Non-local Exchange: Introduce Non-locality via Graph Re-wiring to Graph Neural Networks

Ziquan Wei
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
ziquanw@eamil.unc.edu

Guorong Wu
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
guorong_wu@med.unc.edu

Abstract

Graph is an effective data structure to characterize ubiquitous connections as well as evolving behaviors that emerge from the inter-wined system. Limited by the stereotype of node-to-node connections, learning node representations is often confined in a graph diffusion process where local information has been excessively aggregated as the random walk of graph neural networks (GNN) explores far-reaching neighborhoods layer-by-layer. In this regard, tremendous efforts have been made to alleviate feature over-smoothing issues such that current backbones can lend themselves in a deep network architecture. However, not as popular as designing a new GNN, less attention has been paid to underlying topology by graph re-wiring, which is not only mitigating flaws of the random walk but also the over-smoothing risk by reducing unnecessary diffusion in deep layers. Inspired by the human cognition of non-locality, we propose a non-local information exchange mechanism by establishing an express connection to the distant node, instead of propagating information along the (possibly very long) original pathway node-after-node. Since the seek of express connections throughout the graph could be computationally expensive in real-world applications, we propose a re-wiring framework (coined *express messenger* wrapper) to progressively incorporate express links in a non-local manner, which allows us to capture multi-scale feature without using a deep model, thus free of the over-smoothing challenge. We have integrated our *express messenger* wrapper with existing GNN backbones (either using graph convolution or tokenized transformer) and achieved a new record on the Roman-empire dataset along with SOTA performance on both homophilous and heterophilous datasets.

1 Introduction

Despite various deep models for graph learning, current GNNs are *de facto* closely bonded under the overarching umbrella of a topological message-passing paradigm [4, 6]. Given the observed graph feature representations, the driving factor of GNN is to learn intrinsic feature representations by alternatively (1) seeking an optimal feature subspace and (2) aggregating the information within a local neighborhood, where the learned feature representations are supposed to have a better alignment with the existing labels (supervised manner [18]) or exhibit a more structured behavior (unsupervised manner [12]). Since the node-to-node information exchange fundamentally underlines the graph topology, the wiring pattern presented in the graph becomes a pivotal factor steering the heterophily of graph leading to over-smoothing issues [20].

In the cliché of GNN, there is a converging consensus that the message-passing mechanism allows us to capture global graph feature representations, in a layer-by-layer fashion, by progressively aggregating the feature representations from the nearby nodes to the distant nodes. As demonstrated

NeurIPS 2024 Workshop on Behavioral Machine Learning.

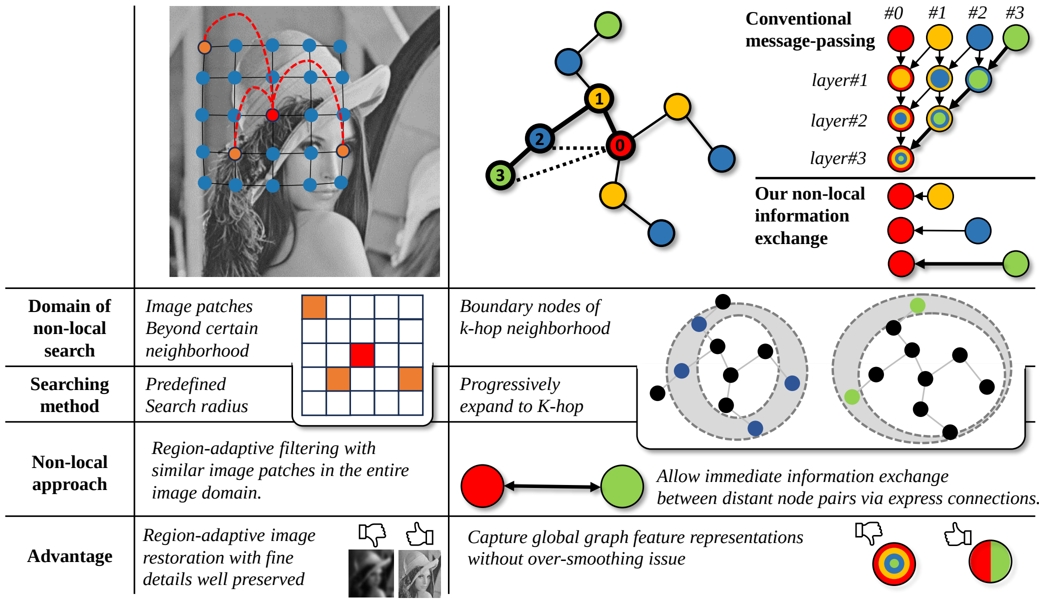

Figure 1: Non-local information exchange mechanism (right), a reminiscent of non-local mean technique for image processing (left), is able to capture global information by express connections (dashed links), which reduces the over-smoothing risk in GNN. Both ideas integrate information beyond either a spatial or topological neighbor, in order to preserve distinctive feature representations.

