Overcoming Information Bottlenecks in Directed Graph Neural Networks through Rewiring

Yu He
Stanford University
heyu@stanford.edu

Ishani Karmarkar
Stanford University
ishanik@stanford.edu

Ellen Vitercik
Stanford University
vitercik@stanford.edu

Abstract

Graph Neural Networks (GNNs) have become an essential tool in relational learning. However, most research has focused on undirected graphs, overlooking the directional structure that shapes many real-world systems. Directionality encodes important asymmetries in domains such as traffic routing, social networks, and causal modeling, where ignoring edge orientation may lose semantic information and hinder predictive performance. At the same time, message passing in GNNs is limited by over-squashing and long-range dependency, challenges that can be alleviated by modifying the graph topology. While recent rewiring techniques based on spectral graph theory have shown promise for undirected graphs, they do not extend naturally to directed settings. In this work, we introduce a rewiring framework for directed graphs that improves information flow while preserving inherent asymmetries and reachability constraints. Our method directly tackles over-squashing and long-range dependency issues, advancing graph rewiring into the directed regime and broadening the use of GNNs in critical real-world tasks.

1 Introduction

Graph Neural Networks (GNNs) [1–3] have achieved remarkable success in domains such as recommendation systems [4], molecular property prediction [5], and material sciences [6]. Their power lies in the message-passing paradigm [7], where nodes iteratively update their representations by aggregating information from their neighbors. Despite this success, most GNN research and benchmarks focus on undirected graphs. Standard datasets [8, 9] are either inherently undirected or preprocessed by symmetrizing edges, thereby erasing the directional nature of information flow that is fundamental in many real-world systems.

In applications such as traffic routing, knowledge graphs, and online influence modeling, edge direction encodes critical asymmetries [10]. Ignoring this structure discards semantic richness and risks undermining predictive performance. Indeed, directionality is central not only in graph data but also in broader machine learning contexts: sequence and causal models operate over directed acyclic structures [11], from word order in language [12] to algorithm execution [13], and temporal graphs [14]. Recent theory further highlights the unique challenges of directed graphs [11, 12], including oversquashing in high in-degree regions and difficulties generalizing out-of-distribution. These observations underscore the need for GNNs explicitly designed for directed settings.

At the same time, message passing faces two long-standing limitations: oversquashing [15], where information is compressed through narrow structural bottlenecks, and poor long-range dependency modeling [9], since many rounds are required for signals to traverse distant nodes. A growing body of work addresses these issues through graph rewiring, i.e., modifying the topology to improve information flow. For undirected graphs, strategies based on spectral properties [16] such as curvature [17] and effective resistance [18, 19] have shown promise. Yet these tools do not naturally extend to directed graphs, which lack symmetry, straightforward spectral decompositions, and mutual reachability. As a result, principled rewiring techniques for directed graphs remain largely unexplored.

He et al., Overcoming Information Bottlenecks in Directed Graph Neural Networks through Rewiring (Extended Abstract). Presented at the Fourth Learning on Graphs Conference (LoG 2025), Hybrid Event, December 10–12, 2025.

In this work, we close this gap by introducing a rewiring framework tailored to directed graphs. Our approach quantifies information flow in directed structures and proposes rewiring strategies that alleviate oversquashing and extend long-range dependency modeling. By handling challenges unique to directed graphs, such as inherent asymmetries and reachability constraints, our method expands the scope of graph rewiring and enables GNNs to better capture the structure of real-world data.

2 Information Bottlenecks in Directed Message Passing

We represent a directed graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with nodes $\mathcal{V}$, directed edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and node features $\mathbf{X} \in \mathbb{R}^{n \times d}$. Each edge $(u, v) \in \mathcal{E}$ indicates a one-way connection from u to v, and $u \to v$ means that node v is reachable from node u.

