Geodesic Distributions Reveal How Heterophily and Bottlenecks Limit the Expressive Power of Message Passing Neural Networks

Jonathan Rubin
Department of Mathematics
Department of Computing
UKRI Centre for Doctoral Training in AI for Healthcare
Imperial College London
jonathan.rubin19@imperial.ac.uk

Sahil Loomba
Department of Mathematics
Imperial College London
s.loomba18@imperial.ac.uk
MIT IDSS
sloomba@mit.edu

Nick S. Jones
Department of Mathematics
Imperial I-X
EPSRC Centre for the Mathematics of Precision Healthcare
Imperial College London
nick.jones@imperial.ac.uk

Abstract

Whilst having shown great success in graph representation learning, message passing neural networks (MPNNs) are known to encounter difficulties in node classification tasks when learning expressive feature representations on certain unfavourable graph structures, especially heterophilic and bottlenecked graphs that have previously been the subject of extensive, but separate, studies. In this paper we develop a theoretical framework to understand the combined effect of heterophily and bottlenecking on the expressive power of MPNNs. We provide a statistical perspective on the performance of the MPNN that decomposes into its expressive power—as measured by "signal sensitivity" that encodes its maximal sensitivity to changes in the mean input features of each node class and ought to be maximised—and generalisation power—as measured by its "noise sensitivity" that ought to be minimised. We then relate signal sensitivity to the graph structure through ℓ-order homophily, a quantity that captures both homophily and bottlenecking behaviour of graphs in a phenomenon we refer to as "homophilic bottlenecking". Pushing the statistical view further by assuming a distribution over graph structures yields a natural decoupling of bottlenecking into two terms measuring underreaching and oversquashing respectively in an ℓ-layer MPNN which makes use of the distribution of geodesic distances up to length ℓ in the graph. Using an asymptotic distribution of geodesic distances in a very general random graph family we can derive tight bounds on ℓ-order homophily, thus providing a complete analytic characterisation of homophilic bottlenecking in MPNNs. Notably, we show that our statistic accurately tracks empirical node classification performance. Our findings offer an interpretable statistical approach for understanding MPNN performance across a variety of graph families, and suggest potentially promising ways to design more powerful MPNNs.

1 Introduction

Message passing graph neural networks. In recent years, graph neural networks (GNNs) have emerged as a powerful paradigm for learning representations of complex structured data [1–3] since

J. Rubin et al., Geodesic Distributions Reveal How Heterophily and Bottlenecks Limit the Expressive Power of Message Passing Neural Networks. *Proceedings of the Second Learning on Graphs Conference (LoG 2023)*, PMLR, Virtual Event, November 27–30, 2023.

they can leverage the rich relational information embedded in graph structures to discover hidden patterns and dependencies. Most GNNs—like Graph Convolutional Networks (GCN) [4] and Graph Attention Networks (GAT) [5]—adhere to the message-passing framework [6, 7] which employs learnable functions to propagate information across the edges of a graph, allowing GNNs to efficiently update each node's representation by transforming and aggregating information from its neighbours. Consider an attributed graph denoted by the tuple $(G, \mathbf{X})$ such that $G = (V, E)$ is the underlying (undirected and simple) graph encoded by its $n \times n$ adjacency matrix $\mathbf{A}$, i.e. $A_{ij} = 1$ if node i has an edge to node j and $A_{ij} = 0$ otherwise, and $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^T$ is the $n \times d_{\text{in}}$ matrix of node feature vectors. Then each node $i \in V$ has an initial feature representation $\mathbf{H}_i^{(0)} \coloneqq \mathbf{X}_i$ which is iteratively updated through a message passing process to obtain a length-d_{out} hidden state representation $\mathbf{H}_i^{(\ell)}$ after ℓ passes according to the following update rule:

$$\mathbf{H}_i^{(\ell)} \coloneqq \phi_\ell \left(\mathbf{H}_i^{(\ell-1)}, \sum_{j \in N(i)} \hat{A}_{ij} \psi_\ell \left(\mathbf{H}_i^{(\ell-1)}, \mathbf{H}_j^{(\ell-1)} \right) \right), \tag{1}$$

where $N(i)$ is the set of neighbours of node i, ℓ represents the layer number, $\hat{\mathbf{A}}$ is a choice of graph shift operator—usually the symmetric normalised adjacency matrix

$$\hat{\mathbf{A}}_{\text{sym}} \coloneqq \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}, \tag{2}$$

where $\mathbf{D} \coloneqq \operatorname{diag}(\mathbf{A}\mathbf{1}_n)$ is the diagonal degree matrix and $\mathbf{1}_n$ is the size-n vector of ones—and the functions ϕ_ℓ and ψ_ℓ are called the update and message functions respectively. However, as we discuss next, message passing neural networks (MPNNs) can be hindered in their ability to capture patterns in certain kinds of graph structures that leads to sub-optimal performance in various applications.

Learning on heterophilic graphs. One challenge that has received much attention is MPNNs' mixed performance on heterophilic graphs [8–10]. Homophily and heterophily measure the tendency of nodes to connect to other nodes who are, respectively, similar and dissimilar to them. Empirically it has been shown that MPNNs perform badly on graphs with lower homophily (i.e. heterophilic graphs), which has heuristically been attributed to a homophilic inductive bias inherent to the message-passing framework. However, the theoretical underpinnings of this phenomenon are still unclear, with recent work suggesting that some cases of heterophily are not necessarily detrimental to MPNN performance [8]. In this regard, the characterisation of harmful heterophily is still an open problem.

Graphs with informational bottlenecks. Another major problem that MPNNs face is of graph bottlenecks [7, 11] that capture the difficulty of propagating information between distant nodes in the graph, owing to their method of *local* message aggregation. This problem is caused by the "oversquashing" of information—into fixed-size vectors due to the exponentially growing receptive field size of the MPNN when aggregating messages across a long path—and underreaching in the MPNN—wherein the limited receptive field size prevents the MPNN from spreading information sufficiently to distant nodes. In this context, a "graph bottleneck" between two nodes is conceptualised as some topological properties of the graph that leads to poor information propagation between them—because of oversquashing and underreaching. For MPNNs, such a graph bottleneck has been previously defined through the Jacobian of the GNN between pairs of nodes, with low values of the Jacobian indicating poor information flow. As the full Jacobian of an ℓ-layer MPNN is proportional to the powers of the graph shift operator [7, 12], powers of the operator capture the topological properties of the graph that lead to bottlenecking.

Prior work. There has been extensive research into both the problem of learning on heterophilic graphs and the graph bottleneck problem, but in separate research strands following disparate approaches. Recent work studying the effects of bottlenecking has primarily employed spectral methods—like using effective resistance and commute times [12, 13]—and mostly targeted graph classification tasks involving small graphs with short diameters. Meanwhile, theoretical accounts of the effects of homophily on MPNNs have been of a statistical nature—for example, by modelling graphs as stochastic block models (SBMs) [8]—and are restricted to node classification tasks since homophily is inherently a measure of class-wise connectivity. In node classification tasks the graphs are typically large and sparse, which limits the tightness of spectrum-derived bounds in Refs. [12, 13].

Our contributions. In this paper we present a consistent theoretical account of the combined effects of heterophily and bottlenecking on MPNN performance through a phenomenon we refer to as *homophilic bottlenecking*—bottlenecking between nodes of the same type. More specifically, we focus on the asymptotically large and sparse regime—where spectral approaches are inadequate—and demonstrate that homophilic bottlenecking restricts the expressive power of MPNNs in node classification tasks. We do so by pursuing a *statistical* framework that allows us to quantify feature expressivity through the "signal sensitivity" of the MPNN, which is its maximal sensitivity to coherent changes in the node feature distribution, that we show is related to graph homophily; see Sec. 2. Modelling the graph itself as a sample from a large family of general random graphs enables a bound for the sensitivity in expectation, providing analytic approximations of bottlenecking that we show tracks empirical classification accuracy very closely; see Sec. 3. We achieve this by using the shortest path length distribution (SPLD) in sparse general graph ensembles [14] to decompose the bottlenecking in terms of oversquashing—that has been previously studied in the literature—and underreaching—in an interpretable geodesic-based formulation that has not appeared in previous analyses [7, 11]—thus providing a complete characterisation of bottlenecking in MPNNs. Appendices B, C, and D provide a complete set of definitions, theorems, and proofs, respectively.

2 Signal sensitivity determines expressivity and is related to homophily

Signal sensitivity of an MPNN. Consider *class-wise* attributed graphs $(G, \mathbf{X})$ wherein each node i has an associated class variable $c_i \in C \coloneqq \{1, 2, \dots, k\}$ such that conditional on it features have the mean vector $\boldsymbol{\mu}_{c_i} \in \mathbb{R}^{d_{\text{in}}}$, and features of a node pair (i, j) have the $d_{\text{in}} \times d_{\text{in}}$ covariance matrix $\boldsymbol{\Sigma}_{ij}$:

$$\mathbb{E}\left[\mathbf{X}_i \mid c_i = c\right] \coloneqq \boldsymbol{\mu}_c, \tag{3a}$$

$$\mathbb{E}\left[\left(\mathbf{X}_i - \boldsymbol{\mu}_c\right)\left(\mathbf{X}_j - \boldsymbol{\mu}_b\right)^T \,\middle|\, c_i = c, c_j = b\right] \coloneqq \boldsymbol{\Sigma}_{ij}, \tag{3b}$$

that is, $\mathbf{X}_i$ are from an arbitrary distribution with finite mean and covariance. Then signal sensitivity $\mathcal{S}_\mu^{(\ell)}(i)$ of the ℓ^{th} MPNN layer for node i is defined as the maximal change in its hidden representation due to changes in the *mean* input representation of the k classes; see Definition 3 for further detail. It can be shown that this definition has an equivalent formulation in terms of graph homophily due to the class structure $\{c_i\}_{i=1}^n$. Let δ_{xy} be the Kronecker delta function, i.e. $\delta_{xy} = 1$ if $x = y$, then:

$$\mathcal{S}_\mu^{(\ell)}(i) = \sup_{\mathbf{X} \in \mathbb{R}^{n \times d_{\text{in}}}} \sum_{j,l=1}^{n} \sum_{p=1}^{d_{\text{out}}} \sum_{q=1}^{d_{\text{in}}} \frac{\partial H_{ip}^{(\ell)}}{\partial X_{jq}} \frac{\partial H_{ip}^{(\ell)}}{\partial X_{lq}} \delta_{c_j c_l}. \tag{4}$$

This provides an initial intuition behind the link between homophily and information propagation through the graph: the product of derivatives $\frac{\partial H_{ip}^{(\ell)}}{\partial X_{jq}} \frac{\partial H_{ip}^{(\ell)}}{\partial X_{lq}}$ measures whether the output dimension p of node i changes in the same or different direction with changes to input dimension q of nodes j and l, while $\delta_{c_j c_l}$ collects terms corresponding to the same class. That is, signal sensitivity is equivalent to the maximal sensitivity to *coherent* changes among features of input nodes of the *same* class. *Importantly, Theorem 1 in Appendix C shows that, for node classification tasks, an ideal model would be the one that* maximises *signal sensitivity (Definition 3) while minimising a related notion of noise sensitivity (Definition 4), that captures the trade-off between generalisation and expressive power.*

Higher-order graph homophily. Lemma 1, that shows how the Jacobian matrix of the output of an MPNN (Eq. (1)) is upper bounded by a function of $\hat{\mathbf{A}}$, alongside Eq. (4) shows that mean signal sensitivity over all nodes, denoted by $\bar{\mathcal{S}}_\mu^{(\ell)}$, is upper bounded by a higher-order generalisation of graph homophily (Definition 5); see Theorem 2. Let $\|\cdot\|$ be the Euclidean norm, and $\nabla_1 f$ and $\nabla_2 f$ be the Jacobian matrices of some function $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with respect to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. Say there exist constants $\alpha_1, \alpha_2, \beta_1, \beta_2$ such that $\forall r \in [\ell]$ the message and update functions satisfy $\|\nabla_1 \phi_r\| \leq \alpha_1$, $\|\nabla_2 \phi_r\| \leq \alpha_2$, $\|\nabla_1 \psi_r\| \leq \beta_1$, and $\|\nabla_2 \psi_r\| \leq \beta_2$. Then, assuming symmetric $\hat{\mathbf{A}}$—as in Eq. (2)—and an isotropic MPNN—for which $\beta_1 = 0$—yields a simpler bound for $\mathcal{S}_\mu^{(\ell)}$:

$$\bar{\mathcal{S}}_\mu^{(\ell)} \leq \sum_{r=0}^{2\ell} \binom{2\ell}{r} \alpha_1^{2\ell - r} (\alpha_2 \beta_2)^r h^r\left(\hat{\mathbf{A}}_{\text{sym}}\right), \tag{5}$$

where $h^r(\cdot)$ is a special case of higher-order weighted homophily (Definition 5):

$$h^r\left(\hat{\mathbf{A}}\right) \coloneqq \frac{1}{n}\sum_{i,j\in V}\left[\hat{\mathbf{A}}^r\right]_{ij}\delta_{c_ic_j}, \tag{6}$$

that we term as r-order homophily. (We remark that the usual notion of node and edge homophily [9] are special cases of $h^1(\cdot)$.) Topping et al. [7] have previously shown how ℓ^{th} power of the graph shift operator relate to performance of the ℓ^{th} MPNN layer—here we show that considering distributional sensitivity globally implicates 2ℓ-order homophily, and that performance is adversely affected by bottlenecks (as characterised by small entries of powers of the operator) between nodes of the *same* class, a phenomenon we refer to as *homophilic bottlenecking*. If we further assume no self-dependence, as in GCNs, then $\alpha_1 = 0$ which yields $\bar{\mathcal{S}}_\mu^{(\ell)} \le (\alpha_2\beta_2)^{2\ell}h^{2\ell}\left(\hat{\mathbf{A}}_{\text{sym}}\right)$. This offers new theoretical insight into the phenomenon of oversmoothing often observed in GCNs: as the number of layers increases, the performance of the GCN deteriorates [15, 16].