in the top-right panel of Figure 1, local features (centered at node $\#0$ and $\#1$) have to be aggregated multiple times until the message-passing process reaches node $\#3$. Limited by the pairwise links in the graph[1], however, the cost of such node-after-node message passing is over-smoothed feature representations due to an excessive number of local aggregations [1, 17]. In this context, the combination of local and global information by conventional node-after-node message-passing mechanism often yields over-smoothed feature representations which are mixed with all information spanning the entire graph pathway.

Inspired by the non-locality of human cognition [10], which is effective in computer vision filed by non-local means technique [3], we introduce a non-local exchanging (NLE) mechanism to the field of GNN by re-wiring the original graph with express connections for distant node pairs. In Figure 1 bottom-right, the selected express connection not only allows us to effectively capture global information but also offers a new window to maintain the distinctive power of the underlying feature representations. It is worth noting that our idea of express connection is simple yet effective. The conjecture is that the graph re-wiring step enhances the expressibility of topology, thus enabling GNN to learn informative global features prior to the incoming features undergoing significant smoothing by conventional message-passing routine. To that end, deep network architecture is no longer the only option to capture global feature representations, thus reducing the over-smoothing risk in GNN. In this context, we further devise a hierarchical re-wiring wrapper, called *express messenger* (*ExM*), that naturally fits the layer-by-layer network architecture of GNN models.

Our main contributions are summarized below:

- We propose a non-local information exchange mechanism to efficiently integrate feature representations from non-local neighborhoods by a collection of express connections.
- Current works focus on the optimization of feature representation with a deep GNN model by overcoming the over-smoothing issue. We address this challenge in a novel perspective of expressibility, i.e., the insight of our NLE mechanism is to directly combine local and global information through express links before the over-smoothed features undermine the discriminative power of feature representation.
- We devise our *ExM* wrapper as a pre-processing step, which generates new topologies in advance and facilitates various GNN models to retain SOTA performance on both homophilous and heterophilous graph data.

[1]High-order network models such as hyper-graph technique use hyperlink to model relationship among multiple nodes. However, most graph applications are conducted on 1-simplex graphs [2].

2 Methods

As explained in Figure 1, our core idea is to capitalize on a collection of express connections for capturing multi-scale graph feature representations, thus being free of over-smoothed features due to the conventional diffusion-based message-passing mechanism.

Supposing the non-local search is performed throughout the entire graph, the NLE mechanism yields a complementary graph topology $hop(\mathbf{A}, k)$ in addition to the original graph $\mathbf{A}$, where k is non-local searching radius. Since most intelligent systems process information in a hierarchical manner [8], we present a progressive *express messenger* (*ExM*) that gradually increases the hopping steps k and includes distant (but topologically connected) nodes v_j for the underlying node v_i ($i \neq j$) by a node selection function (1: selected for express connection; 0: otherwise):

$$hop(\mathbf{A}, k)_{ij} = \begin{cases} 1 & \text{for} \quad j \in \mathcal{N}_i^k - \mathcal{N}_i^{k-1} \\ 0 & \text{for} \quad j \notin \mathcal{N}_i^k - \mathcal{N}_i^{k-1} \end{cases} \tag{1}$$

where the hopping steps k acts as the the search radius from v_i to v_j on the graph, and $\mathcal{N}_i^k$ is the node set of k-hop neighborhoods around i-th node of graph. In other words, hop$(\mathbf{A}, k)_i$ rewires i-th node from its k-hop neighborhood to the boundary non-local nodes. Thus, we can obtain the new topology as this binary adjacency matrix.

ExM lets k in equation 1 progressively enlarge from K_{start} to K_{end} along with the model depth. Meanwhile, it replaces the original adjacency matrix $\mathbf{A}$ to $hop(\mathbf{A}, k)$ in k-th layer of GNN. Consequently, non-local information of the graph is exchanged with those nodes that are more distant when the model is deeper. That is based on the idea of conventional graph convolution the deeper layer of which gains more diffused features.