Directed message passing. Extending message passing neural networks (MPNNs) to directed graphs requires explicitly modeling edge orientation. Approaches include magnetic Laplacian filters in MagNet [20], direction-aware attention in transformer architectures [21–23], and Dir-GNN [10], which separates in- and out-neighbors. For example, the Dir-GNN update at layer k is:

$$\mathbf{m}_{i,\text{in}}^{(k)} = \text{AGGREGATE}_{\text{in}}^{(k)}\left(\{f_{\text{in}}^{(k)}(\mathbf{x}_j^{(k-1)}, \mathbf{x}_i^{(k-1)}, \mathbf{e}_{ji}) : (j, i) \in \mathcal{E}\}\right), \tag{1}$$

$$\mathbf{m}_{i,\text{out}}^{(k)} = \text{AGGREGATE}_{\text{out}}^{(k)}\left(\{f_{\text{out}}^{(k)}(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}, \mathbf{e}_{ij}) : (i, j) \in \mathcal{E}\}\right), \tag{2}$$

$$\mathbf{x}_i^{(k)} = \text{UPDATE}^{(k)}\left(\mathbf{x}_i^{(k-1)}, \mathbf{m}_{i,\text{in}}^{(k)}, \mathbf{m}_{i,\text{out}}^{(k)}\right). \tag{3}$$

On the other hand, rewiring usually modifies the graph itself, offering a model-agnostic way to relieve bottlenecks. Most existing methods, however, are designed for undirected graphs, making it essential to study how such bottlenecks can be tackled in directed graphs.

Long-range dependency. Message passing is inherently local: after ℓ layers, a node only accesses its ℓ-hop receptive field $\mathcal{R}_v^{(\ell)}$. Capturing long-range dependencies therefore requires many layers, which increases cost and risks over-smoothing [24]. Solutions include skip connections, positional encodings, and rewiring strategies that shorten path lengths, such as diffusion-based edges [25, 26] or curvature-based modifications [17, 27]. Yet these rely on spectral tools suited to undirected graphs, making extensions to directed settings challenging due to asymmetry and reachability.

Oversquashing. Oversquashing arises when signals from many nodes are forced through a few intermediaries, compressing them into indistinguishable embeddings [15] which leads to information loss. While oversquashing is more likely to arise with long-range dependencies, their formal relationship remains an open research question [9]. Rewiring has proven effective in undirected graphs—via resistance [18], commute time [28], or curvature [17]—but these approaches again depend on spectral formalisms. In directed or DAG-like structures, oversquashing has been linked to high in-degree nodes and poor dispersion [29, 30]. This motivates our work: model-agnostic strategies tailored to directed graphs, designed to improve information flow while respecting asymmetry.

3 Methodology

We propose a rewiring framework for directed graphs that builds on shortcut sets and edge connectivity to address long-range dependency and oversquashing, while also handling the unique challenges of asymmetry and reachability and maintaining its scalability and efficiency.

3.1 Shortcut set rewiring

We address the *long-range dependency* challenge using the concept of a *shortcut set*. Let $G = (\mathcal{V}, \mathcal{E})$ be a directed graph, and let $\text{TC}(G) = (\mathcal{V}, \mathcal{E}^*)$ denote its transitive closure. A set of edges $S \subseteq \mathcal{E}^*$ is called a *d-shortcut set* if the augmented graph $H = (\mathcal{V}, \mathcal{E} \cup S)$ preserves the same reachability as G, and for every reachable pair $u, v \in \mathcal{V}$, there exists a path from u to v in H of length at most d. In essence, the shortcut set compresses long paths, reducing the graph's diameter while maintaining its asymmetric reachability structure.

To construct such a shortcut set efficiently, we adopt the classic $\sqrt{n}$*-sampling algorithm* [31], which guarantees $d = O(\sqrt{n})$ with only $O(n)$ additional edges. The algorithm proceeds as follows: each node is independently sampled with probability $1/\sqrt{n}$ to form a landmark set L. Then, for every pair $(a, b) \in L \times L$, if b is reachable from a in G, we add the edge (a, b) to the shortcut set S. The intuition is that long paths in G are likely to pass through landmarks near their endpoints, allowing a single shortcut to dramatically reduce path length. With high probability, this yields an augmented graph H with diameter $\tilde{O}(\sqrt{n})$, unchanged reachability, and a linear number of added edges.

In our framework, we treat the shortcut edges in S as rewired edges. This rewiring strategically reduces the diameter and enhances the model's ability to capture long-range dependencies, while preserving directed reachability constraints for modeling asymmetric systems. The linear edge budget further ensures that our approach remains efficient and scalable.