3 Analytic estimates of signal sensitivity track empirical performance

Analytic approximations using the SPLD. Consider the (undirected and simple) graph G to be a sample from a general random graph family with conditionally independent edges, that is, without loss of generality for node indices $i < j : A_{ij} \sim \text{Bernoulli}\left([\mathbb{E}[\mathbf{A}]]_{ij}\right)$ and $A_{ji} = A_{ij}$. In other words, the graph ensemble is completely characterised by the expected adjacency matrix $\mathbb{E}[\mathbf{A}]$ and includes many random graph models like stochastic block models (SBMs [17]), random dot product graphs [18], etc. From Eqs. (5) and (6) we can bound the mean signal sensitivity in expectation:

$$\mathbb{E}\left[\bar{\mathcal{S}}_\mu^{(\ell)}\right] \le \sum_{r=0}^{2\ell}\binom{2\ell}{r}\alpha_1^{2\ell-r}(\alpha_2\beta_2)^r\,\mathbb{E}\left[h^r\left(\hat{\mathbf{A}}_{\text{sym}}\right)\right], \quad \mathbb{E}\left[h^r\left(\hat{\mathbf{A}}\right)\right] = \frac{1}{n}\sum_{i,j\in V}\mathbb{E}\left[\hat{\mathbf{A}}^r\right]_{ij}\delta_{c_ic_j}. \tag{7}$$

Let $\lambda_{ij} \in \mathbb{Z}$ denote the shortest path length between nodes i, j. The ensemble endows the lengths with a distribution, allowing us to decompose the expectation of powers of the graph shift operator as:

$$\mathbb{E}\left[\hat{\mathbf{A}}^r\right]_{ij} = \underbrace{\mathbb{E}\left[\left[\hat{\mathbf{A}}^r\right]_{ij}\middle|\lambda_{ij} \le r\right]}_{\text{oversquashing}}\underbrace{\mathbb{P}\left(\lambda_{ij} \le r\right)}_{\text{underreaching}}, \tag{8}$$

where we use the fact that $\lambda_{ij} > r \implies \left[\hat{\mathbf{A}}^r\right]_{ij} = 0$. The first factor on the RHS measures a form of density within the receptive field of node i while the second factor measures the probability that the target node j is within the receptive field of i, and therefore the two factors can be seen as encoding oversquashing and underreaching, respectively. Assuming that the network is sparse, i.e. $\mathbb{E}[\mathbf{A}] = O\left(n^{-1}\right)$ implying asymptotically (as $n \to \infty$) bounded node degrees, we can use prior results on the asymptotic SPLD [14] in sparse graph ensembles to yield, assuming each node is on the giant component with probability $1 - o(1)$, the underreaching factor in Eq. (8) purely in terms of $\mathbb{E}[\mathbf{A}]$; see Lemma 2. The oversquashing factor is more difficult to calculate, however, similar to Topping et al. [7], we can calculate a tight bound for oversquashing at the boundary of the receptive field $\mathbb{E}\left[\left[\hat{\mathbf{A}}^r\right]_{ij}\middle|\lambda_{ij} = r\right]$ [7] purely in terms of $\mathbb{E}[\mathbf{A}]$; see Theorem 3 for when $\hat{\mathbf{A}} \coloneqq \hat{\mathbf{A}}_{\text{sym}}$. For $\ell = 1$ Eq. (7) suggests that we only need first and second order homophily to determine the MPNN's performance which—using Eq. (8), Lemma 2 and Theorem 3—we can derive (tight) analytic bounds for any such graph ensemble; see Corollary 3.1 for the case of sparse SBMs. For simplicity, consider a 2-block "planted partition" sparse SBM with two equi-sized classes such that $\mathbb{E}[\mathbf{A}]_{ij} \coloneqq \frac{B_{c_ic_j}}{n}$ where $\mathbf{B} \coloneqq 2d\begin{bmatrix} h & (1-h) \\ (1-h) & h \end{bmatrix}$ and $d > 0$ controls the mean degree of every node while $0 \le h \le 1$ controls the (edge) homophily [9]. Application of Corollary 3.1 yields:

$$\mathbb{E}\left[h^1\left(\hat{\mathbf{A}}_{\text{sym}}\right)\right] \lessapprox h, \quad \mathbb{E}\left[h^2\left(\hat{\mathbf{A}}_{\text{sym}}\right)\right] \lessapprox \frac{1}{d} + \left(1 - \frac{1 - e^{-d}}{d}\right)\left(2h^2 - 2h + 1\right). \tag{9}$$

Figure 1 shows that our analytic estimates strongly track empirical node classification performance. Notable is the quadratic variation of performance with homophily (Eq. (9)): In the case of GCNs performance is symmetric around "ambiphily" ($h = 0.5$), and almost equal for extremely heterophilic ($h = 0$) and homophilic ($h = 1$) graphs. We also validate on real-world graphs in Figure 2.

4 Conclusion

In this paper we have presented a statistical framework to show that classification accuracy is theoretically determined by the sensitivity of MPNN to changes to the class-wise mean of node features, as measured through signal sensitivity (Eq. (4)), that we validate empirically. We proved that signal sensitivity is in turn adversely impacted by information bottlenecks between nodes of the same class (Eqs. (7), (6)), in a phenomenon we termed as homophilic bottlenecking, which provides a theoretical explanation for empirically poor MPNN performance on heterophilic graphs. We were able to decompose signal sensitivity into oversquashing and underreaching factors (Eq. (8)) using shortest path length distribution which we used to derive analytic expressions for signal sensitivity (Eq. (9)) that were empirically validated. Overall, our work combines homophily, bottlenecks and feature expressivity in MPNNs within a unifying theoretical framework that we hope enables new MPNN design choices motivated from analytic principles.

Disclosure of Funding

J. R. is supported by the UK Research and Innovation (UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1).

References

[1] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24, 2020. URL https://arxiv.org/abs/1901.00596. 1

[2] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. URL https://arxiv.org/abs/2104.13478.

[3] Lilapati Waikhom and Ripon Patgiri. Graph neural networks: Methods, applications, and opportunities, 2021. URL https://arxiv.org/abs/2108.10733. 1

[4] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. URL http://arxiv.org/abs/1609.02907. 2

[5] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ. 2

[6] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*, 2017. URL https://arxiv.org/abs/1704.01212. 2

[7] 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. URL https://arxiv.org/abs/2111.14522. 2, 3, 4, 9, 13

[8] Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? *arXiv preprint arXiv:2106.06134*, 2021. URL https://arxiv.org/abs/2106.06134. 2

[9] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs, 2020. URL https://arxiv.org/abs/2006.11468. 4

[10] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks, 2022. URL https://arxiv.org/abs/2210.07606. 2

[11] Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications, 2021. URL https://arxiv.org/abs/2006.05205. 2, 3

[12] Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio', and Michael Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology, 2023. URL https://arxiv.org/abs/2302.02941. 2

[13] Mitchell Black, Zhengchao Wan, Amir Nayyeri, and Yusu Wang. Understanding Oversquashing in GNNs through the Lens of Effective Resistance, 2023. URL https://arxiv.org/abs/2302.06835. 2

[14] Sahil Loomba and Nick S. Jones. Geodesic statistics for random network families, 2021. URL https://arxiv.org/abs/2111.02330. 3, 4, 13, 17, 18, 19

[15] Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1909.12223. 4

[16] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 3538–3545, 2018. URL https://arxiv.org/abs/1801.07606. 4

[17] Paul W. Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. ISSN 0378-8733. doi: 10.1016/0378-8733(83)90021-7. URL https://www.sciencedirect.com/science/article/pii/0378873383900217. 4, 8

[18] Stephen J. Young and Edward R. Scheinerman. Random dot product graph models for social networks. In *Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph*, WAW'07, page 138–149, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3540770038. doi: 10.5555/1777879.1777890. 4

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. 8

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114. 9

[21] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Education, 1976. 21

A Figures

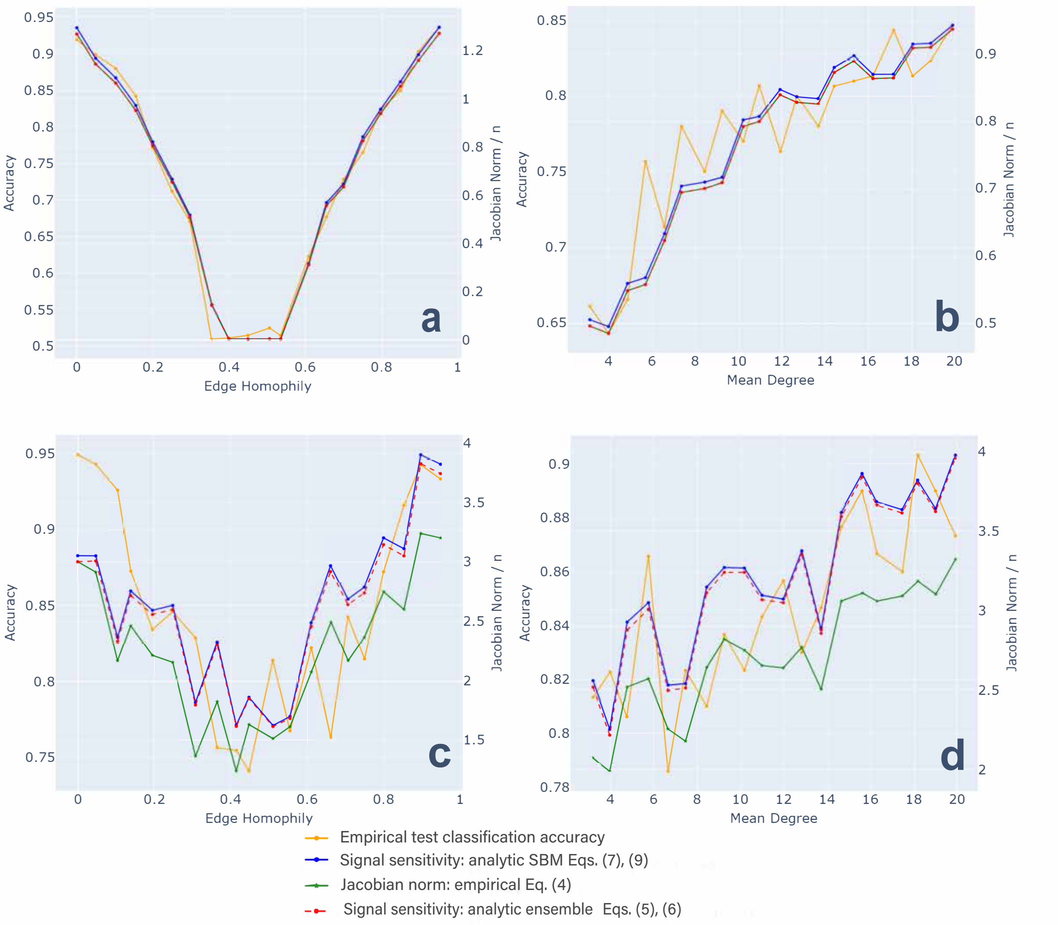


Figure 1: Empirical and analytic estimates of (root) mean signal sensitivity accurately track empirical node classification performance for a 2-block planted partition SBM. For a GCN without self-dependence (**a**, **b**; $\alpha_1 = 0$ in Eq. (5)) and linear isotropic MPNN with self-dependence (**c**, **d**; $\alpha_1 > 0$ in Eq. (5)) in the update function (Eq. (1)) while assuming $\hat{\mathbf{A}}_{\text{sym}}$ (Eq. (2)) as the graph shift operator, when varying the edge homophily while keeping the mean degree fixed at $d = 5$ (**a**, **c**) and when varying the mean degree while keeping the edge homophily fixed at $h = 0.7$ (**b**, **d**). The graph with $n = 3000$ nodes was sampled from a 2-block equi-sized SBM [17] with block matrix $\mathbf{B} := 2d \begin{bmatrix} h & (1-h) \\ (1-h) & h \end{bmatrix}$ and the class-wise feature distribution (Eq. (3)) was assumed to be a bivariate normal with class-wise means $(1, 0)$ and $(0, 1)$ respectively, and class-wise covariance $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ for both blocks. The MPNN models (yellow) were trained with a weight decay of 10^{-3} to minimise generalisation error and ensure the performance difference was due to variation in expressive power. The Jacobian norm (green) was empirically calculated, as the argument to the $\sup$ in Eq. (4), for each experiment instance using the PyTorch Autograd package [19]. The analytic (root) mean signal sensitivity was computed for the full graph (red) using Eqs. (5) and (6) and for the corresponding SBM (blue) using Eqs. (7) and (9). For an analysis of real-world graphs see Figure 2.

