Response to the Reviewers

We thank the reviewers for their insightful comments on our work. We have made modifications to our manuscript according to Reviewer 1's comments.

Reviewer 1

Reviewer Point P 1.1 — In the revised manuscript, the authors have addressed all my concerns, especially adding more computation efficiency discussion. Current version looks very impressive. Just a small issue left. Some discussions about limitations or future efforts should be better added.

Reply: Thank you for the insightful suggesion. According to the suggestion, we have added a paragraph at the end of Section 7 "Conclusion and Future Work" to discuss the limitations of our work and point out potential future research directions. Specifically, in this work, we mainly analyze performance bottlenecks of GNN training/inference in the single-GPU environment on static graphs with the message-passing framework. In fact, performance bottlenecks of GNN training/inference over the multi-GPU or distributed environment, dynamic graphs and other GNN frameworks are also worth studying. In the future, we plan to explore the GNN training/inference performance analysis under the following scenarios:

1. *Multi-GPU or distributed GNN training/inference.* To handle large-scale graph datasets, training/inferring GNNs with the multi-GPU environment or the distributed environment is essential. Multi-GPU and distributed GNN training/inference will inevitably introduce overheads such as inter-GPU and inter-machine communication. How these overheads affect performance bottlenecks is worthy to study.

2. *Spatial-temporal graph datasets.* Spatial-temporal graphs usually have dynamic topology structures. They appear in a variety of applications like traffic speed forecasting [Li et al. (2018)] and human action recognition [Yan et al. (2018)]. Many new GNNs are proposed to handle this kind of dynamic graphs. The differences of performance issues between these GNNs and the classic GNNs are also worthy of in-depth investigation.

3. *Emerging GNN frameworks.* In this work, we analyzed the widely-used message-passing framework in GNN learning systems. However, some emerging GNN learning systems adopt different frameworks like SAGA framework [Ma et al. (2019)]. It is interesting to research whether different frameworks would lead to different performance bottlenecks.

References

Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the 6th International Conference on Learning Representations*.

Ma, L., Yang, Z., Miao, Y., Xue, J., Wu, M., Zhou, L., and Dai, Y. (2019). Neugraph: Parallel deep neural network computation on large graphs. In *Proceedings of the 2019 USENIX Annual Technical Conference*, pages 443–458. USENIX Association.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 7444–7452. AAAI.