3.2 Greedy edge-connectivity rewiring

Another limitation of message passing is *oversquashing*, where information from many nodes is funneled through a few edges and compressed. Prior work has addressed this with influence-based metrics such as effective resistance [18], but these are spectral and not well-suited for directed graphs. We instead use *edge connectivity*, defined as the number of edge-disjoint $u \to v$ paths, which directly captures topological bottlenecks and naturally extends to asymmetric settings.

To keep the method scalable, we restrict computation to *ℓ-hop edge connectivity*, where ℓ matches the number of message-passing layers. Preserving this locality is also beneficial, as locality-aware rewiring has been shown to reduce the loss of structural information from the original graph [19]. For each source node u, we run a BFS to depth ℓ to build the layered subgraph

$$D_\ell(u) = \left(\bigcup_{t=0}^{\ell} V_t(u),\ \{(x, y) \in \mathcal{E} : x \in V_t(u),\ y \in V_{t+1}(u) \text{ for some } t < \ell\} \right),$$

where $V_t(u) = \{v \in \mathcal{V} : d_G(u, v) = t\}$ is the t-hop neighborhood of u. The *directed ℓ-hop edge connectivity* from u to v is then defined as

$$\kappa^{(\ell)}(u \to v) = \max\{k : \text{there exist } k \text{ edge-disjoint } u \to v \text{ paths in } D_\ell(u)\}.$$

By Menger's theorem, this value is equal to the size of the minimum edge cut separating u and v within $D_\ell(u)$. Low values of $\kappa^{(\ell)}$ indicate bottlenecks that are most vulnerable to oversquashing. Given a budget of k edges to add, we first sample $K > k$ candidate node pairs (u, v) and retain only those where v is reachable from u in G. For each remaining pair, we compute $\kappa^{(\ell)}(u \to v)$, rank the pairs by ascending connectivity, and select the top k, excluding edges already in $\mathcal{E}$. Because all shortcuts are chosen from the transitive closure $\mathrm{TC}(G)$, reachability is preserved. In this way, our method reinforces the weakest connections, reducing oversquashing while maintaining locality.

3.3 Combined strategy with R-GNN

Rewiring augments the graph with new edges that play different roles: original edges preserve structural information, while rewired edges improve connectivity. Standard GNNs treat all edges uniformly, leading to suboptimal trade-offs. To address this, we use *relational GNNs* (R-GNNs) [32], which assign separate parameters to different edge types. An R-GNN layer updates node v as

$$\mathbf{h}_v^{(k+1)} = \phi_k\Big(\mathbf{h}_v^{(k)},\ \sum_{r \in \mathcal{R}} \sum_{u \in \mathcal{N}_r(v)} \psi_{k,r}\Big(\mathbf{h}_u^{(k)}, \mathbf{h}_v^{(k)}\Big)\Big), \tag{4}$$

where $\mathcal{R}$ denotes a relation type, such as original or rewired edges. This setup lets the model learn when to rely on local structure and when to exploit rewired edges for long-range communication. By parameterizing $\psi_{k,r}$ separately, the R-GNN balances locality and connectivity, which is especially valuable in directed graphs where both asymmetry and bottlenecks must be respected.

4 Experiments

Additional details of results, datasets and experiment setup are in Appendix B and C.

4.1 Synthetic datasets

Since there are no benchmarks dedicated to long-range dependency and oversquashing in directed graphs, we construct synthetic datasets to probe these properties directly. We design three algorithmic tasks—*cycle detection* (Cycle), *reachability* (Reach.), and *topological sort* (Topo.)—on random directed graphs. These tasks are inherently global: detecting cycles, checking reachability, or recovering a valid topological order cannot be solved using only local information. Directionality is also crucial, as preserving asymmetry is key to faithfully modeling these problems.

Most existing rewiring strategies focus on undirected graphs and rely on spectral properties that do not extend naturally to the directed setting. As a directed baseline, we adopt Stochastic Discrete Ricci Flow (**SDRF**) [17] and compare it with our two proposed methods based on shortcut set **Shortcut** and greedy edge connectivity **Greedy EC**. We use Dir-GCN [10] as the base model. Table 1 shows that both shortcut-based and greedy rewiring outperform the vanilla baseline. Gains are most significant on reachability, where reducing the effective diameter improves information propagation. Greedy edge connectivity achieves the best results across tasks, underscoring the importance of directly targeting structural bottlenecks while preserving locality.