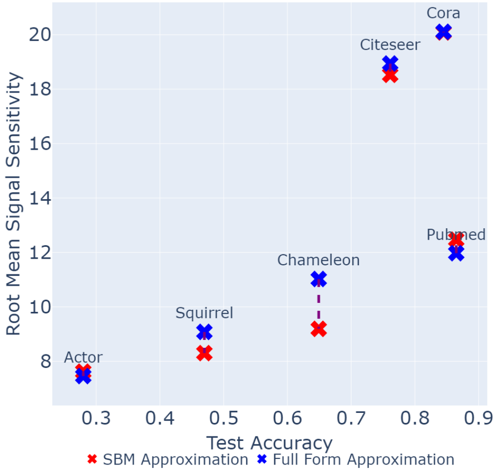


Figure 2: Analytic estimates of (root) mean signal sensitivity positively covary with empirical node classification performance for multiple real-world graphs. Stochastic block models (SBMs) were inferred for the given real-world graphs via maximum likelihood estimation, given the node classes. The analytic (root) mean signal sensitivity was computed for the full graph (blue) using Eqs. (5) and (6) and for the corresponding SBM (red) using Eqs. (7) and (9). Both are strongly correlated with each other and with the empirical test classification accuracy for a GCN, but we note a few deviations. First, the analytics from the SBM vs. the full graph for Chameleon and Squirrel have small differences, that can be partly explained due to the SBM not being as good a model fit as for the other datasets. (We remark that the log-likelihood of the real-world datasets under their assumed SBMs are -0.076, -0.066, -0.011, -0.009, -0.008, and -0.002 for the Squirrel, Chameleon, Cora, Actor, Citeseer, and Pubmed datasets, respectively.) Second, the analytics deviate from the empirical test accuracies for Citeseer and Cora—that appear to have a lower accuracy than what would be predicted by signal sensitivity alone—which can be explained by noting that noise sensitivity also contributes to model performance, as captured by Theorem 1 and supported by additional analyses in Figure 3. (In the real-world datasets considered, noise sensitivity plays a significant role when the total feature variance is higher, say as a result of a large number of feature dimensions when compared to the synthetic datasets considered in Figure 1.)

B Definitions

An MPNN's sensitivity is typically measured by the Jacobian with respect to input features.

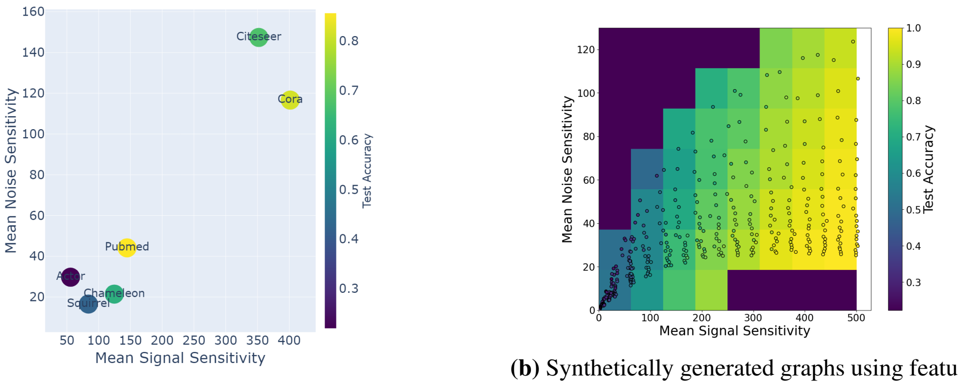
Definition 1 (MPNN Jacobian; Topping et al. [7])**.** *Let $(G, \mathbf{X})$ be an attributed graph, i.e. a graph G of n nodes with node feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d_{\text{in}}}$. Let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{n \times d_{\text{out}}}$ be the matrix of output node embeddings of the ℓ^{th} MPNN layer then the $n \times n \times d_{\text{in}} \times d_{\text{out}}$ Jacobian tensor of the ℓ^{th} MPNN layer is defined as:*

$$\left[\mathcal{J}^{(\ell)}\right]_{ijpq} := \frac{\partial H_{ip}^{(\ell)}}{\partial X_{jq}},$$

where the index i refers to the output node, j refers to the input node, p and q refer to the output and input feature dimensions respectively, and ℓ refers to the layer.

Remark on tensor indexing. Indexing into a four-dimensional tensor $\mathcal{T}$ as $[\mathcal{T}]_i$ corresponds to a three-dimensional tensor and as $[\mathcal{T}]_{ij}$ corresponds to a two-dimensional tensor (i.e. matrix).

Consider a reparameterisation of node i's feature vector in terms of its class-wise mean vector and corresponding residual or "noise" vector $\mathbf{X}_i = \boldsymbol{\mu}_{c_i} + \boldsymbol{\epsilon}_i$, akin to the reparameterisation used in variational autoencoders to learn latent data distributions in a differentiable manner [20]. Analysing



(a) Real-world graphs

(b) Synthetically generated graphs using features and labels from Cora dataset

Figure 3: Together, signal sensitivity and noise sensitivity can account for empirically observed variations in node classification performance of MPNNs. The x-axis measures signal sensitivity as in Eq. (5), the y-axis measures noise sensitivity as in Eq. (14), while the colour encodes node classification test accuracy i.e. empirical model performance for a graph. Figure 3a highlights, for the real-world graphs used in Figure 2, that the lower-than-expected model performance for Citeseer and Cora can be explained by a large noise sensitivity, as predicted by Theorem 1. Figure 3b provides further evidence by considering 400 synthetic graphs of varying mean signal and noise sensitivities—generated under a planted partition SBM with different levels of mean degree and edge homophily but with node features and labels corresponding to those in Cora to match its variance. The emerging trend corroborates the prediction of Theorem 1—superior model performance is associated with heightened signal sensitivity and lowered noise sensitivity.

the Jacobian with respect to the noise—by using $\frac{\partial \mathbf{X}_i}{\partial \boldsymbol{\epsilon}_j} = \boldsymbol{\delta}_{ij}$ and the chain rule—yields:

$$\frac{\partial H_{ip}^{(\ell)}}{\partial \epsilon_{jq}} = \sum_{l=1}^{n} \frac{\partial H_{ip}^{(\ell)}}{\partial X_{lq}} \delta_{lj} = \frac{\partial H_{ip}^{(\ell)}}{\partial X_{jq}}. \tag{10}$$

In other words, the sensitivity of the MPNN's output for node i with respect to the features of node j is—at least in part—due to the residual noise in the feature distribution. Therefore, in this paper, we argue that MPNN performance can be better understood using the Jacobian with respect to the class-wise means instead.

Definition 2 (Class-reduced Jacobian)**.** *Let $(G, \mathbf{X})$ be a class-wise attributed graph with k classes following Eq. (3). Let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{n \times d_{out}}$ be the matrix of output node embeddings of the ℓ^{th} MPNN layer then the $n \times k \times d_{\text{in}} \times d_{\text{out}}$ class-reduced Jacobian tensor of the ℓ^{th} MPNN layer is defined as:*

$$\left[\mathcal{J}_{\mu}^{(\ell)}\right]_{iupq} := \frac{\partial H_{ip}^{(\ell)}}{\partial \mu_{uq}},$$

where index i refers to the output node, u refers to the class of the input node, p and q refer to the output and input feature dimensions respectively, and ℓ refers to the layer.

Using the mean-residual decomposition from above and the chain rule yields:

$$\left[\mathcal{J}_{\mu}^{(\ell)}\right]_{iupq} = \sum_{j=1}^{n} \frac{\partial H_{ip}^{(\ell)}}{\partial X_{jq}} \delta_{c_j u}, \tag{11}$$

which can now be used to study the *distributional* sensitivity of MPNNs, by defining what we term as *signal sensitivity*.

Definition 3 (Signal sensitivity)**.** *Let $(G, \mathbf{X})$ be a class-wise attributed graph with k classes following Eq. (3) and $\mathcal{J}_{\mu}^{(\ell)}$ be the class-reduced Jacobian of the ℓ^{th} MPNN layer (Definition 2) then the signal sensitivity of the ℓ^{th} MPNN layer for node i is defined as:*

$$\mathcal{S}_{\mu}^{(\ell)}(i) := \sup_{\mathbf{M} \in \mathbb{R}^{k \times d_{\text{in}}}} \left\| \left[\mathcal{J}_{\mu}^{(\ell)}\right]_i \right\|^2,$$

where $\|\cdot\|$ is the Euclidean norm and $\mathbf{M}$ refers to the $k \times d_{\text{in}}$ matrix of class-wise mean input vectors. Furthermore, the mean signal sensitivity of the ℓ^{th} MPNN layer is defined as:

$$\bar{\mathcal{S}}_\mu^{(\ell)} := \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_\mu^{(\ell)}(i).$$

We similarly define the notion of *noise sensitivity*.

Definition 4 (Noise sensitivity)**.** *Let $(G, \mathbf{X})$ be a class-wise attributed graph with k classes following Eq. (3) and $\mathcal{J}^{(\ell)}$ be the Jacobian of the ℓ^{th} MPNN layer (Definition 1) then the noise sensitivity of the ℓ^{th} MPNN layer for node i is defined as:*

$$\mathcal{S}_\epsilon^{(\ell)}(i) := \sup_{\mathbf{E}\in\mathbb{R}^{n\times d_{\text{in}}}} \left\| \left[\mathcal{J}^{(\ell)}\right]_i \right\|^2,$$

where $\|\cdot\|$ is the Euclidean norm and $\mathbf{E}$ refers to the $n \times d_{\text{in}}$ matrix of the difference of each node's input feature vector from its class's mean input vector. Furthermore, the mean noise sensitivity of the ℓ^{th} MPNN layer is defined as:

$$\bar{\mathcal{S}}_\epsilon^{(\ell)} := \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_\epsilon^{(\ell)}(i).$$

We show that the graph structural determinants of MPNN performance in node classification tasks can be succinctly captured by the following higher-order weighted generalisation of homophily.

Definition 5 ((r, s)-order (u, v)-weighted homophily)**.** *Let $\hat{\mathbf{A}}$ be a choice of graph shift operator of graph $G = (V, E)$ with n nodes and $\boldsymbol{w}$ be a vector of node weights used to define the matrix:*

$$\mathfrak{H}_u^r\left(\hat{\mathbf{A}}, \boldsymbol{w}\right) := \binom{r+u}{r}^{-1} \sum_{\substack{\boldsymbol{p}\in\{0,1\}^{r+u} \\ \sum_a p_a = r}} \prod_{a=1}^{r+u} \hat{\mathbf{A}}^{p_a} \left[\operatorname{diag}\left(\boldsymbol{w}\right)\right]^{1-p_a}, \tag{12}$$

then (r, s)-order (u, v)-weighted homophily is an (appropriately weighted) average proportion of nodes of the same class that are respectively in the r-order and s-order neighbourhoods of a node:

$$h_{u,v}^{r,s}\left(\hat{\mathbf{A}}, \boldsymbol{w}\right) := \frac{1}{n} \sum_{i,j\in V} \left[\mathfrak{H}_u^r\left(\hat{\mathbf{A}}^T, \boldsymbol{w}\right) \mathfrak{H}_v^s\left(\hat{\mathbf{A}}, \boldsymbol{w}\right)\right]_{ij} \delta_{c_i c_j},$$

where δ_{c_i, c_j} is the Kronecker delta function. In particular, if $\boldsymbol{w} = c\mathbf{1}_n$ for some $c \geq 0$, where $\mathbf{1}_n$ is the vector of ones of size n, then

$$\mathfrak{H}_u^r\left(\hat{\mathbf{A}}, c\mathbf{1}_n\right) = c^u \hat{\mathbf{A}}^r.$$

C Theorems

This section states the key results while Appendix D.1 provides their proofs.

