüés, and S. Martínez. Robust discrete time dynamic average consensus. *Automatica*, 50 (12):3131–3138, Dec. 2014. ISSN 0005-1098. doi: 10.1016/j.automatica.2014.10.005.

W. S. Rossi, G. Como, and F. Fagnani. Threshold models of cascades in large-scale networks. *IEEE Transactions on Network Science and Engineering*, 6(2):158–172, 2019. doi: 10.1109/TNSE.2017.2777941.

F. Sattler, K.-R. Müller, and W. Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3710–3722, 2021. doi: 10.1109/TNNLS.2020.3015958.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

A. Z. Tan, H. Yu, L. Cui, and Q. Yang. Towards Personalized Federated Learning. 34(12):9587–9603, 2023. ISSN 2162-237X, 2162-2388. doi: 10.1109/TNNLS.2022.3160699.

N. Tsoy, A. Mihalkova, T. N. Todorova, and N. Konstantinov. Provable mutual benefits from federated learning in privacy-sensitive domains. In *International Conference on Artificial Intelligence and Statistics*, pages 4798–4806. PMLR, 2024.

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

L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004. ISSN 0167-6911. doi: https://doi.org/10.1016/j.sysconle.2004.02.022.

A Appendix - Proof of Theorem 3

In this Appendix we provide the proof of Theorem 3 and, as a side result, we derive Equation (7). For the sake of clarity, we repeat the statement of the theorems. The proof of the results in Section 3.1 can be found in Asadi et al. [2022].

The first theoretical result provides a bound on the probability that the nodes in the connected component CC_a of a certain node $a \in \mathcal{A}$ misidentify their *true* neighbors $C_{a'} \cap \mathcal{N}_{a'}$, $\forall a' \in CC_a$. 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 γ.

Theorem 9 (Incorrect neighborhood estimation)**.** *For any $\delta \in (0,1)$, employing B-CoLME or C-CoLME we have:*

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

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

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. (3) and (7)) and considering the sample mean $\bar{x}$ computed over n samples, the probability that the *true* mean μ, falls within the confidence interval $\left[\bar{x} - \beta_\gamma(n), \bar{x} + \beta_\gamma(n)\right]$ 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 2γ (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 $a' \in CC_a$ 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 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 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 σ) 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:*

$$\mathbb{P}\left(\exists n \in \mathbb{N}, |\bar{x} - \mu| \geq \beta_\gamma(n)\right) \leqslant 2\gamma \tag{14}$$

Proof. **(σ-sub-Gaussian x_t)** We start from the theoretical guarantees on the parametrization (Eq. 3) 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 μ_t and is σ_t-sub-Gaussian. Then, for all $\gamma \in (0,1)$ it holds:

$$\begin{aligned}
&\mathbb{P}\left(\exists n \in \mathbb{N}, \sum_{t=1}^n (x_t - \mu_t) \geq \sqrt{2\sum_{t=1}^n \sigma_t^2\left(1+\frac{1}{n}\right)\ln\frac{\sqrt{n+1}}{\gamma}}\right) \leqslant \gamma \\
&\mathbb{P}\left(\exists n \in \mathbb{N}, \sum_{t=1}^n (\mu_t - x_t) \geq \sqrt{2\sum_{t=1}^n \sigma_t^2\left(1+\frac{1}{n}\right)\ln\frac{\sqrt{n+1}}{\gamma}}\right) \leqslant \gamma
\end{aligned} \tag{15}$$

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 μ_a and σ-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 Section 3 (which we restate here for completeness):

$$\beta_\gamma(n) := \sigma\sqrt{\frac{2}{n}\left(1+\frac{1}{n}\right)\ln\frac{\sqrt{n+1}}{\gamma}} \tag{16}$$

By a simple substitution, we get:

$$\mathbb{P}\left(\exists n \in \mathbb{N},\ \bar{x} - \mu \geq \beta_\gamma(n)\right) \leqslant \gamma \tag{17}$$

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

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

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. 14):

$$\mathbb{P}\left(\forall n \in \mathbb{N}, |\bar{x} - \mu| < \beta_\gamma(n)\right) \geq 1 - 2\gamma \tag{18}$$

Note also that the probabilistic confidence level γ can be considered as a function of δ. By choosing appropriately the function $\gamma = f(\delta)$ it is possible to provide the desired level δ 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:*

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

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

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

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:

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

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:

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

where $\zeta(z) := \sum_n \frac{1}{n^z}$, with $z \in \mathbb{C}$, denotes the ζ-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 2γ, 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 β_γ) 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 β_γ 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 $E := \left\{\forall a' \in CC_a, \forall t \in \mathbb{N}, \forall a'' \in \mathcal{N}_{a'}, \left|\bar{x}^t_{a',a''} - \mu_{a''}\right| < \beta_\gamma(n)\right\}$, which is equivalent to say that, for every node $a' \in CC_a$ 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 γ:

Lemma 12 (Confidence of β_γ interval)**.** *Considering the interval parametrization $\beta_\gamma(n)$ (Eq. (3) or (7)), setting $\gamma(\delta) = \frac{\delta}{4r|CC_a|}$, it holds:*

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

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

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

Applying a union bound with respect to the nodes $a' \in CC_a$ 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:

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

Now, we set $\gamma = \frac{\delta}{4r|CC_a|}$ and thus we immediately obtain $\mathbb{P}(E) \geq 1 - \frac{\delta}{2}$. This explains the value of the constant γ 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 C_a.