Table 1: Performance of rewiring strategies on synthetic datasets (mean ± std). Best in red, second best in blue.

Method	Cycle	Reach.	Topo.
None	68.46 ± 1.65	71.53 ± 6.93	2.75 ± 0.64
SDRF	68.18 ± 2.11	74.01 ± 4.93	2.51 ± 1.04
Shortcut	68.29 ± 2.18	75.28 ± 8.85	3.01 ± 1.09
Greedy EC	69.09 ± 3.10	77.12 ± 7.23	3.16 ± 0.45

4.2 Real-world datasets

To complement the synthetic experiments, we evaluate on two real-world heterophilic webpage graphs, **Chameleon** and **Squirrel** [33], which are particularly challenging since neighboring nodes often have dissimilar labels and directionality is proved to be informative [10]. We test three base models (Dir-GCN, Dir-GAT, Dir-SAGE [10]). We also include a **Combined** method that applies shortcut set and edge connectivity together, using the same settings as in their individual variants and assigning relation type 1 and 2 in R-GNN, respectively.

Table 2 summarizes the results. On both heterophilic datasets, rewiring consistently improves performance over the baseline, with shortcut sets and greedy edge connectivity providing the largest gains. However, we observe that the Combined variant underperforms the individual strategies, likely due to the difficulty of R-GNN in balancing multiple edge types and the possibility of conflicting signals between the two rewiring methods. Finally, we note that gains on these real-world datasets are smaller than in the synthetic tasks, underscoring the need for benchmarks explicitly designed to test long-range dependency in directed graphs.

Table 2: Performance of different rewiring strategies on real-world datasets (mean ± std). The best is highlighted in red and the second best is highlighted in blue.

Dataset	Model	None	SDRF	Shortcut Set	Greedy EC	Combined
Chameleon	Dir-GCN	78.88 ± 2.25	79.17 ± 1.55	79.60 ± 1.36	79.23 ± 2.13	78.96 ± 1.69
	Dir-SAGE	58.37 ± 1.94	58.91 ± 1.91	58.93 ± 1.79	59.56 ± 1.91	58.29 ± 1.19
	Dir-GAT	63.33 ± 2.17	64.32 ± 2.31	63.46 ± 2.10	64.91 ± 3.34	63.77 ± 3.04
Squirrel	Dir-GCN	73.45 ± 1.25	74.23 ± 1.19	74.59 ± 1.82	75.02 ± 1.58	74.78 ± 1.92
	Dir-SAGE	45.92 ± 2.02	46.09 ± 2.98	46.74 ± 1.99	46.54 ± 2.63	46.42 ± 2.70
	Dir-GAT	65.82 ± 2.90	66.34 ± 2.31	65.92 ± 2.42	66.52 ± 2.19	66.32 ± 2.21

5 Conclusion

We presented two model-agnostic rewiring strategies for directed GNNs—shortcut sets and greedy edge connectivity—that reduce diameter, mitigate oversquashing, and preserve reachability. By integrating these with relational GNNs, the model can adaptively balance local structure with long-range connectivity. Experiments on synthetic tasks tailored for directed long-range reasoning and on real-world citation and webpage graphs demonstrate consistent improvements. These findings establish rewiring as a powerful tool for improving information flow in directed GNNs and underscore the need for benchmarks that more systematically evaluate long-range reasoning in directed settings.

Acknowledgments

YH is supported by a Cubist PhD Fellowship, and EV acknowledges the support of NSF grant CCF-2338226.

References

[1] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017. URL https://arxiv.org/abs/1609.02907. 1

[2] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. URL https://arxiv.org/abs/1710.10903.