3 Experiments

We evaluate our proposed *ExM* wrapper through two types of experiments: (1) Benchmark graph rewiring performance by comparing to prior rewiring methods. (2) Benchmark graph representation learning performance on node classification using our *ExM* wrapper.

3.1 Experimental setting

Dataset. Experiments are carried out on a set of nine publicly accessible graph datasets, encompassing three homophilous graphs (Cora, PubMed, and Citeseer) and six heterophilous graphs (Texas, Wisconsin, Chameleon, Squirrel, Roman-empire (Roman.), and Amazon-ratings (Amazon.)). The initial seven graphs are widely recognized for evaluating graph representation learning techniques, while the last two were introduced by [14]. Our split settings are consistent with those outlined by [13, 14]. Further details can be found in Appendix.

Experiment setup. We conduct experiments using six baseline GNNs to evaluate our *ExM* framework: GAT [18], GT [16], SAGE [9], NAG [5], Jacobi [19], and FSGNN [11]. Among these, GAT, GT, and SAGE implementations are provided by [14], while the rest are implemented by their respective studies. We set the hyperparameters through the best model they provided. Graph re-wiring baselines are DropEdge [15], GDC [7], and SDRF [17]. Our *ExM* wrapper is employed without altering the model architecture to ensure a fair comparison. The performance is demonstrated using the optimal combination of our hyperparameters.

3.2 Benchmark graph re-wiring techniques

Since transformer models have started to prevail in the field of GNN, we select the latest graph transformer model NAG as the reference to benchmark the effect of various graph re-wiring techniques. Table 1 presents the results of graph rewiring performance across nine node classification tasks. Leveraging non-local information has led to improvements in node classification across various graph types. For instance, on homophilous graphs such as Cora (h=0.81) and Pubmed (h=0.80), as well as on heterophilous graphs like Roman-empire (h=0.05), the baseline performance has been improved. Among five heterophilous graphs ($h \leq 0.24$), four are improved significantly (denoted by '*'). With the exception of Squirrel and Citeseer, where our method achieved the second-best performance, it

Table 1: Benchmark results for graph rewiring performance. NAG is the baseline here. Red denotes the first rank, followed by blue (2^{nd}). '–' means the model has no result reported.

	Roman. h=0.05	Texas h=0.11	Wisconsin h=0.20	Squirrel h=0.22	Chameleon h=0.24	Amazon. h=0.38	Citeseer h=0.74	Pubmed h=0.80	Cora h=0.81
Baseline	73.57$_{\pm 1.30}$	70.54$_{\pm 3.07}$	67.45$_{\pm 1.80}$	34.85$_{\pm 0.85}$	46.05$_{\pm 1.10}$	47.08$_{\pm 0.60}$	71.32$_{\pm 0.65}$	87.65$_{\pm 0.24}$	86.08$_{\pm 0.69}$
DropEdge	75.97$_{\pm 0.83}$	66.48$_{\pm 3.24}$	61.18$_{\pm 2.12}$	34.49$_{\pm 1.13}$	47.23$_{\pm 1.09}$	45.34$_{\pm 0.50}$	72.40$_{\pm 0.46}$	87.33$_{\pm 0.46}$	69.08$_{\pm 0.67}$
GDC	75.62$_{\pm 0.26}$	72.43$_{\pm 1.62}$	70.19$_{\pm 3.37}$	33.84$_{\pm 0.93}$	45.48$_{\pm 1.09}$	46.20$_{\pm 0.55}$	71.76$_{\pm 0.58}$	87.80$_{\pm 0.28}$	85.67$_{\pm 0.75}$
SDRF	–	70.35$_{\pm 0.60}$	61.55$_{\pm 0.86}$	37.67$_{\pm 0.23}$	44.46$_{\pm 0.17}$	–	72.58$_{\pm 0.20}$	79.10$_{\pm 0.11}$	82.76$_{\pm 0.23}$
ExM	77.00$^{*}_{\pm 0.59}$	73.51$^{*}_{\pm 2.02}$	72.75$^{*}_{\pm 2.05}$	37.42$^{*}_{\pm 1.19}$	58.60$^{*}_{\pm 1.12}$	49.67$_{\pm 0.48}$	72.43$_{\pm 0.76}$	88.64$_{\pm 0.38}$	86.68$_{\pm 0.55}$

Table 2: Performance by using *ExM* to other SOTA methods. The average improvements by *ExM* across datasets and baselines are reported in the last column and row, respectively.