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

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

Proof. Defined the *optimistic* distance[7] 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:

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

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:

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

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

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

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

(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., $a' \notin C_a \cup \mathcal{N}_a \implies d^t_\gamma(a,a') > 0$). 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 ($|a-b| > ||a| - |b|| \implies |a| - |b| \leqslant |a-b|$):

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

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

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

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:

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

Moreover, consider that by definition[8] 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 ($a' \in C_a \cup \mathcal{N}_a$) 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., $a' \notin C_a \cup \mathcal{N}_a$), 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 $a' \notin C_a \cup \mathcal{N}_a \implies d^t_\gamma(a,a') > 0$. 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 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 $\gamma = \frac{\delta}{4r|CC_a|}$, 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 $\Delta_a = \min_{a' \in \mathcal{A}\setminus C_a} \Delta_{a,a'}$.

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 $t - 1 \leq n^\star_\gamma(a)$ this holds for all the pairs in the connected component CC_a, as $\Delta_a \geq \min_{a' \in CC_a \setminus C_a} \Delta_{a,a'}$.

□

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

B Appendix - Schematic Representation for B-CoLME

We provide a sketch for the functioning of the B-CoLME which, together with Algorithm 1 (in the main text) provides a detailed explanation of the proposed algorithm.

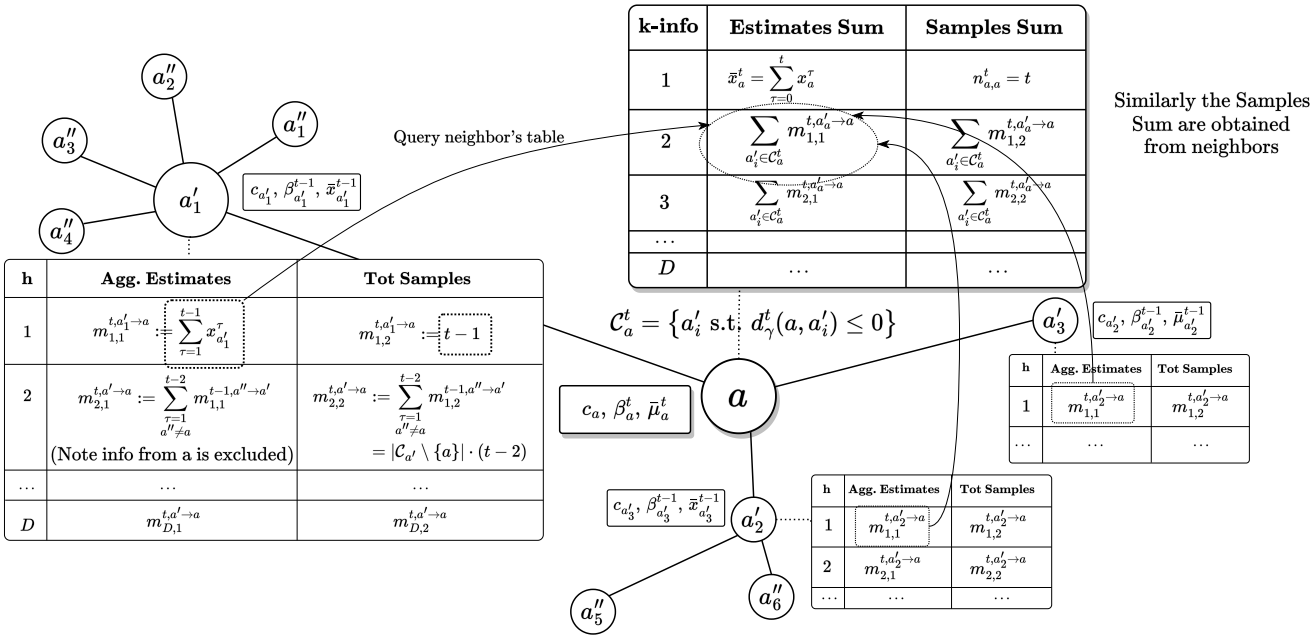


Figure 4: Simplified illustration of the functioning of the B-CoLME algorithm from the point of view of node a (all the quantities are already aggregated in the messages $m_{h,i}^{t,a' \to a}$, so that to exclude the "self" info sent by a).

C Appendix - Proof for B-CoLME - Theorem 4

First, we recall the notion of (ε, δ)-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 1 (PAC-convergence). *An estimation procedure for agent a is called (ε, δ)-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 D.