[3] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017. URL https://arxiv.org/abs/1704.01212. 1

[4] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation, 2019. URL https://arxiv.org/abs/1902.07243. 1

[5] Yuyang Wang, Zijie Li, and Amir Barati Farimani. *Graph Neural Networks for Molecules*, page 21–66. Springer International Publishing, 2023. ISBN 9783031371967. doi: 10.1007/978-3-031-37196-7_2. URL http://dx.doi.org/10.1007/978-3-031-37196-7_2. 1

[6] Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry, 2022. URL https://arxiv.org/abs/2208.09481. 1

[7] Petar Veličković. Message passing all the way up, 2022. URL https://arxiv.org/abs/2202.11097. 1

[8] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs, 2021. URL https://arxiv.org/abs/2005.00687. 1, 9

[9] Vijay Prakash Dwivedi, Ladislav Rampášek, Mikhail Galkin, Ali Parviz, Guy Wolf, Anh Tuan Luu, and Dominique Beaini. Long range graph benchmark, 2023. URL https://arxiv.org/abs/2206.08164. 1, 2, 9

[10] Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael M. Bronstein. Edge directionality improves learning on heterophilic graphs. In Soledad Villar and Benjamin Chamberlain, editors, *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, pages 25:1–25:27. PMLR, 27–30 Nov 2024. URL https://proceedings.mlr.press/v231/rossi24a.html. 1, 2, 4, 8, 10

[11] Alex Vitvitskyi, João G. M. Araújo, Marc Lackenby, and Petar Veličković. What makes a good feedforward computational graph?, 2025. URL https://arxiv.org/abs/2502.06751. 1

[12] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GtvuNrk58a. 1

[13] Petar Veličković, Adrià Puigdomènech Badia, David Budden, Razvan Pascanu, Andrea Banino, Misha Dashevskiy, Raia Hadsell, and Charles Blundell. The clrs algorithmic reasoning benchmark, 2022. URL https://arxiv.org/abs/2205.15659. 1

[14] Antonio Longa, Veronica Lachi, Gabriele Santin, Monica Bianchini, Bruno Lepri, Pietro Lio, Franco Scarselli, and Andrea Passerini. Graph neural networks for temporal graphs: State of the art, open challenges, and opportunities, 2023. URL https://arxiv.org/abs/2302.01018. 1

[15] Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=i80OPhOCVH2. 1, 2, 7

[16] Kedar Karhadkar, Pradeep Kr. Banerjee, and Guido Montufar. FoSR: First-order spectral rewiring for addressing oversquashing in GNNs. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3YjQfCLdrzz. 1

[17] Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=7UmjRGzp-A. 1, 2, 4, 7

[18] Mitchell Black, Zhengchao Wan, Amir Nayyeri, and Yusu Wang. Understanding oversquashing in GNNs through the lens of effective resistance. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2528–2547. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/black23a.html. 1, 2, 3, 7

[19] Federico Barbero, Ameya Velingker, Amin Saberi, Michael M. Bronstein, and Francesco Di Giovanni. Locality-aware graph rewiring in GNNs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4Ua4hKiAJX. 1, 3

[20] Xitong Zhang, Yixuan He, Nathan Brugnone, Michael Perlmutter, and Matthew Hirn. Magnet: A neural network for directed graphs. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=TRDAFiwDq8A. 2, 7

[21] Simon Geisler, Yujia Li, Daniel J Mankowitz, Ali Taylan Cemgil, Stephan Günnemann, and Cosmin Paduraru. Transformers meet directed graphs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11144–11172. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/geisler23a.html. 2, 8

[22] Yuankai Luo, Veronika Thost, and Lei Shi. Transformers over directed acyclic graphs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=g49s1N5nmO. 8

[23] Christian Koke and Daniel Cremers. Holonets: Spectral convolutions do extend to directed graphs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EhmEwfavOW. 2, 8

[24] Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks, 2020. URL https://arxiv.org/abs/2006.13318. 2

[25] Johannes Gasteiger, Stefan Weiß enberger, and Stephan Günnemann. Diffusion improves graph learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/23c894276a2c5a16470e6a31f4618d73-Paper.pdf. 2, 7

[26] Adrian Arnaiz-Rodriguez, Ahmed Begga, Francisco Escolano, and Nuria Oliver. Diffwire: Inductive graph rewiring via the lovász bound, 2022. URL https://arxiv.org/abs/2206.07369. 2, 7

[27] Jakub Bober, Anthea Monod, Emil Saucan, and Kevin N. Webster. Rewiring networks for graph neural network training using discrete geometry, 2022. URL https://arxiv.org/abs/2207.08026. 2, 7

[28] Igor Sterner, Shiye Su, and Petar Veličković. Commute-time-optimised graphs for gnns, 2024. URL https://arxiv.org/abs/2407.08762. 2, 7

