We appreciate the valuable comments from the reviewers. We will revise them accordingly.

## Novelty of DAPG and Brief History of Decentralized Proximal Gradient Descent

The novelty of DAPG is the main concern of reviewers because DAPG is closely related to Mudag [25]. However, reviewers don't know and we don't emphasize that it is hard to extend decentralized gradient descent methods with gradient tracking to proximal counterpart with the same linear convergence rates. Please note that, researchers take *five years* to propose DPA [1], which is the first linearly convergent decentralized proximal gradient algorithm, after the publication of EXTRA [19], the first decentralized gradient descent method of linear convergence rate.

PG-EXTRA [17], the follow-up work of EXTRA, extends EXTRA to the composite setting which has extra non-smooth term, but only obtains the sub-linear convergence rate even for strongly convex $f_i(x)$. In the following years, different decentralized gradient descent algorithms with gradient tracking were proposed such as [13], [14]. However, no evidence shows that these algorithms can be extended to the composite setting but keeping linear convergence. Even the recent work NIDS [8] can only achieve a sub-linear convergence rate for the composite setting. In fact, until five years after EXTRA, DPA [1] is the first linearly convergent proximal gradient algorithm for decentralized optimization which is published in Neurips 2019. Just as the title of [13]-'Harnessing smoothness to accelerate distributed optimization' mentioned, gradient tracking tries to harness the smoothness of the function. However, due to non-smooth term, the convergence analysis of gradient tracking based methods become much harder.

From above history, we can observe that, in the decentralized setting, extending the results of decentralized gradient descent methods to their proximal counterparts is not an easy work. Thus, whether Mudag [25] can be extended to the decentralized proximal setting is not obvious. In fact, just as pointed by Reviewer 2, to deal with the non-smoothness, DAPG takes more consensus steps compared to Mudag. The proof of Lemma 2 is also totally different from the one of Lemma 9 of Mudag. Note that, the proof of DAPG does not rely on the technique of DPA, either.

Furthermore, the results obtained by DAPG are total new for decentralized proximal algorithms. To the best of our knowledge, DAPG is the first accelerated decentralized gradient descent which theoretically outperforms current decentralized proximal algorithms.

Thus, the novelty and contribution of DAPG are substantial.

**Reviewer_1** : Thank you for the comments on figures, we will revise accordingly.

**Reviewer_2**

Q1:Incremental contribution to Mudag

A1: Just as mentioned at the beginning, it is not easy to extend decentralized gradient descent with gradient tracking to the composite setting. Because of the non-smoothness of composite setting, DAPG has more consensus steps than Mudag and the proof of Lemma 2 is totally different from the one of Lemma 9 of Mudag.

**Reviewer_3**

Q1: Experiment and Relation to prior work and Reference to APM-C

A1: We should set $K$ according to the condition number of graph. We will give more experiments on graphs of large condition numbers in our revised paper. We will give more detailed comparison with Mudag and will add the reference of APM-C in our revised paper.

**Reviewer_4**

Q1: Comparison with Accelerated version of decentralized proximal gradient descent methods?

A1: As mentioned at the beginning, DPA [1] is the first decentralized proximal gradient descent method with linear convergence rate. NIDS is shown to achieve linear convergence rate in [24] which is the best convergence rate of decentralized proximal gradient methods can achieve before our work. Work [14] is a decentralized accelerated gradient descent method which can not deal with the non-smooth regularization term and no evidence shows that it can be directly extended to deal with the non-smooth regularization term. To the best of our knowledge, DAPG is the first decentralized accelerated proximal gradient descent method. Thus, we have compared DAPG with state-of-the-art algorithms theoretically and empirically.

Q2: What will happen if all the FastMix steps in Alg.1 are replaced by standard (K rounds of) communication? It is unclear what will happen $f_i(x)$ is $\nu$ strongly convex?

A2: By standard communication, the communication complexity of algorithm will depend on $1/(1-\lambda_2(W))$ instead of $1/\sqrt{1-\lambda_2(W)}$. DAPG does not require $f_i(x)$ to be $\nu$ strongly convex. But if $f_i(x)$ is $\nu$ strongly convex, the results of Theorem 1 of course still hold.