Theorem 14. *Provided that CC_a^d has a tree structure, for any $\delta \in (0,1)$, employing B-CoLME(d), we have:*

$$\mathbb{P}\left(\forall t > \tau_a', |\hat{\mu}_a^t - \mu_a| < \varepsilon\right) \geq 1 - \delta$$
$$\text{with} \quad \tau_a' = \max\left[\zeta_D + d, \frac{\widetilde{n}_{\frac{\delta}{2}}}{|CC_a^d|} + d\right]$$

where

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

with

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

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 σ, 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 $C_a^t = C_a \cap \mathcal{N}_a$ 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 CC_a^d. Such samples are i.i.d. and follows D_a; moreover its number easily lower bounded by $|CC_a^d|(t_0 - d)$. Therefore whenever $|CC_a^d|(t_0 - d) \geq \widetilde{n}_{\frac{\delta}{2}}$ by the definition of $\widetilde{n}_{\frac{\delta}{2}}$ we have

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

The claim descends from the definition of ζ_D in Theorem 3.

Now consider the more general case in which the distribution of samples at nodes are Gaussian with possibly different standard deviations, uniformly bounded by σ, 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 μ_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 σ, 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 σ having the same average μ_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 μ_a have a fourth central moment uniformly bounded bu μ_4, and a variance uniformly bounded by σ^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 μ_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:

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

□

D Appendix - Proofs for C-CoLME - Theorem 6 and Theorem 7

It is convenient to consider an auxiliary system for which the consensus matrix can be arbitrary until time ζ_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 $a' \in C_a \cap \mathcal{N}_a$.

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.

D.1 Preliminaries

After time ζ_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

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

D.2 Technical Results

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

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

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 $\tau_2 \geqslant \tau_1 \geqslant t'$, $\frac{\beta^{-\tau_1}}{\tau_1} \leqslant \frac{\beta^{-\tau_2}}{\tau_2}$.

$$\sum_{\tau=t_0+1}^{t+1} \frac{\beta^{-\tau}}{\tau} = \underbrace{\sum_{\tau=t_0}^{t'-1} \frac{\beta^{-\tau}}{\tau}}_{C} + \sum_{\tau=t'}^{t} \frac{\beta^{-\tau}}{\tau} + \frac{\beta^{-t-1}}{t+1} \tag{31}$$

$$\leqslant C + \frac{\beta^{-t-1}}{t+1} + \int_{t'}^{t+1} \frac{\beta^{-\tau}}{\tau}\,\mathrm{d}\tau \tag{32}$$

$$= C + \frac{\beta^{-t-1}}{t+1} + \int_{t'}^{t+1} \frac{e^{\ln\frac{1}{\beta}\tau}}{\tau}\,\mathrm{d}\tau \tag{33}$$

$$= C + \frac{\beta^{-t-1}}{t+1} + \int_{t'\ln\frac{1}{\beta}}^{(t+1)\ln\frac{1}{\beta}} \frac{e^x}{x}\,\mathrm{d}x \tag{34}$$

$$\leqslant C + \frac{\beta^{-t-1}}{t+1} + \mathrm{Ei}\left((t+1)\ln\frac{1}{\beta}\right) - \mathrm{Ei}\left(t'\ln\frac{1}{\beta}\right) \tag{35}$$

$$\leqslant C + \frac{\beta^{-t-1}}{t+1} + \mathrm{Ei}\left((t+1)\ln\frac{1}{\beta}\right) \tag{36}$$

$$\leqslant C + \frac{\beta^{-t-1}}{t+1} + \frac{e^{(t+1)\ln\frac{1}{\beta}}}{(t+1)\ln\frac{1}{\beta}}\left(1+\frac{3}{(t+1)\ln\frac{1}{\beta}}\right) \tag{37}$$

$$= C + \left(1+\frac{1}{\ln\frac{1}{\beta}} + \frac{3}{(t+1)\ln\frac{1}{\beta}}\right)\frac{\beta^{-t-1}}{t+1}, \tag{38}$$

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 $1 + \sqrt{\pi}\frac{\Gamma(n/2+3/2)}{\Gamma(n/2+1)}$ Digital Library of Mathematical Functions [2023]. This factor is smaller than 3 for $n = 0$.

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

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

□

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 $\mathbf{z}_\tau$.