[29] Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João Guilherme Madeira Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=93HCE8vTye. 2

[30] Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. softmax is not enough (for sharp out-of-distribution), 2024. URL https://arxiv.org/abs/2410.01104. 2

[31] J. Ullman and M. Yannakakis. High-probability parallel transitive closure algorithms. In *Proceedings of the Second Annual ACM Symposium on Parallel Algorithms and Architectures*, SPAA '90, page 200–209, New York, NY, USA, 1990. Association for Computing Machinery. ISBN 0897913701. doi: 10.1145/97444.97686. URL https://doi.org/10.1145/97444.97686. 3

[32] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 3

[33] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding, 2021. URL https://arxiv.org/abs/1909.13021. 4, 9

[34] Francesco Di Giovanni, T. Konstantin Rusch, Michael Bronstein, Andreea Deac, Marc Lackenby, Siddhartha Mishra, and Petar Veličković. How does over-squashing affect the power of GNNs? *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=KJRoQvRWNs. 7

[35] Lukas Fesser and Melanie Weber. Mitigating over-smoothing and over-squashing using augmentations of forman-ricci curvature. In Soledad Villar and Benjamin Chamberlain, editors, *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, pages 19:1–19:28. PMLR, 27–30 Nov 2024. URL https://proceedings.mlr.press/v231/fesser24a.html. 7

[36] Mehdi Azabou, Venkataramana Ganesh, Shantanu Thakoor, Chi-Heng Lin, Lakshmi Sathidevi, Ran Liu, Michal Valko, Petar Veličković, and Eva L Dyer. Half-hop: A graph upsampling approach for slowing down message passing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1341–1360. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/azabou23a.html. 7

[37] Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking, 2018. URL https://arxiv.org/abs/1707.03815. 8, 9

[38] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks, 2018. URL https://arxiv.org/abs/1806.03536. 8

[39] Abdol-Hossein Esfahanian. Connectivity algorithms. http://www.cse.msu.edu/~cse835/Papers/Graph_connectivity_revised.pdf, 2010. 10

A Additional Related Works

Oversquashing and graph rewiring. Oversquashing, where signals from many distant nodes are compressed into fixed-size representations, is a well-known limitation of GNNs, particularly in deep architectures and sparse graphs [15, 34]. This challenge arises directly from the topology of the underlying graph. To mitigate it, a large body of work has explored graph rewiring, which modifies the input structure to ease topological bottlenecks. Most existing approaches focus on undirected graphs and rely heavily on spectral tools. Examples include diffusion-based rewiring [25, 26], curvature-driven modifications [17, 27, 35], rewiring via effective resistance [18], and methods that reduce over-smoothing or commute time [28, 36]. While effective for undirected graphs, these strategies do not translate naturally to directed settings. The lack of a theoretical connection between oversquashing and spectral properties in directed graphs highlights the need for alternative approaches. This motivates the development of rewiring strategies tailored to directed graphs.

Directed graph neural networks. More recently, models explicitly designed for directed graphs have emerged to address the shortcomings of undirected-centric methods. MagNet [20], for instance, introduces direction-aware spectral filters using the magnetic Laplacian and complex-valued convolutions. DirGNN [10] distinguishes between incoming and outgoing edges to improve effective homophily in heterophilic graphs, although its advantages diminish when edges are symmetrized. Transformer-based approaches such as Transformers Meet Directed Graphs [21], HoloNets [23], and Transformers over DAGs [22] extend positional encodings, spectral methods, or causal attention to incorporate directionality. While promising, these models rely on architecture-specific modifications to encode directionality. In contrast, rewiring offers a model-agnostic alternative: instead of redesigning message passing, it modifies the graph itself to alleviate bottlenecks and improve information flow. Our work builds on this perspective, introducing a model-agnostic rewiring framework that can be paired with any directed GNN architecture to reduce oversquashing and enhance learning.

