Supplementary Materials
(Neighbor2Seq: Deep Learning on Massive Graphs by Transforming Neighbors to Sequences)

Meng Liu* Shuiwang Ji†

1 More Discussions

1.1 Information Loss in Neighbor2Seq. Our Neighbor2Seq obtains the sequence by integrating features of nodes in each layer of the neighborhood tree. This transformation may lose the cross-layer dependency information in the tree. Specifically, the Neighbor2Seq ignores the identities of nodes that each walk passes through and only considers what are the nodes in each layer of the neighborhood tree. Nevertheless, this information can neither be captured by message passing methods because the used aggregation is usually permutation-invariant. This implies that messages from different neighbors cannot be distinguished, as pointed out in [5]. According to our experimental results, our models can outperform existing message passing methods, such as GCN, under the situation that both cannot capturing such information. It is intriguing to have an in-depth exploration of whether such information is useful and how it can be captured effectively.

1.2 Relations with the Weisfeiler-Lehman Hierarchy. As shown in [9], most current GNNs are at most as powerful as the Weisfeiler-Lehman (WL) graph isomorphism test [8] in distinguishing graph structures. Our Neighbor2Seq is still under the WL hierarchy because the neighborhood tree used to obtain the sequence is indeed the one that the WL test uses to distinguish different graphs. We would be interested in exploring how Neighbor2Seq can be extended to go beyond the WL hierarchy as a future direction.

2 Dataset Descriptions

ogbn-papers100M [3] is the existing largest benckmark dataset for node classification. It is a directed citation graph of 111 million papers indexed by Microsoft Academic Graph (MAG) [7]. Each node is a paper and each directed edge indicates that one paper cites another one. Each node is associated with a 128-dimensional feature vector obtained by averaging the word2vec [4] embeddings of words in its title and abstract. Among the node set, approximately 1.5 millione of them are ARXIV papers, each of which has a label with one of ARXIV's subject areas. The rest nodes (*i.e.*, non-ARXIV papers) are not associated with label information. The task is to leverage the entire citation graph to infer the labels of the ARXIV papers. The time-based splitting is used as the splitting strategy. To be more specific, the training nodes are all ARXIV papers published until 2017, while the validation nodes are the ARXIV papers published in 2018, and the ARXIV papers published since 2019 are treated as test nodes.

ogbn-products [3] is an undirected Amazon product co-purchasing graph [1]. Nodes denote products and edges between two nodes indicate that the corresponding products are purchased together. Node features are derived by extracting bag-of-words representations from the product descriptions. Further, a Principal Component Analysis is applied to these features to reduce the dimension to 100. The task is to predict the categories of the products. A realistic splitting scheme is used in this data. Specifically, the products are firstly sorted according to their sales ranking, and then the top 10% products are used for training, next 2% for validation, and the rest for testing. This strategy matches the real-world situation where manual labeling is prioritized to important nodes and models are subsequently deployed to predict the less important ones.

Reddit [2], *Yelp* [10], and *Flickr* [10] are widely used datasets for inductive learning. During training, only the node features of training nodes and the edges between training nodes are visible. *Reddit* is a social network extracted from Reddit forum. Nodes represent posts and edges between two posts indicate the same user comments on both posts. Node features are fromed by GloVe CommonCrawl word vectors [6] of the posts. The task is to predict which community different posts belong to. The splitting is also time-based. *Yelp* is a social network constructed from Yelp website. Nodes are users and edges between two nodes indicate

*Department of Computer Science and Engineering, Texas A&M University. Email: mengliu@tamu.edu

†Department of Computer Science and Engineering, Texas A&M University. Email: sji@tamu.edu

they are friends. Node features of users are obtained by the word2vec embeddings of their corresponding reviews. The task is to predict the categories of businesses reviewed by different users, which is a multi-label classification task. *Flickr* is a social network based on Flickr, a photo sharing website. Nodes represent images and there is an edge between two nodes if two images share some common properties. The node features are fromed by the bag-of-words representations of the images. The task is to predict the category each image belongs to.

References

[1] K. Bhatia, K. Dahiya, H. Jain, Y. Prabhu, and M. Varma. The extreme classification repository: multi-label datasets & code. *URL http://manikvarma.org/downloads/XC/XMLRepository. html*, 2016.

[2] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

[3] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in Neural Information Processing Systems*, 2020.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[5] H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, and B. Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020.

[6] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[7] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, 2020.

[8] B. Weisfeiler and A. A. Lehman. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsia*, 2(9):12–16, 1968.

[9] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[10] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. Prasanna. Graphsaint: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020.