[General response] We thank the reviewers for their useful comments. Before we go through our specific response to each reviewer, we would like to make two general comments that can be useful for all reviewers. **First**, The results from the paper by Woodworth and Srebro (NIPS 2016) [arxiv:1605.08003] imply that decentralized proximal algorithms cannot achieve linear convergence in the presence of more than one non-smooth term in the worst case. This conclusion illustrates the importance of the global common regularizer structure in our problem set-up. **Second**, We can *relax* the strong-convexity assumption and extend our analysis to cover the scenario in which the aggregate cost function $\bar{J}(w) = \frac{1}{K}\sum_{k=1}^{K} J_k(w) : \mathbb{R}^M \to \mathbb{R}$ is restricted strongly-convex (see [1] for details) while each individual cost $J_k(w)$ is not necessarily convex. This extension requires minor modifications and can be included in the revised manuscript. Let us briefly explain how to extend our analysis to this case. We know from the EXTRA paper [1] that the aggregate cost $\bar{J}(w)$ is $\bar{\nu}$-restricted-strongly-convex if, and only if, the penalized augmented cost $\mathcal{J}(\mathcal{w}) + \frac{\rho}{2}\|\mathcal{w}\|^2_{\mathcal{B}}$ ($\mathcal{J}(\mathcal{w}) = \frac{1}{K}\sum_{k=1}^{K} J_k(w_k) : \mathbb{R}^{MK} \to \mathbb{R}$) is $\bar{\nu}_\rho$-restricted-strongly-convex for some $\bar{\nu}_\rho > 0$ and any scalar $\rho > 0$. This means that for any $\mathcal{w}$, it holds [1]: $(\mathcal{w} - \mathcal{w}^\star)^{\mathsf{T}}\big(\nabla\mathcal{J}(\mathcal{w}) - \nabla\mathcal{J}(\mathcal{w}^\star)\big) + \rho\|\mathcal{w} - \mathcal{w}^\star\|^2_{\mathcal{B}} \geq \bar{\nu}_\rho\|\mathcal{w} - \mathcal{w}^\star\|^2$. Since $\mathcal{J}_\mu(\mathcal{w}) = \mathcal{J}(\mathcal{w}) + \frac{1}{2\mu}\|\mathcal{w}\|^2_{\mathcal{B}}$ has $\delta_\mu = (\delta + \frac{1}{\mu}\sigma_{\max})$-Lipschitz gradients it satisfies (see inequality (2.1.8) from [46]) $\|\nabla\mathcal{J}_\mu(\mathcal{w}) - \nabla\mathcal{J}_\mu(\mathcal{w}^\star)\|^2 \leq \delta_\mu\big(\nabla\mathcal{J}_\mu(\mathcal{w}) - \nabla\mathcal{J}_\mu(\mathcal{w}^\star)\big)^{\mathsf{T}}(\mathcal{w} - \mathcal{w}^\star) = \delta_\mu\big(\nabla\mathcal{J}(\mathcal{w}) - \nabla\mathcal{J}(\mathcal{w}^\star)\big)^{\mathsf{T}}\widetilde{\mathcal{w}} + \frac{1}{\mu}\delta_\mu\|\widetilde{\mathcal{w}}\|^2_{\mathcal{B}}$. Using the above two inequalities we reach $\|\widetilde{\mathcal{w}}_{i-1} - \mu\big(\nabla\mathcal{J}_\mu(\mathcal{w}_{i-1}) - \nabla\mathcal{J}_\mu(\mathcal{w}^\star)\big)\|^2 \leq (1 - \mu(2 - \mu\delta_\mu)\bar{\nu}_\rho)\|\widetilde{\mathcal{w}}_{i-1}\|^2 - (2 - \sigma_{\max} - \mu\delta - \mu\rho)\|\widetilde{\mathcal{w}}_{i-1}\|^2_{\mathcal{B}}$. Replacing the bound in Lemma 2 by this bound and following the same proof technique in the paper we can establish linear convergence under this more relaxed condition for some small enough constant $\rho > 0$ that depends on μ and $\sigma_{\max}$. For the smooth case, we can guarantee linear convergence for $\mu \leq O((1-\sigma_{\max})/\delta)$, which is still tighter than EXTRA $O(\bar{\nu}_\rho(1 - \sigma_{\max})/\delta^2)$ step-size. Note that the restricted strongly-convex condition is weaker than strong-convexity and it is met in sparse optimization settings and others – see Hui and Yin. [arXiv:1303.4645], Hui [arXiv:1511.01635] and references therein.

[Reviewer 1] Thank you for the positive comments on the paper. We agree it is better to move the simulation section to the supplementary material since the contributions are theoretical. In L142, p is an integer, which will be highlighted in the revision. In Eq. (7), we can cite [1] or [2]. The remarks in early section 2 are relevant to section 3 of the work of Loizou and Richtarik, which we were not aware of. We will include this work in the reference list.

[Reviewer 2] We thank the reviewer for his/her insightful and encouraging comments. We agree with all of the reviewers' comments and will make further clarifications in the revision: 1) DIGing can be fundamentally different from EXTRA as highlighted by the reviewer. That said, for a static and undirected network, both DIGing and EXTRA belong to a class of primal-descent dual-ascent methods applied to the augmented Lagrangian function. They only differ in choosing the weight matrix and augmented Lagrangian penalty matrix, which are highlighted in the supplementary material. To avoid misleading the readers, we can rephrase that remark by stating that our technique covers the class of augmented Lagrangian methods without mentioning DIGing. 2) As highlighted in the general response above, we can prove linear convergence under the same assumptions as EXTRA, and still provide tighter bounds. 3) The proposed algorithm in this work does not cover PG-EXTRA when $R(.) \neq 0$. The difference lies in the order by which the agents conduct their updates and the proximal mapping. We will clarify this point in the revision.

[Reviewer 4] We thank the reviewer for his/her comments and for verifying the correctness of the proof. Please see the second point in the general response above regarding relaxing our assumption.

[Reviewer 5] We thank the reviewer for his/her comments. We respectively disagree that our contribution is incremental. Our work closes the gap between the convergence rate of the centralized and decentralized methods for composite optimization problems. As observed by the reviewer, with a *slight* modification (yet novel) on the algorithm structure, we resolve a long-standing open question, which is a significant contribution in our opinion. In addition, our proof technique delicately handles the proximal mapping and, moreover, it is shorter and more constructive than existing proofs related to this work. For example, one can check that our proof is simpler to verify than those in [1], [2] and [26]. Third, while non-convex and asynchronous settings are of great practical value, they are beyond the scope of this paper. Note that the gap between the understanding of centralized and decentralized algorithms still exists even in the synchronous and convex scenario. Any theoretical breakthrough in the synchronous and convex optimization can be beneficial. For example, the novel work [29] closed the gap between decentralized and centralized optimal algorithms under similar conditions to this work and left the more practical conditions for future work – see [Section 4.3, 29]. Later, other works studied the more practical conditions – see [37] and [arXiv:1810.02660]. Another example is the work [20], which still has impact and is useful to current decentralized optimization works. Similarly, with the idea established in this work and [arXiv:1612.00150, arXiv:1810.02660], it is possible to establish the linear convergence for composite problems under the asynchronous setting. Finally, to partially resolve the reviewer's suggested improvements, we can relax the strong-convexity assumption so that only the aggregate cost, rather than each local $J_k(w)$, is required to be restricted-strongly-convex, see point two in the general response above.