B Additional Experiment Results

We additionally report results on the homophilic citation graphs Cora-ML and Citeseer-Full [37]. Unlike the heterophilic datasets in the main text, these benchmarks exhibit strong local label homophily, meaning that neighboring nodes already provide highly informative features. As a result, rewiring yields only minor improvements across all base models, consistent with prior observations that directionality is less critical in homophilic settings [10]. These findings reinforce the complementary nature of our study: while rewiring proves most valuable in heterophilic graphs where long-range dependencies play a larger role, its impact is limited when local neighborhoods already align well with the task. More broadly, this underscores the lack of real-world directed benchmarks that truly stress-test long-range reasoning and oversquashing, highlighting the need for datasets that go beyond citation graphs and webpages and better capture the structural challenges of directed settings.

Table 3: Performance of different rewiring strategies on real-world datasets (mean $\pm$ std).

Dataset	Model	None	SDRF	Shortcut Set	Greedy EC	Combined
Chameleon	Dir-GCN	78.88 ± 2.25	79.17 ± 1.55	79.60 ± 1.36	79.23 ± 2.12	78.96 ± 1.69
	Dir-SAGE	58.37 ± 1.94	58.91 ± 1.91	58.93 ± 1.79	59.56 ± 1.91	58.29 ± 1.19
	Dir-GAT	63.33 ± 2.17	64.32 ± 2.31	63.46 ± 2.10	64.91 ± 3.34	63.77 ± 3.04
Squirrel	Dir-GCN	73.45 ± 1.25	74.23 ± 1.19	74.59 ± 1.82	75.02 ± 1.58	74.78 ± 1.92
	Dir-SAGE	45.92 ± 2.02	46.09 ± 2.98	46.74 ± 1.99	46.54 ± 2.63	46.42 ± 2.70
	Dir-GAT	65.82 ± 2.90	66.34 ± 2.31	65.92 ± 2.42	66.52 ± 2.19	66.32 ± 2.21
Cora-ML	Dir-GCN	84.56 ± 1.54	84.45 ± 0.79	84.40 ± 1.24	84.55 ± 1.10	84.60 ± 1.29
	Dir-SAGE	86.99 ± 1.19	87.16 ± 1.21	87.24 ± 1.01	87.25 ± 1.19	86.99 ± 1.04
	Dir-GAT	86.77 ± 1.25	86.73 ± 1.10	87.03 ± 1.11	86.44 ± 1.31	86.60 ± 1.01
Citeseer-Full	Dir-GCN	93.15 ± 0.35	93.28 ± 0.45	93.36 ± 0.40	93.29 ± 0.59	93.40 ± 0.44
	Dir-SAGE	94.21 ± 0.60	93.88 ± 0.71	93.99 ± 0.75	93.88 ± 0.63	94.09 ± 0.81
	Dir-GAT	94.24 ± 0.68	94.43 ± 0.49	94.29 ± 0.55	94.19 ± 0.60	94.30 ± 0.51

C Experiment Settings

The base models we use are Dir-GCN, Dir-GAT, and Dir-SAGE, as proposed by Rossi et al. [10]. For real-world datasets, we follow their hyperparameter settings. For synthetic datasets, we adopt the following configuration: learning rate of 0.001, hidden dimension of 64, and 3 layers. Following Rossi et al. [10], we apply L2 normalization after each convolutional layer and use max aggregation for jumping knowledge [38].

For rewiring-specific parameters, we use the $\sqrt{n}$-sampling algorithm for shortcut sets on synthetic datasets, where n is the number of nodes. For greedy edge connectivity, we add $k = 100$ edges from a pool of $K = 5000$ sampled candidate pairs. On real-world datasets, we increase the candidate pool to $K = 100{,}000$. In addition, we perform a grid search over key hyperparameters: the number of added edges k (ranging from 10 to 5000), the local neighborhood size ℓ (ranging from 3 to 10), and the aggregation function used in R-GNNs to combine messages across edge relation types, considering `attention`, `sum`, `concat`, `max`, and `mean`.

C.1 Synthetic Datasets

All synthetic datasets are generated from 1,000 random directed acyclic graphs (DAGs), each with 100 nodes. To ensure that oversquashing effects are present, we deliberately inject bottlenecks into the graph generation process. The DAG construction proceeds as follows. Nodes are first partitioned into a random number of clusters (between 5 and 20), with cluster sizes balanced so that all nodes are assigned. Within each cluster, edges are added while preserving acyclicity: a backbone chain connects consecutive nodes, and additional forward edges are included between earlier and later nodes with probability 0.5. This yields dense intra-cluster connectivity while maintaining the DAG property. To link clusters, a small number of "bridge" edges (between 1 and 5) are added from one cluster to the next. These sparse inter-cluster connections act as bottlenecks, forcing information to pass through narrow pathways. The resulting graphs are specifically designed to test long-range dependencies and oversquashing in directed settings.