	Roman. h=0.05	Texas h=0.11	Wisconsin h=0.20	Squirrel h=0.22	Chameleon h=0.24	Citeseer h=0.74	Pubmed h=0.80	Avg. improve
GAT	80.92$_{\pm 0.68}$	58.92$_{\pm 5.81}$	60.39$_{\pm 3.67}$	62.00$_{\pm 1.29}$	69.85$_{\pm 1.72}$	74.03$_{\pm 1.23}$	87.86$_{\pm 0.42}$	7.80$_{\pm 6.82}$
w/ *ExM*	86.06$_{\pm 0.35}$	75.14$_{\pm 7.73}$	72.35$_{\pm 7.36}$	69.54$_{\pm 1.30}$	73.18$_{\pm 1.60}$	74.25$_{\pm 1.27}$	88.04$_{\pm 0.42}$	
GT	85.70$_{\pm 0.99}$	62.43$_{\pm 7.80}$	57.65$_{\pm 4.91}$	58.09$_{\pm 1.50}$	68.99$_{\pm 2.55}$	74.68$_{\pm 1.46}$	87.51$_{\pm 0.52}$	2.95$_{\pm 3.29}$
w/ *ExM*	89.20$_{\pm 0.63}$	68.11$_{\pm 10.1}$	67.06$_{\pm 8.71}$	58.09$_{\pm 1.50}$	70.92$_{\pm 1.44}$	74.83$_{\pm 1.22}$	87.51$_{\pm 0.52}$	
SAGE	86.96$_{\pm 0.56}$	80.27$_{\pm 5.70}$	81.37$_{\pm 5.49}$	44.53$_{\pm 1.08}$	62.92$_{\pm 1.70}$	75.19$_{\pm 1.51}$	88.59$_{\pm 0.38}$	1.55$_{\pm 1.42}$
w/ *ExM*	90.34$_{\pm 0.42}$	82.97$_{\pm 6.38}$	84.71$_{\pm 2.23}$	45.49$_{\pm 1.17}$	62.92$_{\pm 1.70}$	75.51$_{\pm 1.46}$	88.75$_{\pm 0.38}$	
NAG	73.57$_{\pm 1.30}$	70.54$_{\pm 3.07}$	67.45$_{\pm 1.80}$	34.85$_{\pm 0.85}$	46.05$_{\pm 1.10}$	71.32$_{\pm 0.65}$	87.65$_{\pm 0.24}$	4.13$_{\pm 3.69}$
w/ *ExM*	77.00$_{\pm 0.59}$	73.51$_{\pm 2.02}$	72.75$_{\pm 2.05}$	37.42$_{\pm 1.19}$	58.60$_{\pm 1.12}$	72.43$_{\pm 0.68}$	88.64$_{\pm 0.38}$	
Jacobi	71.25$_{\pm 0.45}$	76.49$_{\pm 6.74}$	76.67$_{\pm 5.00}$	50.21$_{\pm 2.39}$	67.94$_{\pm 1.13}$	76.00$_{\pm 1.44}$	88.95$_{\pm 0.46}$	1.73$_{\pm 1.11}$
w/ *ExM*	72.75$_{\pm 0.64}$	79.46$_{\pm 3.86}$	79.22$_{\pm 4.49}$	53.07$_{\pm 2.54}$	69.89$_{\pm 1.50}$	76.08$_{\pm 1.54}$	89.14$_{\pm 0.42}$	
FSGNN	67.93$_{\pm 0.53}$	85.95$_{\pm 5.77}$	85.88$_{\pm 4.94}$	73.55$_{\pm 2.16}$	78.16$_{\pm 1.11}$	76.52$_{\pm 1.76}$	89.63$_{\pm 0.40}$	1.04$_{\pm 1.53}$
w/ *ExM*	72.61$_{\pm 0.57}$	86.49$_{\pm 4.52}$	87.06$_{\pm 3.84}$	73.85$_{\pm 2.08}$	78.22$_{\pm 0.81}$	76.95$_{\pm 1.25}$	89.72$_{\pm 0.49}$	
Avg. improve	3.61$_{\pm 1.16}$	5.18$_{\pm 5.16}$	5.62$_{\pm 3.85}$	4.04$_{\pm 6.13}$	3.30$_{\pm 4.29}$	0.39$_{\pm 0.34}$	0.27$_{\pm 0.33}$	3.20$_{\pm 4.25}$

secured the top-1 position in all other cases. In contrast, previous graph re-wiring approaches faced challenges with homophilous graphs and exhibited lower accuracy on graphs with $h > 0.24$.