Theorem 1 (Signal sensitivity bounds MPNN performance)**.** *Let $(G, \mathbf{X})$ and $(G, \widetilde{\mathbf{X}})$ be class-wise attributed graphs—with k classes following Eq. (3)—consisting of the same graph G with n nodes but different input features $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, corresponding to (potentially different) class-wise means $\{\boldsymbol{\mu}_u\}_{u=1}^k$, $\{\widetilde{\boldsymbol{\mu}}_u\}_{u=1}^k$ and node-wise covariances $\{\boldsymbol{\Sigma}_{ii}\}_{i=1}^n$, $\{\widetilde{\boldsymbol{\Sigma}}_{ii}\}_{i=1}^n$. Let the hidden feature representation for node i be considered as a differentiable function $H_i^{(\ell)} : \mathbb{R}^{n\times d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ of the input features matrix. Then the expected squared distance between the output embeddings of the ℓ^{th} MPNN layer, given the graph G, is bounded by:*

$$\mathbb{E}\left[\left\|H_i^{(\ell)}(\mathbf{X}) - H_i^{(\ell)}(\widetilde{\mathbf{X}})\right\|^2 \,\middle|\, G\right] \leq \mathcal{S}_\mu^{(\ell)}(i) \sum_{u=1}^{k} \|\boldsymbol{\mu}_u - \widetilde{\boldsymbol{\mu}}_u\|^2 + \mathcal{S}_\epsilon^{(\ell)}(i) \sum_{j=1}^{n} \operatorname{Tr}\left(\boldsymbol{\Sigma}_{jj} + \widetilde{\boldsymbol{\Sigma}}_{jj}\right), \tag{13}$$

where $\mathcal{S}_\mu^{(\ell)}(i)$ and $\mathcal{S}_\epsilon^{(\ell)}(i)$ refer to the signal sensitivity (Definition 3) and noise sensitivity (Definition 4) of the ℓ^{th} MPNN layer for node i, respectively.

Interpretation for Theorem 1. Consider $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ sampled from two different class-wise distributions with class-wise means $\boldsymbol{\mu}_u$ and $\widetilde{\boldsymbol{\mu}}_u$ being very close. Then, ideally, the resulting output embeddings should also be close, and the bound in Theorem 1 enforces a small distance between the output embeddings when the noise sensitivity $\mathcal{S}_{\epsilon}^{(\ell)}(i)$ is small provided—thus, minimising noise sensitivity would ensure that the output embeddings would be close. On the other hand, when $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ are sampled from two distributions with very different class-wise means $\boldsymbol{\mu}_u$ and $\widetilde{\boldsymbol{\mu}}_u$, then the resulting output embeddings should ideally also be distant. However, the bound in Theorem 1 means that when both $\mathcal{S}_{\epsilon}^{(\ell)}(i)$ and $\mathcal{S}_{\mu}^{(\ell)}(i)$ are small then the output embeddings are close together, which is undesirable when $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ are sampled from two very different distributions. Therefore, an ideal model minimises $\mathcal{S}_{\epsilon}^{(\ell)}(i)$ whilst maximising $\mathcal{S}_{\mu}^{(\ell)}(i)$ to have the best discriminative power. The trade-off described here precisely reflects the trade-off between generalisation and expressive power of a given model—between sensitivity to noise which indicates overfitting and sensitivity to true changes in the feature distribution which indicates an expressive model.

Lemma 1 (Bound for MPNN Jacobian). *Let $\mathcal{J}^{(\ell)}$ be the Jacobian of the ℓ^{th} layer of an MPNN (Definition 1) that uses the graph shift operator $\hat{\mathbf{A}}$ with message and update functions $\{\psi_k(\cdot,\cdot)\}_{k=1}^{\ell}$ and $\{\phi_k(\cdot,\cdot)\}_{k=1}^{\ell}$, as in Eq. (1). Let $\|\cdot\|$ be the Euclidean norm, and $\nabla_1 f$ and $\nabla_2 f$ be the Jacobian matrices of some function $f(\boldsymbol{x}_1,\boldsymbol{x}_2)$ with respect to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. Assuming that there exist constants $\alpha_1,\alpha_2,\beta_1,\beta_2$ such that $\forall r\in[\ell]$ the message and update functions satisfy $\|\nabla_1\phi_r\|\le\alpha_1$, $\|\nabla_2\phi_r\|\le\alpha_2$, $\|\nabla_1\psi_r\|\le\beta_1$, and $\|\nabla_2\psi_r\|\le\beta_2$ then:*

$$\left\|\left[\mathcal{J}^{(\ell)}\right]_{ij}\right\| \le \left[\left(\alpha_2\beta_2\hat{\mathbf{A}}+\alpha_2\beta_1\operatorname{diag}\left(\hat{\mathbf{A}}\mathbf{1}_n\right)+\alpha_1\mathbf{I}_n\right)^{\ell}\right]_{ij},$$

where $\mathbf{1}_n$ is the size-n vector of ones and $\mathbf{I}_n$ is the identity matrix of size n.

Theorem 2 (Higher-order weighted homophily bounds signal sensitivity). *Let $(G,\mathbf{X})$ be a class-wise attributed graph with k classes following Eq. (3) and $\bar{\mathcal{S}}_{\mu}^{(\ell)}$ be the mean signal sensitivity of the ℓ^{th} layer (Definition 3) of an MPNN that uses the graph shift operator $\hat{\mathbf{A}}$ with message and update functions $\{\psi_k(\cdot,\cdot)\}_{k=1}^{\ell}$ and $\{\phi_k(\cdot,\cdot)\}_{k=1}^{\ell}$, as in Eq. (1). Let $\|\cdot\|$ be the Euclidean norm, $\nabla_1 f$ and $\nabla_2 f$ be the Jacobian matrices of some function $f(\boldsymbol{x}_1,\boldsymbol{x}_2)$ with respect to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. Assuming that there exist constants $\alpha_1,\alpha_2,\beta_1,\beta_2$ such that $\forall r\in[\ell]$ the message and update functions satisfy $\|\nabla_1\phi_r\|\le\alpha_1$, $\|\nabla_2\phi_r\|\le\alpha_2$, $\|\nabla_1\psi_r\|\le\beta_1$, and $\|\nabla_2\psi_r\|\le\beta_2$ then:*

$$\bar{\mathcal{S}}_{\mu}^{(\ell)} \le \sum_{r,s=0}^{\ell}\sum_{u=0}^{r}\sum_{v=0}^{s}\binom{\ell}{r}\binom{\ell}{s}\binom{r}{u}\binom{s}{v}\alpha_1^{2\ell-r-s}\left(\alpha_2\beta_2\right)^{r+s}h_{u,v}^{r-u,s-v}\left(\hat{\mathbf{A}},\beta_1\beta_2^{-1}\hat{\mathbf{A}}\mathbf{1}_n\right),$$

where $h_{u,v}^{r,s}(\cdot,\cdot)$ is the (r,s)-order (u,v)-weighted homophily given by Definition 5.

Interpretation for Theorem 2. This theorem shows how the signal sensitivity of an MPNN—which by Theorem 1 would determine its node classification performance—depends on the graph structure as encoded by the graph shift operator $\hat{\mathbf{A}}$—in particular by a higher-order homophily that accounts for an appropriately weighted sum of walks between nodes of the same class.

Corollary 2.1 (r-order homophily bounds signal sensitivity in isotropic MPNNs with a symmetric graph shift operator). *Let $(G,\mathbf{X})$ be a class-wise attributed graph with k classes following Eq. (3) and $\bar{\mathcal{S}}_{\mu}^{(\ell)}$ be the mean signal sensitivity of the ℓ^{th} layer (Definition 3) of an MPNN that uses a symmetric graph shift operator $\hat{\mathbf{A}}$ with message and update functions $\{\psi_k(\cdot,\cdot)\}_{k=1}^{\ell}$ and $\{\phi_k(\cdot,\cdot)\}_{k=1}^{\ell}$, as in Eq. (1), that satisfy the conditions in Theorem 2 with the additional constraint that $\beta_1=0$ (for isotropic MPNNs), then:*

$$\bar{\mathcal{S}}_{\mu}^{(\ell)} \le \sum_{r=0}^{2\ell}\binom{2\ell}{r}\alpha_1^{2\ell-r}\left(\alpha_2\beta_2\right)^{r}h^{r}\left(\hat{\mathbf{A}}\right),$$

where $h^r(\cdot)$ is the r-order homophily defined in Eq. (6).

Corollary 2.2 (Weighted sum of closed walks bounds noise sensitivity in isotropic MPNNs with a symmetric graph shift operator). *Let $(G,\mathbf{X})$ be a class-wise attributed graph with k classes following Eq. (3) and $\bar{\mathcal{S}}_{\epsilon}^{(\ell)}$ be the mean noise sensitivity of the ℓ^{th} layer (Definition 4) of an MPNN that uses a symmetric graph shift operator $\hat{\mathbf{A}}$ with message and update functions $\{\psi_k(\cdot,\cdot)\}_{k=1}^{\ell}$ and*

$\{\phi_k(\cdot,\cdot)\}_{k=1}^{\ell}$, *as in Eq.* (1), *that satisfy the conditions in Theorem* 2 *with the additional constraint that* $\beta_1 = 0$ *(for isotropic MPNNs), then:*

$$\bar{\mathcal{S}}_{\epsilon}^{(\ell)} \leq \frac{1}{n} \sum_{r=0}^{2\ell} \binom{2\ell}{r} \alpha_1^{2\ell-r} \left(\alpha_2 \beta_2\right)^r \operatorname{Tr}\left(\hat{\mathbf{A}}^r\right), \tag{14}$$

where $\operatorname{Tr}(\cdot)$ *is the matrix trace.*

Lemma 2 (Underreaching in MPNNs for sparse graph ensembles; Loomba and Jones [14])**.** *For an undirected and simple graph* G *with* n *nodes encoded by the adjacency matrix* $\mathbf{A}$*, sampled from a general random graph family with conditionally independent edges and expected adjacency matrix* $\mathbb{E}[\mathbf{A}]$*, if the network is sparse in the sense that* $\mathbb{E}[\mathbf{A}] = O\left(n^{-1}\right)$*, each node is on the giant component with probability* $1 - o(1)$*, then asymptotically the cumulative distribution function of the length of the shortest path* λ_{ij} *between nodes* i *and* $j \neq i$ *is given by:*

$$\mathbb{P}\left(\lambda_{ij} \leq r\right) \approx \left[\sum_{s=1}^{r} \mathbb{E}[\mathbf{A}]^s\right]_{ij}$$

where "$\approx$*" indicates an asymptotic first-order approximation as* $n \to \infty$*.*

Theorem 3 (Boundary oversquashing in MPNNs for sparse graph ensembles)**.** *Assume the same conditions as in Lemma* 2*, and additionally assume large expected node degrees encoded in the diagonal matrix* $\langle \mathbf{D} \rangle := \operatorname{diag}\left(\mathbb{E}[\mathbf{A}]\,\mathbf{1}_n\right)$ *where* $\mathbf{1}_n$ *is the length-*n *vector of ones. Then for the symmetric normalised adjacency matrix* $\hat{\mathbf{A}}_{\mathrm{sym}}$*(Eq.* (2)*) the boundary oversquashing between nodes* i *and* $j \neq i$*, where* λ_{ij} *is the shortest path distance from* i *to* j*, is asymptotically bounded by:*

$$\mathbb{E}\left[\left[\hat{\mathbf{A}}_{\mathrm{sym}}^r\right]_{ij} \middle| \lambda_{ij} = r\right] \lessapprox \frac{\left[\langle \mathbf{D} \rangle^{-\frac{1}{2}} \mathbb{E}[\mathbf{A}] \left(\left\{\langle \mathbf{D} \rangle^{-1} - \langle \mathbf{D} \rangle^{-2}\left(\mathbf{I}_n - e^{-\langle \mathbf{D} \rangle}\right)\right\} \mathbb{E}[\mathbf{A}]\right)^{r-1} \langle \mathbf{D} \rangle^{-\frac{1}{2}}\right]_{ij}}{\left[\mathbb{E}[\mathbf{A}]^r\right]_{ij}}, \tag{15}$$

where $\mathbf{I}_n$ *is the size-*n *identity matrix, and the bound gets tighter with larger mean degrees.*

Interpretation for Theorem 3. This theorem provides an alternate bound to the bound given in Theorem 4 in Topping et al. [7] where, instead of an absolute bound in terms of edge curvature, we bound the boundary oversquashing in expectation. It can effectively lend itself to an efficient variational rewiring procedure for alleviating oversquashing, by inferring a random graph model that maximises the likelihood of observing the given graph whilst also minimsing oversquashing through maximising Eq. (15). We emphasise that as the mean node degrees and network size become appropriately large then the bound becomes an asymptotic equality.