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

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

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

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

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

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

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

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:*

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

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

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

$$\leqslant 2\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\left(\sum_{\tau_1,\tau_2,\tau_3,\tau_4=1}^{t}\frac{(\beta\lambda_2(B))^{4t-\tau_1-\tau_2-\tau_3-\tau_4}}{\tau_2\tau_3}+\sum_{\tau_1,\tau_2,\tau_3,\tau_4=1}^{t}\frac{(\beta\lambda_2(B))^{4t-\tau_1-\tau_2-\tau_3-\tau_4}}{\tau_2\tau_4}\right) \tag{51}$$

$$= 4\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\sum_{\tau_1,\tau_2,\tau_3,\tau_4=1}^{t}\frac{(\beta\lambda_2(B))^{4t-\tau_1-\tau_2-\tau_3-\tau_4}}{\tau_2\tau_3} \tag{52}$$

$$= 4\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\left(\sum_{\tau_1=1}^{t}(\beta\lambda_2(B))^{t-\tau_1}\right)^2\left(\sum_{\tau_2=1}^{t}\frac{(\beta\lambda_2(B))^{t-\tau_2}}{\tau_2}\right)^2 \tag{53}$$

$$\leq 4\,\mathbb{E}\left[\|\mathbf{z}\|^4\right]\frac{1}{(1-\beta\lambda_2(B))^2}\left(\sum_{\tau_2=1}^{t}\frac{(\beta\lambda_2(B))^{t-\tau_2}}{\tau_2}\right)^2 \tag{54}$$

$$\in O\left(\mathbb{E}\left[\|\mathbf{z}\|^4\right]\frac{1}{(1-\beta\lambda_2(B))^2}\left(1+\frac{1}{\ln\frac{1}{\beta\lambda_2(B)}}\right)^2\frac{1}{(t+1)^2}\right), \tag{55}$$

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 [2001][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 [2001][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 $a\geq 1$.

$$x\geqslant a\ln^2 a \implies \sqrt{x}\geqslant\sqrt{a}|\ln a|\underset{a\geq 1}{=}\sqrt{a}\ln a \tag{56}$$

$$\implies \sqrt{x}\geqslant\sqrt{a}\ln\sqrt{x}=\frac{\sqrt{a}}{2}\ln x \tag{57}$$

$$\implies x\geqslant\frac{a}{4}\ln^2 x, \tag{58}$$

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. □

D.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:*

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{t+1}-\boldsymbol{\mu}\|^4\right]\in O\left(\sup_{W_1,W_2,\cdots,W_{\zeta_D}}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]\alpha^{4t}\right)$$

$$+O\left(\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}\|^4\right]\frac{(1-\alpha)^4}{(1-\alpha\lambda_2(W))^2}\left(1+\frac{1}{\ln\frac{1}{\alpha\lambda_2(W)}}\right)^2\frac{1}{(t+1)^2}\right)$$

$$+O\left(\mathbb{E}\left[\|P\mathbf{x}-\boldsymbol{\mu}\|^4\right](1-\alpha)^2\left(1+\frac{1}{\ln\frac{1}{\alpha}}\right)^2\frac{1}{(t+1)^2}\right), \qquad \text{if } \alpha_t=\alpha, \tag{59}$$

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

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):

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

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:

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{t+1}-\boldsymbol{\mu}\|^4\right]\leq 27\cdot\alpha_{\zeta_D,t}^4\cdot\mathbb{E}\left[\|W\|^{4(t+1-\zeta_D)}\cdot\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]$$
$$+27\,\mathbb{E}\left[\left\|\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(W-P)^{t-\tau}(\bar{\mathbf{x}}^{\tau+1}-P\bar{\mathbf{x}}^{\tau+1})\right\|^4\right]$$
$$+27\,\mathbb{E}\left[\left\|\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(P\bar{\mathbf{x}}^{\tau+1}-\boldsymbol{\mu})\right\|^4\right] \tag{62}$$
$$\leq 27\cdot\underbrace{\alpha_{\zeta_D,t}^4\cdot\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}^{\zeta_D}-\boldsymbol{\mu}\|^4\right]}_{C_1}$$
$$+27\,\underbrace{\mathbb{E}\left[\left\|\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(W-P)^{t-\tau}(\bar{\mathbf{x}}^{\tau+1}-P\bar{\mathbf{x}}^{\tau+1})\right\|^4\right]}_{C_2}$$
$$+27\,\underbrace{\mathbb{E}\left[\left\|\sum_{\tau=\zeta_D}^{t}(1-\alpha_\tau)\alpha_{\tau+1,t}(P\bar{\mathbf{x}}^{\tau+1}-\boldsymbol{\mu})\right\|^4\right]}_{C_3}, \tag{63}$$