3.3 Evaluation on graph feature representation learning

Table 2 illustrates the contrast between the baseline GNN model and baseline+*ExM* wrapper, where our *ExM* wrapper facilitates most baseline methods retaining SOTA performance. It's worth noting that in the majority of cases, the *ExM* wrapper outperforms the baseline. Specifically, the *ExM* wrapper allows us to successfully secure a top-3 ranking across seven different graph datasets characterized by diverse homophily ratios h. The *ExM* wrapper is seamlessly integrated into various baselines, ranging from GAT to the recent FSGNN. Meanwhile, it's worth noting that SAGE+*ExM* achieves a new performance record on the Roman-empire dataset, as reported in the current leaderboard by [14].

4 Conclusion

In this work, we introduce the non-locality of human cognition to GNN through a novel perspective of topological re-wiring. We put the spotlight on the efficiency of capturing global information by a deep GNN model to the extent that can be effectively through a set of express connections between two distant but topologically connected nodes in the graph. By doing so, the new re-wired graph holds non-local connections and allows GNN to learn global features while reducing the chance of over-smoothing features in the conventional node-after-node graph diffusion process. Following this notion, we present an *express messenger* wrapper, serving as an agnostic pre-processing, to re-wire the graph topology via non-local information exchange, which is reminiscent of non-local mean technique prevailing in image processing area more than a decade ago. In practice, the new graph topology by our *ExM* can enhance the global feature of representation learning and instantly boost the node classification performance for current GNN models. Experiments show that our *ExM* wrapper can outperform previous graph re-wiring methods and help various GNN backbones to achieve SOTA performance on nine homophilous/heterophilous graph datasets, indicating the great potential in other graph learning applications such as brain connectomes and drug medicine data.

References

[1] Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205*, 2020.

[2] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1–92, 2020.

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005.

[4] Shaofei Cai, Liang Li, Jincan Deng, Beichen Zhang, Zheng-Jun Zha, Li Su, and Qingming Huang. Rethinking graph neural architecture search from message-passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6657–6666, June 2021.

[5] Jinsong Chen, Kaiyuan Gao, Gaichao Li, and Kun He. Nagphormer: A tokenized graph transformer for node classification in large graphs. In *The Eleventh International Conference on Learning Representations*, 2022.

[6] Jiarui Feng, Yixin Chen, Fuhai Li, Anindya Sarkar, and Muhan Zhang. How powerful are k-hop message passing graph neural networks. *Advances in Neural Information Processing Systems*, 35:4776–4790, 2022.

[7] Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. *Advances in neural information processing systems*, 32, 2019.

[8] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

[9] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[10] Stuart Kauffman and Sudip Patra. Human cognition surpasses the nonlocality tsirelson bound: Is mind outside of spacetime? *arXiv preprint arXiv:2301.12931*, 2022.

[11] Sunil Kumar Maurya, Xin Liu, and Tsuyoshi Murata. Simplifying approach to node classification in graph neural networks. *Journal of Computational Science*, 62:101695, 2022.

[12] Yujie Mo, Liang Peng, Jie Xu, Xiaoshuang Shi, and Xiaofeng Zhu. Simple unsupervised graph representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7797–7805, 2022.

[13] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*, 2020.

[14] Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: are we really making progress? *arXiv preprint arXiv:2302.11640*, 2023.

[15] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

[16] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.

[17] Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. *arXiv preprint arXiv:2111.14522*, 2021.

[18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[19] Xiyuan Wang and Muhan Zhang. How powerful are spectral graph neural networks. In *International Conference on Machine Learning*, pages 23341–23362. PMLR, 2022.

[20] Jiong Zhu, Ryan A Rossi, Anup Rao, Tung Mai, Nedim Lipka, Nesreen K Ahmed, and Danai Koutra. Graph neural networks with heterophily. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11168–11176, 2021.

A Appendix

A.1 Dataset details

In Table 3, we show the profile of all graph datasets used in this paper.

Table 3: Graph data profiles.

	h	Node number	Edge number	Class number
Cora	0.81	2,708	10,556	7
Pubmed	0.80	19,717	88,648	3
Citeseer	0.74	3,327	9,104	6
Amazon.	0.38	24,492	93,050	5
Chameleon	0.24	2,277	31,421	5
Squirrel	0.22	5,201	198,493	5
Wisconsin	0.20	251	515	5
Texas	0.11	183	325	5
Roman.	0.05	22,662	32,927	18