Corollary 3.1 (Bound for first and second order homophily in sparse SBMs)**.** *Consider an undirected and simple graph* G *with* n *nodes encoded by the adjacency matrix* $\mathbf{A}$ *sampled from a sparse stochastic block model (SBM) such that node classes are i.i.d. as per* $c \sim \operatorname{Categorical}(\boldsymbol{\pi})$ *where* $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)^T$ *is the probability distribution over the* k *node classes and nodes connect with probability* $\mathbb{E}[\mathbf{A}]_{ij} := \frac{B_{c_i c_j}}{n}$ *encoded in the* $k \times k$ *block matrix* $\mathbf{B}$*. Let* $\boldsymbol{\Pi} := \operatorname{diag}(\boldsymbol{\pi})$ *and* $\mathbf{D} := \operatorname{diag}(\mathbf{B}\boldsymbol{\pi})$ *be diagonal matrices encoding the probability of class membership and mean class-wise degrees respectively. Then, assuming that the other conditions of Lemma* 2 *hold, the first and second order homophily (Eq.* (6)*) with the symmetric normalised adjacency matrix* $\hat{\mathbf{A}}_{\mathrm{sym}}$ *as the graph shift operator (Eq.* (2)*) can be tightly bounded in expectation by:*

$$\mathbb{E}\left[h^1\left(\hat{\mathbf{A}}_{\mathrm{sym}}\right)\right] \lessapprox \operatorname{Tr}\left(\mathbf{D}^{-1}\boldsymbol{\Pi}\mathbf{B}\boldsymbol{\Pi}\right),$$

$$\mathbb{E}\left[h^2\left(\hat{\mathbf{A}}_{\mathrm{sym}}\right)\right] \lessapprox \boldsymbol{\pi}^T\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}\boldsymbol{\pi} + \operatorname{Tr}\left(\mathbf{D}^{-1}\boldsymbol{\Pi}\mathbf{B}\left\{\mathbf{D}^{-1} - \mathbf{D}^{-2}\left(\mathbf{I}_k - e^{-\mathbf{D}}\right)\right\}\boldsymbol{\Pi}\mathbf{B}\boldsymbol{\Pi}\right),$$

where $\mathbf{I}_k$ *is the size-*k *identity matrix, and the bound gets tighter with larger class-wise mean degrees.*

Interpretation for Corollary 3.1. This corollary provides a fully analytic characterisation of first and second order homophily in terms of parameters of a graph model, which can be used with Eq. (7) to bound the mean signal sensitivity of a single MPNN layer in expectation. In particular, using an SBM to model the graph ensemble, one can analyse how class-wise connectivity affects the emergence of homophilic bottlenecks. We demonstrate the validity of these bounds and their tightness in Figure 1.

D Proofs

D.1 Main

In this section we restate the key results and provide their proofs.

Theorem 1 (Signal sensitivity bounds MPNN performance)**.** *Let $(G, \mathbf{X})$ and $(G, \widetilde{\mathbf{X}})$ be class-wise attributed graphs—with k classes following Eq.* (3)*—consisting of the same graph G with n nodes but different input features $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, corresponding to (potentially different) class-wise means $\{\boldsymbol{\mu}_u\}_{u=1}^k$, $\{\widetilde{\boldsymbol{\mu}}_u\}_{u=1}^k$ and node-wise covariances $\{\boldsymbol{\Sigma}_{ii}\}_{i=1}^n$, $\{\widetilde{\boldsymbol{\Sigma}}_{ii}\}_{i=1}^n$. Let the hidden feature representation for node i be considered as a differentiable function $H_i^{(\ell)} : \mathbb{R}^{n \times d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ of the input features matrix. Then the expected squared distance between the output embeddings of the ℓ^{th} MPNN layer, given the graph G, is bounded by:*

$$\mathbb{E}\left[\left\|H_i^{(\ell)}(\mathbf{X}) - H_i^{(\ell)}(\widetilde{\mathbf{X}})\right\|^2 \,\middle|\, G\right] \le \mathcal{S}_\mu^{(\ell)}(i) \sum_{u=1}^{k} \|\boldsymbol{\mu}_u - \widetilde{\boldsymbol{\mu}}_u\|^2 + \mathcal{S}_\epsilon^{(\ell)}(i) \sum_{j=1}^{n} \operatorname{Tr}\left(\boldsymbol{\Sigma}_{jj} + \widetilde{\boldsymbol{\Sigma}}_{jj}\right), \tag{13}$$

where $\mathcal{S}_\mu^{(\ell)}(i)$ and $\mathcal{S}_\epsilon^{(\ell)}(i)$ refer to the signal sensitivity (Definition 3*) and noise sensitivity (Definition* 4*) of the ℓ^{th} MPNN layer for node i, respectively.*

Proof. We begin by separating the variation in input features due to the class-wise means and residuals. Let $\mathbf{C}$ be an $n \times k$ assignment matrix such that $C_{iu} := 1$ if $c_i = u$ and $C_{iu} := 0$ otherwise. Let $\mathbf{M} := (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)^T$ and $\widetilde{\mathbf{M}} := (\widetilde{\boldsymbol{\mu}}_1, \ldots, \widetilde{\boldsymbol{\mu}}_k)^T$ be matrices of size $k \times d_{\text{in}}$ that collect the class-wise means and $\mathbf{E} := (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^T$ and $\widetilde{\mathbf{E}} := (\widetilde{\boldsymbol{\epsilon}}_1, \ldots, \widetilde{\boldsymbol{\epsilon}}_n)^T$ be matrices of size $n \times d_{\text{in}}$ that collect the residuals such that $\mathbf{X} = \mathbf{CM} + \mathbf{E}$ and $\widetilde{\mathbf{X}} = \mathbf{C}\widetilde{\mathbf{M}} + \widetilde{\mathbf{E}}$. Then we can write:

$$\begin{aligned} \left\|H_i^{(\ell)}(\mathbf{X}) - H_i^{(\ell)}(\widetilde{\mathbf{X}})\right\|^2 &= \left\|H_i^{(\ell)}(\mathbf{CM} + \mathbf{E}) - H_i^{(\ell)}(\mathbf{C}\widetilde{\mathbf{M}} + \widetilde{\mathbf{E}})\right\|^2 \\ \le \left\|H_i^{(\ell)}(\mathbf{CM} + \mathbf{E}) - H_i^{(\ell)}(\mathbf{C}\widetilde{\mathbf{M}} + \mathbf{E})\right\|^2 &+ \left\|H_i^{(\ell)}(\mathbf{C}\widetilde{\mathbf{M}} + \mathbf{E}) - H_i^{(\ell)}(\mathbf{C}\widetilde{\mathbf{M}} + \widetilde{\mathbf{E}})\right\|^2. \end{aligned} \tag{16}$$

Considering $H_i^{(\ell)}(\mathbf{CM} + \mathbf{E})$ first as a function of $\mathbf{M}$ and then as a function of $\mathbf{E}$ we can use the mean value theorem to bound both terms in Eq. (16) using the signal sensitivity and noise sensitivity of the GNN. In particular, consider the function $\hat{H}_i^{(\ell)} : \mathbb{R}^{k \times d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ given by $\hat{H}_i^{(\ell)}(\mathbf{M}) = H_i^{(\ell)}(\mathbf{CM} + \mathbf{E})$. Assuming continuous differentiability of $\hat{H}_i^{(\ell)}$ over $\mathbb{R}^{k \times d_{\text{in}}}$, we apply the mean value inequality for matrices from Proposition 2:

$$\left\|\hat{H}_i^{(\ell)}(\mathbf{M}) - \hat{H}_i^{(\ell)}(\widetilde{\mathbf{M}})\right\|^2 \le \left\|\mathbf{M} - \widetilde{\mathbf{M}}\right\|^2 \sup_{\mathbf{M} \in \mathbb{R}^{k \times d_{\text{in}}}} \left\|\nabla \hat{H}_i^{(\ell)}\left(\mathbf{M}\right)\right\|^2 \implies$$

$$\left\|H_i^{(\ell)}(\mathbf{CM} + \mathbf{E}) - H_i^{(\ell)}(\mathbf{C}\widetilde{\mathbf{M}} + \mathbf{E})\right\|^2 \le \left\|\mathbf{M} - \widetilde{\mathbf{M}}\right\|^2 \sup_{\mathbf{M} \in \mathbb{R}^{k \times d_{\text{in}}}} \sum_{q=1}^{d_{\text{out}}} \left\|\mathbf{C}^T \left[\nabla H_i^{(\ell)}\left(\mathbf{CM} + \mathbf{E}\right)\right]_q\right\|^2.$$

By the definition of $\mathbf{C}$ we have:

$$\left[\mathbf{C}^T \nabla H_i\left(\mathbf{CM} + \mathbf{E}\right)\right]_q]_{up} = \sum_{j=1}^{n} \frac{\partial H_{ip}}{\partial X_{jq}} \delta_{ju} = \left[\mathcal{J}_\mu^{(\ell)}\right]_{iupq}$$

$$\implies \sup_{\mathbf{M} \in \mathbb{R}^{k \times d_{\text{in}}}} \sum_{q=1}^{d_{\text{out}}} \left\|\mathbf{C}^T \left[\nabla H_i^{(\ell)}\left(\mathbf{CM} + \mathbf{E}\right)\right]_q\right\|^2 = \sup_{\mathbf{M} \in \mathbb{R}^{k \times d_{\text{in}}}} \left\|\left[\mathcal{J}_\mu^{(\ell)}\right]_i\right\|^2 = \mathcal{S}_\mu^{(\ell)}(i),$$

where we use the definition of the class-reduced Jacobian (Definition 2) and signal sensitivity (Definition 3), and Eq. (11).

Similarly, consider the function $\widetilde{H}_i^{(\ell)} : \mathbb{R}^{n \times d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ given by $\widetilde{H}_i^{(\ell)}(\mathbf{E}) = H_i^{(\ell)}(\mathbf{C}\widetilde{\mathbf{M}} + \mathbf{E})$, then by the mean value inequality for matrices in Proposition 2 we have:

$$\begin{aligned}
\left\| \widetilde{H}_i^{(\ell)}(\mathbf{E}) - \widetilde{H}_i^{(\ell)}(\widetilde{\mathbf{E}}) \right\|^2 &\leq \left\| \mathbf{E} - \widetilde{\mathbf{E}} \right\|^2 \sup_{\mathbf{E} \in \mathbb{R}^{n \times d_{\text{in}}}} \left\| \nabla \widetilde{H}_i^{(\ell)}\left(\mathbf{E}\right) \right\|^2 \\
\implies \left\| H_i^{(\ell)}(\mathbf{C}\widetilde{\mathbf{M}} + \mathbf{E}) - H_i^{(\ell)}(\mathbf{C}\widetilde{\mathbf{M}} + \widetilde{\mathbf{E}}) \right\|^2 &\leq \left\| \mathbf{E} - \widetilde{\mathbf{E}} \right\|^2 \sup_{\mathbf{E} \in \mathbb{R}^{n \times d_{\text{in}}}} \nabla H_i^{(\ell)}\left(\mathbf{C}\widetilde{\mathbf{M}} + \mathbf{E}\right) \\
&= \left\| \mathbf{E} - \widetilde{\mathbf{E}} \right\|^2 \mathcal{S}_\epsilon^{(\ell)}(i),
\end{aligned}$$

where we use the definition of noise sensitivity (Definition 4) and Eq. (10). Substituting both these bounds into Eq. (16) yields:

$$\left\| H_i^{(\ell)}(\mathbf{X}) - H_i^{(\ell)}(\widetilde{\mathbf{X}}) \right\|^2 \leq \mathcal{S}_\mu^{(\ell)}(i) \left\| \mathbf{M} - \widetilde{\mathbf{M}} \right\|^2 + \mathcal{S}_\epsilon^{(\ell)}(i) \left\| \mathbf{E} - \widetilde{\mathbf{E}} \right\|^2.$$

For a given graph G the terms $\mathcal{S}_\mu^{(\ell)}(i)$ and $\mathcal{S}_\epsilon^{(\ell)}(i)$ are deterministic, then taking the expectation conditioned on G gives:

$$\mathbb{E}\left[\left\| H_i^{(\ell)}(\mathbf{X}) - H_i^{(\ell)}(\widetilde{\mathbf{X}}) \right\|^2 \middle| G \right] \leq \mathcal{S}_\mu^{(\ell)}(i) \left\| \mathbf{M} - \widetilde{\mathbf{M}} \right\|^2 + \mathcal{S}_\epsilon^{(\ell)}(i) \mathbb{E}\left[\left\| \mathbf{E} - \widetilde{\mathbf{E}} \right\|^2 \right]. \tag{17}$$

Let $\boldsymbol{\sigma} \coloneqq \{\text{Tr}(\boldsymbol{\Sigma}_{ii})\}_{i=1}^n$ and $\widetilde{\boldsymbol{\sigma}} \coloneqq \{\text{Tr}(\widetilde{\boldsymbol{\Sigma}}_{ii})\}_{i=1}^n$ be length-n vectors encoding node-wise variances then:

$$\begin{aligned}
\mathbb{E}\left[\left\| \mathbf{E} - \widetilde{\mathbf{E}} \right\|^2 \right] &= \mathbb{E}\left[\text{Tr}\left((\mathbf{E} - \widetilde{\mathbf{E}})(\mathbf{E} - \widetilde{\mathbf{E}})^T \right) \right] = \left(\mathbb{E}\left[\text{Tr}(\mathbf{E}\mathbf{E}^T) \right] + \mathbb{E}\left[\text{Tr}(\widetilde{\mathbf{E}}\widetilde{\mathbf{E}})^T \right] \right) \\
&= \sum_{i=1}^n \mathbb{E}\left[\boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i \right] + \mathbb{E}\left[\widetilde{\boldsymbol{\epsilon}}_i^T \widetilde{\boldsymbol{\epsilon}}_i \right] = \sum_{i=1}^n \left(\sigma_i + \widetilde{\sigma}_i \right),
\end{aligned}$$

where we use Eq. (3). Substituting in Eq. (17) yields the RHS of Eq. (13). □

Lemma 1 (Bound for MPNN Jacobian). *Let $\mathcal{J}^{(\ell)}$ be the Jacobian of the ℓ^{th} layer of an MPNN (Definition 1) that uses the graph shift operator $\hat{\mathbf{A}}$ with message and update functions $\{\psi_k(\cdot,\cdot)\}_{k=1}^\ell$ and $\{\phi_k(\cdot,\cdot)\}_{k=1}^\ell$, as in Eq. (1). Let $\|\cdot\|$ be the Euclidean norm, and $\nabla_1 f$ and $\nabla_2 f$ be the Jacobian matrices of some function $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with respect to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. Assuming that there exist constants $\alpha_1, \alpha_2, \beta_1, \beta_2$ such that $\forall r \in [\ell]$ the message and update functions satisfy $\|\nabla_1 \phi_r\| \leq \alpha_1$, $\|\nabla_2 \phi_r\| \leq \alpha_2$, $\|\nabla_1 \psi_r\| \leq \beta_1$, and $\|\nabla_2 \psi_r\| \leq \beta_2$ then:*

$$\left\| \left[\mathcal{J}^{(\ell)} \right]_{ij} \right\| \leq \left[\left(\alpha_2 \beta_2 \hat{\mathbf{A}} + \alpha_2 \beta_1 \text{diag}\left(\hat{\mathbf{A}} \mathbf{1}_n \right) + \alpha_1 \mathbf{I}_n \right)^\ell \right]_{ij},$$

where $\mathbf{1}_n$ is the size-n vector of ones and $\mathbf{I}_n$ is the identity matrix of size n.