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

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

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

$$
\begin{aligned}
C_2 = \mathbb{E}\Bigg[&\left(\sum_{\tau_1=\check{\zeta}_D}^{t}(1-\alpha_{\tau_1})\alpha_{\tau_1+1,t}(W-P)^{t-\tau_1}(\bar{\mathbf{x}}^{\tau_1+1}-P\bar{\mathbf{x}}^{\tau_1+1})\right)^\top\\
&\left(\sum_{\tau_2=\check{\zeta}_D}^{t}(1-\alpha_{\tau_2})\alpha_{\tau_2+1,t}(W-P)^{t-\tau_2}(\bar{\mathbf{x}}^{\tau_2+1}-P\bar{\mathbf{x}}^{\tau_2+1})\right)\\
&\left(\sum_{\tau_3=\check{\zeta}_D}^{t}(1-\alpha_{\tau_3})\alpha_{\tau_3+1,t}(W-P)^{t-\tau_3}(\bar{\mathbf{x}}^{\tau_3+1}-P\bar{\mathbf{x}}^{\tau_3+1})\right)^\top\\
&\left(\sum_{\tau_4=\check{\zeta}_D}^{t}(1-\alpha_{\tau_4})\alpha_{\tau_4+1,t}(W-P)^{t-\tau_4}(\bar{\mathbf{x}}^{\tau_4+1}-P\bar{\mathbf{x}}^{\tau_4+1})\right)\Bigg] \qquad (67)\\
= \mathbb{E}\Bigg[&\sum_{\tau_1,\tau_2,\tau_3,\tau_4=\check{\zeta}_D}^{t}(1-\alpha_{\tau_1})\alpha_{\tau_1+1,t}(1-\alpha_{\tau_2})\alpha_{\tau_2+1,t}\ (1-\alpha_{\tau_3})\alpha_{\tau_3+1,t}(1-\alpha_{\tau_4})\alpha_{\tau_4+1,t}\\
&(\bar{\mathbf{x}}^{\tau_1+1}-P\bar{\mathbf{x}}^{\tau_1+1})^\top(W-P)^{2t-\tau_1-\tau_2}(\bar{\mathbf{x}}^{\tau_1+1}-P\bar{\mathbf{x}}^{\tau_1+1})\\
&(\bar{\mathbf{x}}^{\tau_3+1}-P\bar{\mathbf{x}}^{\tau_3+1})^\top(W-P)^{2t-\tau_3-\tau_4}(\bar{\mathbf{x}}^{\tau_4+1}-P\bar{\mathbf{x}}^{\tau_4+1})\Bigg]\\
\in &\begin{cases}
O\left(\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}\|^4\right]\frac{(1-\alpha)^4}{(1-\alpha\lambda_2(W))^2}\left(1+\frac{1}{\ln\frac{1}{\alpha\lambda_2(W)}}\right)^2\frac{1}{(t+1)^2}\right), & \text{if } \alpha_t=\alpha,\\
O\left(\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}\|^4\right]\frac{1}{(1-\lambda_2(W))^2}\left(1+\frac{1}{\ln\frac{1}{\lambda_2(W)}}\right)^2\frac{1}{(t+1)^4}\right), & \text{if } \alpha_t=\frac{t}{t+1},
\end{cases} \qquad (68)
\end{aligned}
$$

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:

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

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*

$$\mathbb{E}\left[\left\|{}_c\mathbf{x}-{}_c\boldsymbol{\mu}\right\|^4\right] = n_c\kappa\sigma^4+n_c(n_c-1)\sigma^4, \qquad (70)$$

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

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

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

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

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

D.4 (ε, δ)-Bounds: Proof of Theorem 7

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

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

for $\alpha_t = \alpha$, and

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

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 ζ_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

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

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

$$\begin{aligned}
\sum_{t=t'}^{\infty}\left(\frac{\ln(1+t)}{t+1}\right)^2 &\leqslant \int_{t'}^{\infty}\left(\frac{\ln t}{t}\right)^2 \mathrm{d}t && (84)\\
&= \int_{\ln t'}^{\infty} x^2 e^{-x}\,\mathrm{d}x && (85)\\
&= \left(e^{-x}\left(x^2+2x+2\right)\right)\Big|_{\infty}^{\ln t'} && (86)\\
&= \frac{(\ln t')^2 + 2\ln t' + 2}{t'}, && (87)\\
&= \frac{(\ln t' + 1)^2 + 1}{t'}, && (88)\\
&\leqslant 2\frac{(\ln t' + 1)^2}{t'}, && (89)\\
&= 2\frac{\ln^2(et')}{t'} && (90)
\end{aligned}$$

from which we conclude:

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

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

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

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

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

Let then define

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

Finally, let us consider the system of interest. With probability $1-\delta/2$ the agents will have identified the correct links by time ζ_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 ${}_c\tau^C$ with probability $1-\delta/2$. It follows that the probability that the estimates in the stochastic have not the required precision after time ${}_c\tau^C$ is at most δ. □

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

E.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. [9] 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 dv 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 et al. [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. E.1

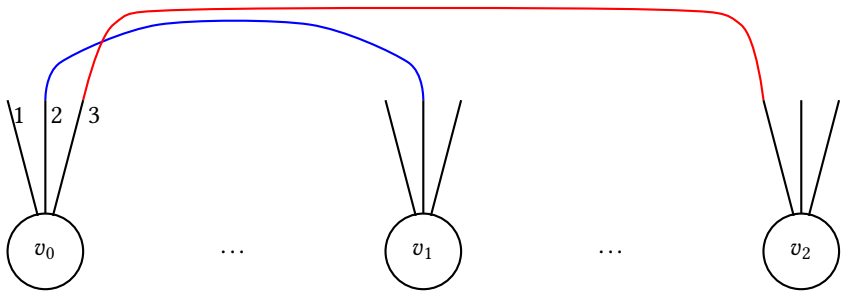


Figure 5: Random matching of stubs, $r = 3$

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

[9] 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])**.**

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

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].