Cycle Detection. In this task, the goal is to predict for each node whether it participates in a directed cycle, formulated as a binary classification problem. Detecting cycles requires reasoning over potentially long paths, since cycles can span distant parts of the graph. Directionality plays a key role, as reversing even a single edge can alter the presence or absence of a cycle. This task directly probes the ability of a model to capture global structural dependencies.

Reachability. Here we randomly sample a source node and ask the model to predict, for every other node, whether it is reachable from the source. This is framed as a binary node classification problem. Because reachability may depend on paths that extend across the entire graph, the task is highly sensitive to bottlenecks and oversquashing, and therefore serves as a strong test of whether rewiring improves global information flow.

Topological Sort. In this task, the objective is to recover a valid topological ordering of a DAG. Unlike local node-level classification tasks, topological sorting requires capturing global ordering constraints that span the entire graph. Success on this task reflects a model's ability to propagate long-range signals and maintain global structural consistency, making it one of the most challenging and comprehensive evaluations of directed message passing.

C.2 Real-World Datasets

Currently, there is a lack of dedicated directed benchmarks for evaluating directed graph neural networks. Most commonly used datasets convert directed edges into undirected ones [8], which removes important asymmetry in information flow. Even more scarce are real-world datasets that require long-range reasoning. For example, the Long-Range Graph Benchmark [9] has become a standard for testing rewiring methods such as molecular prediction, but it is restricted to the undirected setting. In this work, we select four widely used citation and webpage datasets where directed edges are available. However, these graphs are large and often dominated by local structures, meaning that long-range dependencies may not play a central role and oversquashing bottlenecks may be limited. This further underscores the need for new, dedicated benchmarks designed explicitly for directed graphs and long-range reasoning, a gap that remains largely overlooked despite the growing importance of directed settings.

Cora and Citeseer. These citation networks consist of scientific publications as nodes, with directed citation links as edges [37]. Node features are derived from document word vectors, and labels correspond to research topics. Both datasets are highly homophilic, meaning that papers tend to cite and be cited by others within the same field.

Chameleon and Squirrel. These webpage networks contain nodes representing pages and directed edges corresponding to hyperlinks [33]. Node features are bag-of-words representations of page content, and labels denote categories such as topic or function. Unlike citation graphs, these datasets are heterophilic: linked webpages often belong to different categories, making long-range information more valuable than local neighborhoods.

Table 4: Statistics of the real-world datasets reported by Rossi et al. [10].

Dataset	# Nodes	# Edges	# Features	# Classes	Unidirectional Edges	Homophily
CiteSeer-Full	4,230	5,358	602	6	99.61%	0.949
Cora-ML	2,995	8,416	2,879	7	96.84%	0.792
Chameleon	2,277	36,101	2,325	5	85.01%	0.235
Squirrel	5,201	217,073	2,089	5	90.60%	0.223

D Computational Complexity

The shortcut-set construction follows a $\sqrt{n}$-sampling algorithm that yields $\mathcal{O}(n)$ edges and runs in linear time, where n is the number of nodes. The greedy edge-connectivity computation is more expensive, but we mitigate this by sampling K node pairs and restricting the search to ℓ-hop BFS subgraphs, where K is a hyperparameter and ℓ (typically ≤ 5) is the number of GNN layers. The main bottleneck is computing edge connectivity using Algorithm 1 in [39], which requires $\mathcal{O}(n_\ell \cdot m_\ell)$ time and $\mathcal{O}(n_\ell + m_\ell)$ memory, where n_l and m_l are the number of nodes and edges in the ℓ-hop subgraph. While faster approximation algorithms exist, we leave them for future work. Additionally, we preserve reachability, which can be computed in $\mathcal{O}(n + m)$ time and $\mathcal{O}(n)$ memory via BFS, with more efficient computations for future work. We note that the rewired edges are precomputed once before training and can be further accelerated through CPU parallelism.