Proof. By applying the chain rule to Eq. (1) the Jacobian of the ℓ^{th} MPNN layer is given by:

$$\begin{aligned}
\left[\mathcal{J}^{(\ell)} \right]_{ij} &= \nabla_1 \phi_\ell \left[\nabla \mathbf{H}_i^{(\ell-1)} \right]_j + \nabla_2 \phi_\ell \sum_{l \in N(i)} \hat{A}_{il} \left(\nabla_1 \psi_\ell \left[\nabla \mathbf{H}_i^{(\ell-1)} \right]_j + \nabla_2 \psi_\ell \left[\nabla \mathbf{H}_l^{(\ell-1)} \right]_j \right), \\
&= \left(\nabla_1 \phi_\ell + \nabla_2 \phi_\ell \nabla_1 \psi_\ell \sum_{l \in N(i)} \hat{A}_{il} \right) \left[\mathcal{J}^{(\ell-1)} \right]_{ij} + \nabla_2 \phi_\ell \sum_{l \in N(i)} \hat{A}_{il} \nabla_2 \psi_\ell \left[\mathcal{J}^{(\ell-1)} \right]_{lj}.
\end{aligned}$$

By norm sub-additivity and sub-multiplicativity we have:

$$\begin{aligned}
\left\|\left[\mathcal{J}^{(\ell)}\right]_{ij}\right\| &\leq \left(\|\nabla_1\phi_\ell\| + \|\nabla_2\phi_\ell\|\,\|\nabla_1\psi_\ell\| \sum_{l\in N(i)} \hat{A}_{il}\right)\left\|\left[\mathcal{J}^{(\ell-1)}\right]_{ij}\right\| \\
&\qquad + \|\nabla_2\phi_\ell\|\,\|\nabla_2\psi_\ell\| \sum_{l\in N(i)} \hat{A}_{il}\left\|\left[\mathcal{J}^{(\ell-1)}\right]_{lj}\right\| \\
&\leq \left(\alpha_1 + \alpha_2\beta_1 \sum_{l\in N(i)} \hat{A}_{il}\right)\left\|\left[\mathcal{J}^{(\ell-1)}\right]_{ij}\right\| + \alpha_2\beta_2 \sum_{l\in N(i)} \hat{A}_{il}\left\|\left[\mathcal{J}^{(\ell-1)}\right]_{lj}\right\| \\
&= \sum_{l=1}^{n}\left[\alpha_2\beta_2\hat{\mathbf{A}} + \alpha_2\beta_1\mathrm{diag}\left(\hat{\mathbf{A}}\mathbf{1}_n\right) + \alpha_1\mathbf{I}_n\right]_{il}\left\|\left[\mathcal{J}^{(\ell-1)}\right]_{lj}\right\| \\
\implies \mathbf{J}^{(\ell)} &\leq \left(\alpha_2\beta_2\hat{\mathbf{A}} + \alpha_2\beta_1\mathrm{diag}\left(\hat{\mathbf{A}}\mathbf{1}_n\right) + \alpha_1\mathbf{I}_n\right)\mathbf{J}^{(\ell-1)},
\end{aligned}$$

where $J_{ij}^{(\ell)} := \left\|\left[\mathcal{J}^{(\ell)}\right]_{ij}\right\|$. Applying this bound recursively yields $\mathbf{J}^{(\ell)} \leq \left(\alpha_2\beta_2\hat{\mathbf{A}} + \alpha_2\beta_1\mathrm{diag}\left(\hat{\mathbf{A}}\mathbf{1}_n\right) + \alpha_1\mathbf{I}_n\right)^{\ell}$, where we use the initial condition $\left[\mathcal{J}^{(0)}\right]_{ijpq} := \frac{\partial X_{ip}}{\partial X_{jq}} = \delta_{ij}\delta_{pq} \implies \left\|\left[\mathcal{J}^{(0)}\right]_{ij}\right\| = \delta_{ij}$. □

Theorem 2 (Higher-order weighted homophily bounds signal sensitivity)**.** *Let $(G, \mathbf{X})$ be a class-wise attributed graph with k classes following Eq. (3) and $\bar{\mathcal{S}}_{\mu}^{(\ell)}$ be the mean signal sensitivity of the ℓ^{th} layer (Definition 3) of an MPNN that uses the graph shift operator $\hat{\mathbf{A}}$ with message and update functions $\{\psi_k(\cdot,\cdot)\}_{k=1}^{\ell}$ and $\{\phi_k(\cdot,\cdot)\}_{k=1}^{\ell}$, as in Eq. (1). Let $\|\cdot\|$ be the Euclidean norm, $\nabla_1 f$ and $\nabla_2 f$ be the Jacobian matrices of some function $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with respect to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. Assuming that there exist constants $\alpha_1, \alpha_2, \beta_1, \beta_2$ such that $\forall r \in [\ell]$ the message and update functions satisfy $\|\nabla_1\phi_r\| \leq \alpha_1$, $\|\nabla_2\phi_r\| \leq \alpha_2$, $\|\nabla_1\psi_r\| \leq \beta_1$, and $\|\nabla_2\psi_r\| \leq \beta_2$ then:*

$$\bar{\mathcal{S}}_{\mu}^{(\ell)} \leq \sum_{r,s=0}^{\ell}\sum_{u=0}^{r}\sum_{v=0}^{s}\binom{\ell}{r}\binom{\ell}{s}\binom{r}{u}\binom{s}{v}\alpha_1^{2\ell-r-s}\left(\alpha_2\beta_2\right)^{r+s} h_{u,v}^{r-u,s-v}\left(\hat{\mathbf{A}}, \beta_1\beta_2^{-1}\hat{\mathbf{A}}\mathbf{1}_n\right),$$

where $h_{u,v}^{r,s}(\cdot,\cdot)$ is the (r,s)-order (u,v)-weighted homophily given by Definition 5.

Proof. Applying the Cauchy–Schwarz inequality to the input to the $\sup$ on the RHS of Eq. (4) yields:

$$\left\|\left[\mathcal{J}_{\mu}^{(\ell)}\right]_{i}\right\|^2 = \sum_{j,l=1}^{n}\sum_{p=1}^{d_{\text{out}}}\sum_{q=1}^{d_{\text{in}}}\frac{\partial H_{ip}^{(\ell)}}{\partial X_{jq}}\frac{\partial H_{ip}^{(\ell)}}{\partial X_{lq}}\delta_{c_j c_l} \leq \sum_{j,l=1}^{n}\left\|\left[\mathcal{J}^{(\ell)}\right]_{ij}\right\|\left\|\left[\mathcal{J}^{(\ell)}\right]_{il}\right\|\delta_{c_j c_l}.$$

Since the conditions of Lemma 1 are satisfied, applying it above and summing over i yields:

$$\begin{aligned}
\sum_{i=1}^{n}\left\|\left[\mathcal{J}_{\mu}^{(\ell)}\right]_{i}\right\|^2 &\leq \sum_{i,j,l=1}^{n}\left[\left(\alpha_2\beta_2\hat{\mathbf{A}} + \alpha_2\beta_1\hat{\mathbf{D}} + \alpha_1\mathbf{I}_n\right)^{\ell}\right]_{ij}\left[\left(\alpha_2\beta_2\hat{\mathbf{A}} + \alpha_2\beta_1\hat{\mathbf{D}} + \alpha_1\mathbf{I}_n\right)^{\ell}\right]_{il}\delta_{c_j c_l} \\
&= \sum_{j,l=1}^{n}\left[\left(\alpha_2\beta_2\hat{\mathbf{A}}^T + \alpha_2\beta_1\hat{\mathbf{D}} + \alpha_1\mathbf{I}_n\right)^{\ell}\left(\alpha_2\beta_2\hat{\mathbf{A}} + \alpha_2\beta_1\hat{\mathbf{D}} + \alpha_1\mathbf{I}_n\right)^{\ell}\right]_{jl}\delta_{c_j c_l},
\end{aligned} \tag{18}$$

where we define $\hat{\mathbf{D}} := \mathrm{diag}\left(\hat{\mathbf{A}}\mathbf{1}_n\right)$. Consider the matrix series expansion:

$$\begin{aligned}
\left(\alpha_2\beta_2\hat{\mathbf{A}}^T + \alpha_2\beta_1\hat{\mathbf{D}} + \alpha_1\mathbf{I}_n\right)^{\ell} &= \sum_{r=0}^{\ell}\binom{\ell}{r}\alpha_1^{\ell-r}\left(\alpha_2\beta_2\right)^{\ell}\left(\hat{\mathbf{A}}^T + \beta_1\beta_2^{-1}\mathbf{D}\right)^{r} \\
&= \sum_{r=0}^{\ell}\sum_{u=0}^{r}\binom{\ell}{r}\binom{r}{u}\alpha_1^{\ell-r}\left(\alpha_2\beta_2\right)^{r}\mathfrak{H}_u^{r-u}\left(\hat{\mathbf{A}}^T, \beta_1\beta_2^{-1}\hat{\mathbf{A}}\mathbf{1}_n\right),
\end{aligned}$$

where in the first equality we use the binomial expansion and in the second equality we use the definition in Eq. (12). A similar expansion can be derived using $\hat{\mathbf{A}}$ which, upon substitution in Eq. (18) and dividing by n yields:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left\|\left[\mathcal{J}_{\mu}^{(\ell)}\right]_{i}\right\|^{2} \leq \frac{1}{n}\sum_{j,l=1}^{n}\sum_{r,s=0}^{\ell}\sum_{u=0}^{r}\sum_{v=0}^{s}\binom{\ell}{r}\binom{\ell}{s}\binom{r}{u}\binom{s}{v}\alpha_1^{2\ell-r-s}\left(\alpha_2\beta_2\right)^{r+s} & \\ \times\left[\mathfrak{H}_{u}^{r-u}\left(\hat{\mathbf{A}}^{T},\beta_1\beta_2^{-1}\hat{\mathbf{A}}\mathbf{1}_n\right)\mathfrak{H}_{v}^{s-v}\left(\hat{\mathbf{A}},\beta_1\beta_2^{-1}\hat{\mathbf{A}}\mathbf{1}_n\right)\right]_{jl}\delta_{c_jc_l} & \\ = \sum_{r,s=0}^{\ell}\sum_{u=0}^{r}\sum_{v=0}^{s}\binom{\ell}{r}\binom{\ell}{s}\binom{r}{u}\binom{s}{v}\alpha_1^{2\ell-r-s}\left(\alpha_2\beta_2\right)^{r+s}h_{u,v}^{r-u,s-v}\left(\hat{\mathbf{A}},\beta_1\beta_2^{-1}\hat{\mathbf{A}}\mathbf{1}_n\right), &\end{aligned}$$

where we use Definition 5 for higher-order weighted homophily. Note that the RHS does not depend on the features or their means. Therefore, taking the supremum of the LHS over the space of mean matrices gives, alongside the definition of mean signal sensitivity (Definition 3), produces the desired result. □