E.2 The structure of CC_a^d and CC_a

Now we investigate on the structure of 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 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$\}$. 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 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. E.2. 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 $\{O_{a'}\}_{a' \in CC_a^d}$ are independent.

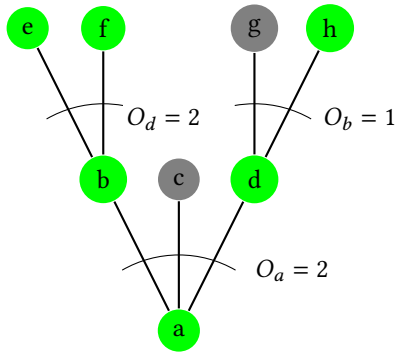


Figure 6: Structure $|CC_a^d|$: an example for $r = 3$, $d = 2$; green nodes belong to $|CC_a^d|$.

We can now recall and adapt a few standard results 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 8.

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*[10] *GW process $\{Z_i\}_i$ is bound to either extinction or explosion, i.e.,*

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

Previous result cam 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:

[10] 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 2: 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*

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

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:

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

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 8, we have $m_0 m^{d-1} = \widetilde{\Theta}(|\mathcal{A}|^{\frac{1}{2}(1+\log_{r-1} p_{k_a})})$, therefore, recalling that by construction $|CC_a^d| > Z_d$, by (96), we can always select $f(|\mathcal{A}|)$ that satisfies jointly $f(|\mathcal{A}|) = o(m_0 m^{d-1})$ and $f(|\mathcal{A}|) = \widetilde{\Theta}(|\mathcal{A}|^{\frac{1}{2}(1+\log_{r-1} p_{k_a})})$, such that:

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

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 CC_a, since by construction we have $CC_a^d \subseteq CC_a \; \forall d, a$, rather immediately we have:

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

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

F Appendix - Proof of the results in Table 1

The results in Table 1 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.

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

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

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

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

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

from which the asymptotics for ζ_a and τ_a can be derived opportunely replacing x and γ.

F.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

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

G Appendix - Additional Performance Evaluation of 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.

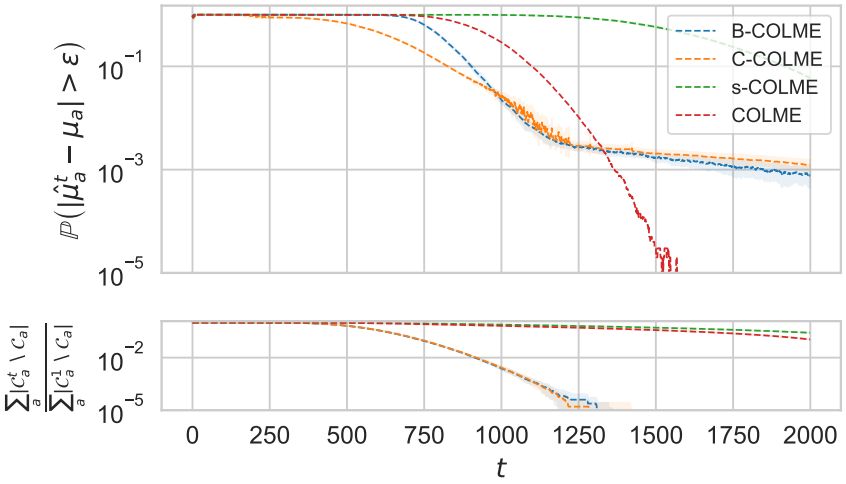


Figure 7: All approaches compared considering a smaller r compared to Figure 2 ($r = 5$), this shows that r needs to be chosen appropriately (large enough).

G.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.