Corollary 2.1 (r-order homophily bounds signal sensitivity in isotropic MPNNs with a symmetric graph shift operator)**.** *Let* $(G, \mathbf{X})$ *be a class-wise attributed graph with* k *classes following Eq.* (3) *and* $\bar{\mathcal{S}}_{\mu}^{(\ell)}$ *be the mean signal sensitivity of the* ℓ^{th} *layer (Definition 3) of an MPNN that uses a symmetric graph shift operator* $\hat{\mathbf{A}}$ *with message and update functions* $\{\psi_k(\cdot,\cdot)\}_{k=1}^{\ell}$ *and* $\{\phi_k(\cdot,\cdot)\}_{k=1}^{\ell}$*, as in Eq.* (1)*, that satisfy the conditions in Theorem 2 with the additional constraint that* $\beta_1 = 0$ *(for isotropic MPNNs), then:*

$$\bar{\mathcal{S}}_{\mu}^{(\ell)} \leq \sum_{r=0}^{2\ell}\binom{2\ell}{r}\alpha_1^{2\ell-r}\left(\alpha_2\beta_2\right)^{r}h^{r}\left(\hat{\mathbf{A}}\right),$$

where $h^{r}(\cdot)$ *is the* r*-order homophily defined in Eq.* (6)*.*

Proof. The proof follows from the result in Theorem 2 by setting $\beta_1 = 0$, using the symmetry of $\hat{\mathbf{A}}$, i.e. $\hat{\mathbf{A}} = \hat{\mathbf{A}}^T$, and applying Vandermonde's identity in the form $\sum_{r+s=t}\binom{\ell}{r}\binom{\ell}{s} = \binom{2\ell}{t}$. □

Corollary 2.2 (Weighted sum of closed walks bounds noise sensitivity in isotropic MPNNs with a symmetric graph shift operator)**.** *Let* $(G, \mathbf{X})$ *be a class-wise attributed graph with* k *classes following Eq.* (3) *and* $\bar{\mathcal{S}}_{\epsilon}^{(\ell)}$ *be the mean noise sensitivity of the* ℓ^{th} *layer (Definition 4) of an MPNN that uses a symmetric graph shift operator* $\hat{\mathbf{A}}$ *with message and update functions* $\{\psi_k(\cdot,\cdot)\}_{k=1}^{\ell}$ *and* $\{\phi_k(\cdot,\cdot)\}_{k=1}^{\ell}$*, as in Eq.* (1)*, that satisfy the conditions in Theorem 2 with the additional constraint that* $\beta_1 = 0$ *(for isotropic MPNNs), then:*

$$\bar{\mathcal{S}}_{\epsilon}^{(\ell)} \leq \frac{1}{n}\sum_{r=0}^{2\ell}\binom{2\ell}{r}\alpha_1^{2\ell-r}\left(\alpha_2\beta_2\right)^{r}\operatorname{Tr}\left(\hat{\mathbf{A}}^{r}\right), \tag{14}$$

where $\operatorname{Tr}(\cdot)$ *is the matrix trace.*

Proof. Using Definition 4, one has the following:

$$\mathcal{S}_{\epsilon}^{(\ell)}(i) = \sup_{\mathbf{X}\in\mathbb{R}^{n\times d_{\text{in}}}}\sum_{j=1}^{n}\sum_{p=1}^{d_{\text{out}}}\sum_{q=1}^{d_{\text{in}}}\left(\frac{\partial H_{ip}^{(\ell)}}{\partial X_{jq}}\right)^{2} = \sup_{\mathbf{X}\in\mathbb{R}^{n\times d_{\text{in}}}}\sum_{j,l=1}^{n}\sum_{p=1}^{d_{\text{out}}}\sum_{q=1}^{d_{\text{in}}}\frac{\partial H_{ip}^{(\ell)}}{\partial X_{jq}}\frac{\partial H_{ip}^{(\ell)}}{\partial X_{lq}}\delta_{jl}. \tag{19}$$

By noting the similarity to Eq. (4), the proof follows similarly to that of Theorem 2, except by replacing every instance of $\delta_{c_ic_j}$ with δ_{ij}. □

Lemma 2 (Underreaching in MPNNs for sparse graph ensembles; Loomba and Jones [14])**.** *For an undirected and simple graph* G *with* n *nodes encoded by the adjacency matrix* $\mathbf{A}$*, sampled from a general random graph family with conditionally independent edges and expected adjacency matrix* $\mathbb{E}[\mathbf{A}]$*, if the network is sparse in the sense that* $\mathbb{E}[\mathbf{A}] = O\left(n^{-1}\right)$*, each node is on the giant* *component with probability* $1 - o(1)$*, then asymptotically the cumulative distribution function of the length of the shortest path* λ_{ij} *between nodes* i *and* $j \neq i$ *is given by:*

$$\mathbb{P}\left(\lambda_{ij} \leq r\right) \approx \left[\sum_{s=1}^{r} \mathbb{E}\left[\mathbf{A}\right]^{s}\right]_{ij}$$

where "$\approx$*" indicates an asymptotic first-order approximation as* $n \to \infty$*.*

Proof. The proof follows by considering the first-order asymptotic approximation of Eq. (30) in Loomba and Jones [14] and computing the matrix series sum by assuming $\mathbb{E}\left[\mathbf{A}\right] - \mathbf{I}_n$ is invertible. □

Theorem 3 (Boundary oversquashing in MPNNs for sparse graph ensembles)**.** *Assume the same conditions as in Lemma 2, and additionally assume large expected node degrees encoded in the diagonal matrix* $\langle \mathbf{D} \rangle := \operatorname{diag}\left(\mathbb{E}\left[\mathbf{A}\right] \mathbf{1}_n\right)$ *where* $\mathbf{1}_n$ *is the length-*n *vector of ones. Then for the symmetric normalised adjacency matrix* $\hat{\mathbf{A}}_{\text{sym}}$*(Eq. (2)) the boundary oversquashing between nodes* i *and* $j \neq i$*, where* λ_{ij} *is the shortest path distance from* i *to* j*, is asymptotically bounded by:*

$$\mathbb{E}\left[\left[\hat{\mathbf{A}}^{r}_{\text{sym}}\right]_{ij} \middle| \lambda_{ij} = r\right] \lessapprox \frac{\left[\langle \mathbf{D} \rangle^{-\frac{1}{2}} \mathbb{E}\left[\mathbf{A}\right] \left(\left\{\langle \mathbf{D} \rangle^{-1} - \langle \mathbf{D} \rangle^{-2} \left(\mathbf{I}_n - e^{-\langle \mathbf{D} \rangle}\right)\right\} \mathbb{E}\left[\mathbf{A}\right]\right)^{r-1} \langle \mathbf{D} \rangle^{-\frac{1}{2}}\right]_{ij}}{\left[\mathbb{E}\left[\mathbf{A}\right]^{r}\right]_{ij}}, \tag{15}$$

where $\mathbf{I}_n$ *is the size-*n *identity matrix, and the bound gets tighter with larger mean degrees.*

Proof. Using Eq. (2), the LHS of Eq. (15) we can be written as:

$$\begin{aligned}
\mathbb{E}\left[\left[\hat{\mathbf{A}}^{r}_{\text{sym}}\right]_{ij} \middle| \lambda_{ij} = r\right] &= \mathbb{E}\left[\frac{1}{\sqrt{D_{ii} D_{jj}}} \sum_{k_1, k_2, \ldots, k_{r-1}=1}^{n} \frac{A_{ik_1} A_{k_1 k_2} \ldots A_{k_{r-1} j}}{D_{k_1 k_1} D_{k_2 k_2} \ldots D_{k_{r-1} k_{r-1}}} \middle| \lambda_{ij} = r\right] \\
&= \sum_{\substack{k_1, k_2, \ldots, k_{r-1}=1 \\ i \neq k_1 \neq k_2 \ldots \neq k_{r-1} \neq j}}^{n} \mathbb{E}\left[\frac{1}{\sqrt{D_{ii} D_{jj}}} \frac{A_{ik_1} A_{k_1 k_2} \ldots A_{k_{r-1} j}}{D_{k_1 k_1} D_{k_2 k_2} \ldots D_{k_{r-1} k_{r-1}}} \middle| \lambda_{ij} = r\right],
\end{aligned} \tag{20}$$

where we use the linearity of expectation and the fact that if the shortest path distance from i to j is r then a walk of length r from i to j via nodes $k_1, k_2, \ldots k_{r-1}$ must be a path, i.e. $i \neq k_1 \neq k_2 \ldots \neq k_{r-1} \neq j$. For brevity we will define $k_0 := i, k_r := j$ and refer to the sequence $\{k_l\}_{l=0}^{r}$ as the length-r path of interest. Given the definition of the adjacency matrix, we can write the conditional expectation on the RHS of Eq. (20) as:

$$\mathbb{E}\left[\left(\sqrt{D_{ii} D_{jj}} \prod_{l=1}^{r-1} D_{ll}\right)^{-1} \middle| \prod_{l=0}^{r-1} A_{k_l k_{l+1}} = 1, \lambda_{ij} = r\right] \mathbb{P}\left(\prod_{l=0}^{r-1} A_{k_l k_{l+1}} = 1 \middle| \lambda_{ij} = r\right). \tag{21}$$

Consider the first factor in Eq. (21). Knowing that $\prod_{l=0}^{r-1} A_{k_l k_{l+1}} = 1$ tell us that there *must exist* edges between nodes k_l and k_{l+1}. Knowing further that $\lambda_{ij} = r$ tell us that the path $\{k_l\}_{l=0}^{r}$ is a shortest path, i.e. there *cannot exist* paths shorter than length m between nodes k_l and k_{l+m}. Asymptotically, the probability of paths shorter than length m (for any finite m) *not existing* between any two nodes in a sparse graph is already $1 - o(1)$ (see Lemma 2 or Loomba and Jones [14]), i.e. the latter asymptotically does not inform the expectation of our quantity of interest. Furthermore, since edges are added between every node pair (conditionally) independently they affect—and can *only* affect—the degree of the nodes to which the edges are attached. This, alongside the fact that every node in the path $\{k_l\}_{l=0}^{r}$ is unique, permits us to asymptotically approximate the first factor in Eq. (21) as:

$$\mathbb{E}\left[D_{ii}^{-\frac{1}{2}} \middle| A_{ik_1}\right] \mathbb{E}\left[D_{jj}^{-\frac{1}{2}} \middle| A_{k_{r-1} j}\right] \prod_{l=1}^{r-1} \mathbb{E}\left[D_{k_l k_l}^{-1} \middle| A_{k_{l-1} k_l} A_{k_l k_{l+1}}\right].$$

Asymptotically, ignoring a single or two nodes has a vanishing effect on the degree of another node in a sparse graph with (conditionally) independent edges. In other words, knowing about the existence of a single or two edges attached to a given node merely shifts its degree distribution by one or two, respectively:

$$\begin{aligned}\mathbb{E}\left[D_{ii}^{-\frac{1}{2}}\,\middle|\,A_{ik_1}\right] &\approx \mathbb{E}\left[(D_{ii}+1)^{-\frac{1}{2}}\right],\\ \mathbb{E}\left[D_{jj}^{-\frac{1}{2}}\,\middle|\,A_{k_{l-1}j}\right] &\approx \mathbb{E}\left[(D_{jj}+1)^{-\frac{1}{2}}\right],\\ \mathbb{E}\left[D_{k_lk_l}^{-1}\,\middle|\,A_{k_{l-1}k_l}A_{k_lk_{l+1}}\right] &\approx \mathbb{E}\left[(D_{k_lk_l}+2)^{-1}\right].\end{aligned}$$

Asymptotically, the degree of a given node in a sparse graph with (conditionally) independent edges is Poisson distributed whose rate is given by its mean degree [14]. This allows us to apply the results in Eqs. (30b) and (30c) in Proposition 3 to write the first factor of Eq. (21) as:

$$\begin{aligned}\mathbb{E}&\left[\left(\sqrt{D_{ii}D_{jj}}\prod_{l=1}^{r-1}D_{ll}\right)^{-1}\,\middle|\,\prod_{l=0}^{r-1}A_{k_lk_{l+1}}=1,\lambda_{ij}=r\right] \lessapprox \left(\langle D_{ii}\rangle\langle D_{jj}\rangle\right)^{-\frac{1}{2}}\\ &\times\prod_{l=1}^{r-1}\left(\langle D_{k_lk_l}\rangle^{-1}-\langle D_{k_lk_l}\rangle^{-2}\left(1-e^{-\langle D_{k_lk_l}\rangle}\right)\right),\end{aligned} \tag{22}$$

and the bound is tight for large node mean degrees. Consider the second factor in Eq. (21) that can be rewritten as:

$$\mathbb{P}\left(\lambda_{ij}=r\,\middle|\,\prod_{l=0}^{r-1}A_{k_lk_{l+1}}=1\right)\frac{\mathbb{P}\left(\prod_{l=0}^{r-1}A_{k_lk_{l+1}}=1\right)}{\mathbb{P}\left(\lambda_{ij}=r\right)}.$$