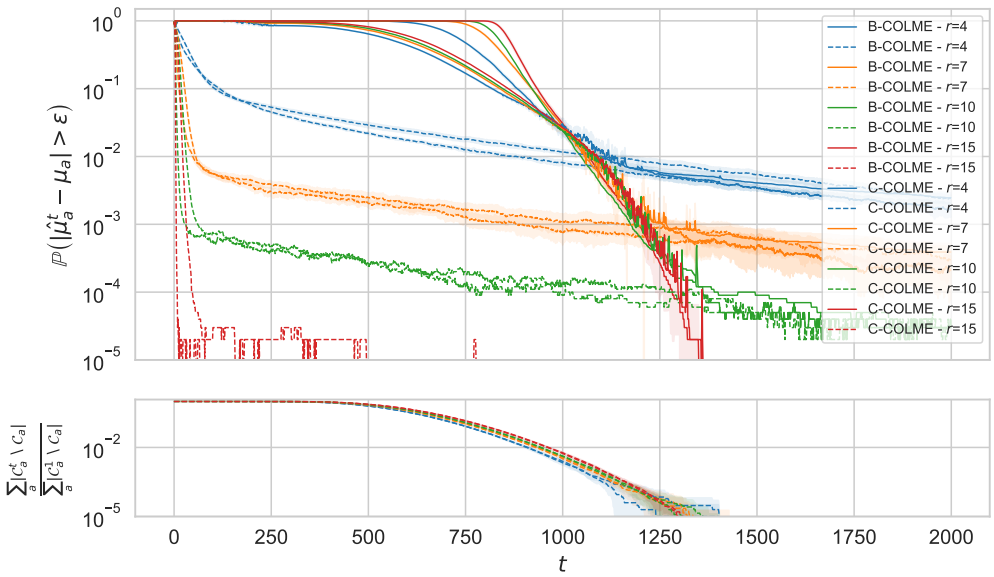


Figure 8: Performance comparison as a function of the parameter r of $G_0(N, r)$ for B-CoLME and C-CoLME. Dashed line is used for the *oracle* while solid line for the algorithm.

G.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.

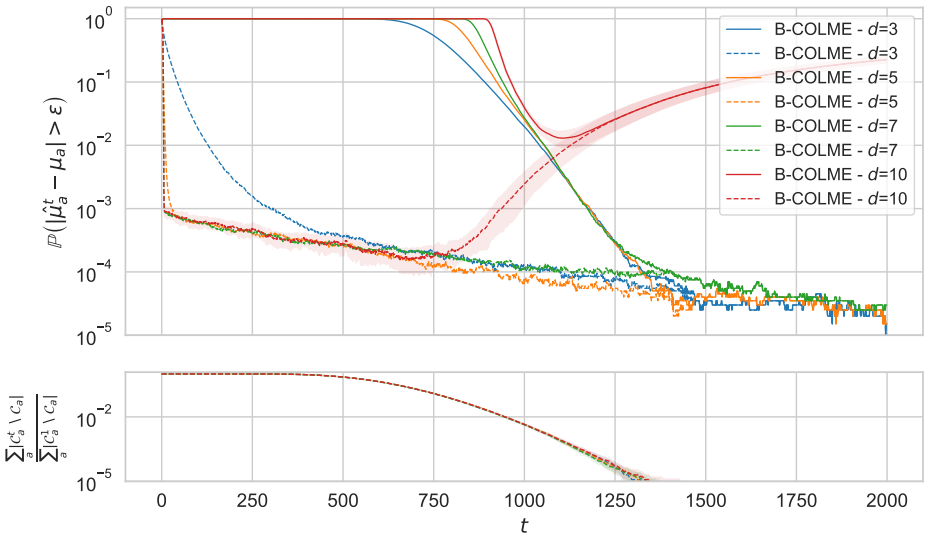


Figure 9: Performance comparison of B-CoLME as a function of the *depth* κ of the info kept. Dashed line is used for the *oracle* while solid line for the algorithm.

G.3 C-CoLME (Constant α) as a Function of the Weight α

We report some experiments considering α 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.

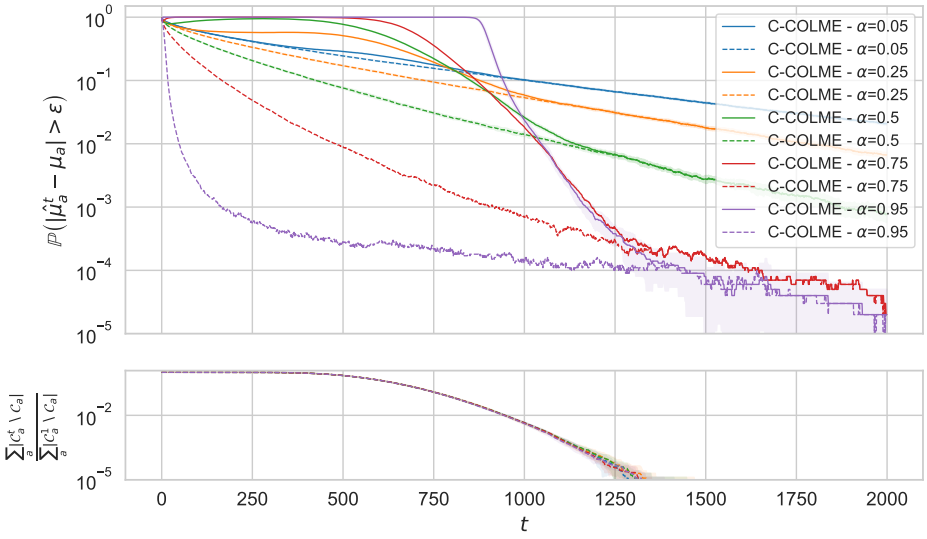


Figure 10: Performance comparison of C-CoLME as a function of the weight α when it is considered as constant. Dashed line is used for the *oracle* while solid line for the algorithm.

H Appendix - Additional Details on Decentralized Federated Learning (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: $\mathcal{G} = (\mathcal{A}, \mathcal{E})$, $D_a \in \{D_c\}_{c=1}^{C}\ \forall a \in \mathcal{A}$, θ_0
Output: collaborative FL-SG model θ_a, $\forall a \in \mathcal{A}$
$C_a^0 \leftarrow \mathcal{N}_a^t\ \forall a \in \mathcal{A}, \quad \theta_a^0 \leftarrow \theta_0, \quad \theta_l^0 \leftarrow \theta_0\ \forall l \in \mathcal{E}$
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 $\mathcal{N}_a \cap C_a^t$ **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**
      $C_a^{t+1} \leftarrow C_a^t \setminus a'$, and $C_{a'}^{t+1} \leftarrow C_{a'}^t \setminus a$
    end if
  end for
  // Model Updating Phase
  for node a in $\mathcal{A}$ **do**
    $\theta_a^{t+1} \leftarrow \frac{1}{|C_a^{t+1}|+1}\theta_a^{t+1}$
    for opt neighbor a' in C_a^{t+1} **do**
      $\theta_a^{t+1} \leftarrow \theta_a^{t+1} + \frac{1}{|C_a^{t+1}|+1}\theta_{a'}^{t+1}$
    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 θ_a^t, for agent a at time t.

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 $|M_{a,0}^0| = 30$),

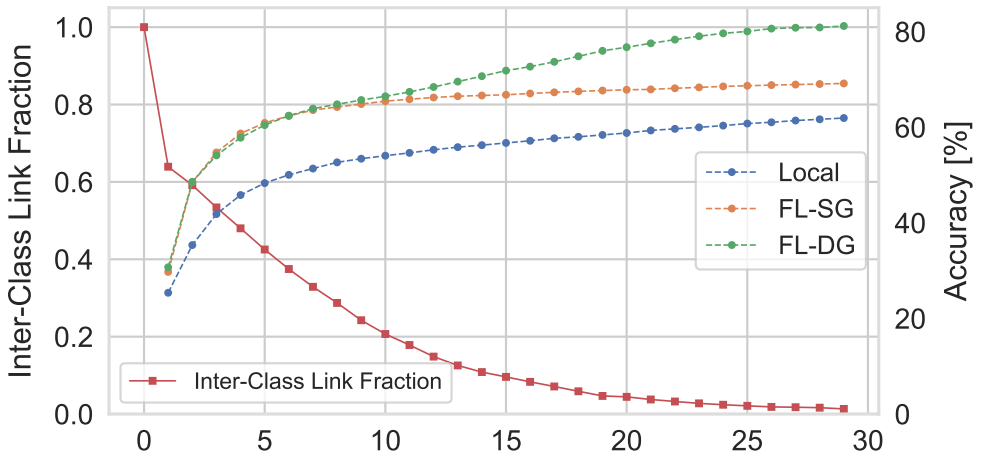


Figure 11: Accuracy of a local model (Local), a decentralized FL over a static graph (FL-SG), and our approach over a dynamic graph (FL-DG). We show also the fraction of links between communities over time for FL-DG.

and with the new samples they construct two overlapping minibatch $M^t_{a,0}$ and $M^t_{a,1}$. We consider a time horizon leading to two non-overlapping minibatch, i.e., $t = |M^0_{a,0}|$. 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 et al. [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 et al. [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 $\theta^t_{\{a,a'\}}$ (for each unordered pair (a, a'), "shared" between the nodes) and a node-link model $\theta^t_{a,a'}$ 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., $\theta^t_{a,a'}$. Then, at each training round, node a retrieves the shared model $\theta^t_{\{a,a'\}}$ and, starting from those parameters, trains the node-link model $\theta^t_{a,a'}$ 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 $\omega^t_{a,a'}$, which allows them to determine whether to collaborate with a neighbor or not. We can compute $\omega^t_{a,a'}$ as:

$$\omega^t_{\{a,a'\}} = \frac{\langle \Delta\theta^t_{a,a'}, \Delta\theta^t_{a',a} \rangle}{||\Delta\theta^t_{a,a'}|| \cdot ||\Delta\theta^t_{a',a}||} \tag{109}$$

This metric is updated at each iteration by making an average with the previous value (see Algorithm 3). Whenever $\omega^t_{a,a'}$ goes below a certain threshold ε_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 θ^t_a, averaging the parameters of the agents a' in their estimated similarity class C^t_a. Moreover, all the link models $\theta^t_{\{a,a'\}}$ are updated averaging the two node-link models of the nodes at the ends of the link, i.e, $\theta^{t+1}_{\{a,a'\}} = \theta^t_{\{a,a'\}} + \frac{\Delta\theta^t_{a,a'} + \Delta\theta^t_{a',a}}{2}$. 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.