Knowing that $\prod_{l=0}^{r-1}A_{k_lk_{l+1}}=1$, i.e. there exists a path of length r between i and j, tell us that the shortest path between i and j cannot be longer than r. Asymptotically, it tells us nothing about whether there exists a path shorter than length r between them. Since, *a priori*, the probability of the shortest path being less than length r is asymptotically vanishing (see Lemma 2 or Loomba and Jones [14]), this implies that $\mathbb{P}\left(\lambda_{ij}=r\,\middle|\,\prod_{l=0}^{r-1}A_{k_lk_{l+1}}=1\right)=1-o(1)$. Finally, due to conditional independence of edges, and considering the first-order approximation of Eq. (30) in Loomba and Jones [14], allows us to write the second factor of Eq. (21) as:

$$\mathbb{P}\left(\prod_{l=0}^{r-1}A_{k_lk_{l+1}}=1\,\middle|\,\lambda_{ij}=r\right)\approx\frac{\prod_{l=0}^{r-1}\mathbb{E}\left[A_{k_lk_{l+1}}\right]}{\left[\mathbb{E}\left[\mathbf{A}\right]^r\right]_{ij}}. \tag{23}$$

Putting Eqs. (22) and (23) in Eq. (20) yields:

$$\mathbb{E}\left[\left[\hat{\mathbf{A}}_{\mathrm{sym}}^r\right]_{ij}\,\middle|\,\lambda_{ij}=r\right]\lessapprox\frac{\left(\langle D_{ii}\rangle\langle D_{jj}\rangle\right)^{-\frac{1}{2}}}{\left[\mathbb{E}\left[\mathbf{A}\right]^r\right]_{ij}}\sum_{\substack{k_1,k_2,\dots,k_{r-1}=1\\ i\neq k_1\neq k_2\dots\neq k_{r-1}\neq j}}^{n}S\left(i,j,\{k_l\}_{l=1}^{r-1}\right),\ \text{where} \tag{24a}$$

$$S\left(i,j,\{k_l\}_{l=1}^{r-1}\right):=\mathbb{E}\left[A_{ik_1}\right]\prod_{l=1}^{r-1}\left(\langle D_{k_lk_l}\rangle^{-1}-\langle D_{k_lk_l}\rangle^{-2}\left(1-e^{-\langle D_{k_lk_l}\rangle}\right)\right)\mathbb{E}\left[A_{k_lk_{l+1}}\right]. \tag{24b}$$

Consider the term on the RHS of Eq. (24b). Due to the sparsity assumption $\mathbb{E}[\mathbf{A}]=O\left(n^{-1}\right)$ we have $S\left(i,j,\{k_l\}_{l=1}^{r-1}\right)=O\left(n^{-r}\right)$. We separately consider what happens when $S(i,j,\{k_l\}_{l=1}^{r}-1)$ is summed over different kinds of index combinations $\{k_l\}_{l=1}^{r-1}$.

First, consider unique index combinations $\{k_l\}_{l=1}^{r-1}$ of size $r-1$ from $[n]\setminus\{i,j\}$, as in the RHS of Eq. (24a) since $\{k_l\}_{l=0}^{r}$ encodes a shortest path. There are $\frac{(n-2)!}{(n-r-1)!}=O\left(n^{r-1}\right)$ such index combinations which yields a total contribution of order $O\left(n^{-1}\right)$ to the RHS of Eq. (24a).

Next, consider unique index combinations $\{k_l\}_{l=1}^{r-1}$ of size $r-1$ from $[n]$, such that exactly one of the $r-1$ indices is either i or j, which *do not* contribute to the RHS of Eq. (24a). There are $2(r-1)\frac{(n-2)!}{(n-r)!}=O\left(n^{r-2}\right)$ such index combinations which yields a total contribution of $O\left(n^{-2}\right)$.

Now, consider unique index combinations $\{k_l\}_{l=1}^{r-1}$ of size $r-1$ from $[n]$, such that exactly one of the $r-1$ indices is i and exactly another one is j, which *do not* contribute to the RHS of Eq. (24a). There are $(r-1)(r-2)\frac{(n-2)!}{(n-r+1)!} = O\left(n^{r-3}\right)$ such index combinations which yields a total contribution of $O\left(n^{-3}\right)$.

Finally, consider non-unique index combinations $\{k_l\}_{l=1}^{r-1}$ of size $r-1$ from n, such that there are $1 \leq m < r-1$ unique indices in the sequence $\{k_l\}_{l=1}^{r-1}$ repeated $t_1, t_2, \ldots, t_m$ number of times such that $\forall l \in [m] : t_l \geq 1$ and $\sum_{l=1}^{m} t_l = r-1$, which *do not* contribute to the RHS of Eq. (24a). There can be $\frac{(r-1)!}{t_1!t_2!\ldots t_m!}\frac{n!}{(n-m)!} = O\left(n^m\right)$ such index combinations which yields a total contribution of $O\left(n^{-r+m}\right)$. Since $1 \leq m < r-1$, considering a sum over all possible values of m yields a total contribution of all non-unique index combinations as $O\left(n^{-2}\right)$.

This exhausts all possible index combinations, which leads us to conclude that asymptotically only the unique index combinations contribute relatively non-vanishingly. In other words, replacing the sum over *unique* index combinations by a sum over *all* index combinations makes a vanishing difference to the RHS of Eq. (24a), allowing us to rewrite it as a product of matrices which yields the RHS of Eq. (20). □

Corollary 3.1 (Bound for first and second order homophily in sparse SBMs)**.** *Consider an undirected and simple graph G with n nodes encoded by the adjacency matrix $\mathbf{A}$ sampled from a sparse stochastic block model (SBM) such that node classes are i.i.d. as per $c \sim \text{Categorical}\left(\boldsymbol{\pi}\right)$ where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)^T$ is the probability distribution over the k node classes and nodes connect with probability $\mathbb{E}\left[\mathbf{A}\right]_{ij} := \frac{B_{c_i c_j}}{n}$ encoded in the $k \times k$ block matrix $\mathbf{B}$. Let $\boldsymbol{\Pi} := \operatorname{diag}\left(\boldsymbol{\pi}\right)$ and $\mathbf{D} := \operatorname{diag}\left(\mathbf{B}\boldsymbol{\pi}\right)$ be diagonal matrices encoding the probability of class membership and mean class-wise degrees respectively. Then, assuming that the other conditions of Lemma 2 hold, the first and second order homophily (Eq. (6)) with the symmetric normalised adjacency matrix $\hat{\mathbf{A}}_{\text{sym}}$ as the graph shift operator (Eq. (2)) can be tightly bounded in expectation by:*

$$\begin{aligned}
&\mathbb{E}\left[h^1\left(\hat{\mathbf{A}}_{\text{sym}}\right)\right] \lessapprox \operatorname{Tr}\left(\mathbf{D}^{-1}\boldsymbol{\Pi}\mathbf{B}\boldsymbol{\Pi}\right), \\
&\mathbb{E}\left[h^2\left(\hat{\mathbf{A}}_{\text{sym}}\right)\right] \lessapprox \boldsymbol{\pi}^T\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}\boldsymbol{\pi} + \operatorname{Tr}\left(\mathbf{D}^{-1}\boldsymbol{\Pi}\mathbf{B}\left\{\mathbf{D}^{-1} - \mathbf{D}^{-2}\left(\mathbf{I}_k - e^{-\mathbf{D}}\right)\right\}\boldsymbol{\Pi}\mathbf{B}\boldsymbol{\Pi}\right),
\end{aligned}$$

where $\mathbf{I}_k$ is the size-k identity matrix, and the bound gets tighter with larger class-wise mean degrees.

Proof. For brevity throughout the proof, we drop the subscript sym and use $\hat{\mathbf{A}}$ to refer to $\hat{\mathbf{A}}_{\text{sym}}$. Given the block membership c_i, c_j of nodes $i \neq j$, we have $\left[\mathbb{E}\left[\mathbf{A}\right]^r\right]_{ij} = \left[\mathbf{B}(\boldsymbol{\Pi}\mathbf{B})^{r-1}\right]_{ij}/n$. First, consider Eq. (8) with $r = 1$, i.e. $\mathbb{E}\left[\hat{\mathbf{A}}\right]$ which is given by:

$$\mathbb{E}\left[\hat{A}_{ij}\right] = \mathbb{E}\left[\hat{A}_{ij} \,\middle|\, \lambda_{ij} = 0\right]\mathbb{P}\left(\lambda_{ij} = 0\right) + \mathbb{E}\left[\hat{A}_{ij} \,\middle|\, \lambda_{ij} = 1\right]\mathbb{P}\left(\lambda_{ij} = 1\right) \lessapprox n^{-1}D_{c_i c_i}^{-\frac{1}{2}}B_{c_i c_j}D_{c_j c_j}^{-\frac{1}{2}},$$

where (a) for $\lambda_{ij} = 0 \implies i = j$ we use the fact that there are *no* self-loops i.e. $A_{ij} = 0 \implies \hat{A}_{ij} = 0$, and (b) for $\lambda_{ij} = 1 \implies i \neq j$ we use Lemma 2 and Theorem 3 with $r = 1$, and the bound gets tighter for larger class-wise mean degrees. Substituting in Eq. (7) yields the desired expression for $h^1\left(\hat{\mathbf{A}}_{\text{sym}}\right)$.

Next, consider Eq. (8) with $r = 2$, i.e. $\mathbb{E}\left[\hat{\mathbf{A}}^2\right]$ which is given by:

$$\mathbb{E}\left[\left[\hat{\mathbf{A}}^2\right]_{ij}\right] = \sum_{s=0}^{2}\mathbb{E}\left[\left[\hat{\mathbf{A}}^2\right]_{ij} \,\middle|\, \lambda_{ij} = s\right]\mathbb{P}\left(\lambda_{ij} = s\right). \tag{25}$$

For $\lambda_{ij} = 0 \implies i = j$, using d_i to denote the degree of node i, we get

$$\begin{aligned}
\mathbb{E}\left[\left[\hat{\mathbf{A}}^2\right]_{ii}\right] &= \mathbb{E}\left[\sum_j (d_i d_j)^{-1} A_{ij}\right] = \sum_j \mathbb{E}\left[(d_i d_j)^{-1} A_{ij}\right] \\
&= \sum_j \mathbb{E}\left[(d_i d_j)^{-1} \,\middle|\, A_{ij} = 1\right] \mathbb{P}\left(A_{ij} = 1\right) \\
&\approx \sum_j \mathbb{E}\left[(d_i + 1)^{-1}\right] \mathbb{E}\left[(d_j + 1)^{-1}\right] \mathbb{P}\left(A_{ij} = 1\right) \lessapprox \sum_j \mathbb{E}\left[d_i\right]^{-1} \mathbb{E}\left[d_j\right]^{-1} \mathbb{P}\left(A_{ij} = 1\right) \\
&= D^{-1}_{c_i c_i} [\mathbf{B}]_{c_i,:} \mathbf{D}^{-1} \boldsymbol{\pi},
\end{aligned} \tag{26}$$

where the second equality makes use of the linearity of expectation, the asymptotic approximation is due to an identical argument as in the proof for Theorem 3 for sparse networks, the bound is due to Eq. (30a) in Proposition 3 which becomes tighter for larger class-wise mean degrees, and $[\mathbf{X}]_{u,:}$ indicates the u^{th} row-vector of a matrix $\mathbf{X}$. For $\lambda_{ij} = 1 \implies i \neq j$ we get:

$$\begin{aligned}
\mathbb{E}\left[\left[\hat{\mathbf{A}}^2\right]_{ij} \,\middle|\, \lambda_{ij} = 1\right] = \mathbb{E}\left[\left[\hat{\mathbf{A}}^2\right]_{ij} \,\middle|\, A_{ij} = 1\right] &= \mathbb{E}\left[\sum_l (d_i d_j)^{-\frac{1}{2}} d_l^{-1} A_{il} A_{lj} \,\middle|\, A_{ij} = 1\right] \\
&= \sum_l \mathbb{E}\left[(d_i d_j)^{-\frac{1}{2}} d_l^{-1} \,\middle|\, A_{il} A_{lj} A_{ij} = 1\right] \mathbb{P}\left(A_{il} = 1, A_{lj} = 1 \,|\, A_{ij} = 1\right) \\
&= \sum_l \mathbb{E}\left[(d_i d_j)^{-\frac{1}{2}} d_l^{-1} \,\middle|\, A_{il} A_{lj} A_{ij} = 1\right] \mathbb{P}\left(A_{il} = 1\right) \mathbb{P}\left(A_{lj} = 1\right),
\end{aligned} \tag{27}$$

where the third equality makes use of the linearity of expectation, and the fourth equality uses the assumption of conditionally independent edges. We emphasise that, due to sparsity, the RHS of Eq. (27) is of the order $O\left(n^{-1}\right)$. For $\lambda_{ij} = 2 \implies i \neq j$ we get, using Eq. (15) from Theorem 3:

$$\mathbb{E}\left[\left[\hat{\mathbf{A}}^2\right]_{ij} \,\middle|\, \lambda_{ij} = 2\right] \lessapprox \frac{\left(D_{c_i c_i} D_{c_j c_j}\right)^{-\frac{1}{2}} \left[\mathbf{B}\left\{\mathbf{D}^{-1} - \mathbf{D}^{-2}\left(\mathbf{I}_k - e^{-\mathbf{D}}\right)\right\} \mathbf{\Pi} \mathbf{B}\right]_{c_i c_j}}{[\mathbf{B}\mathbf{\Pi}\mathbf{B}]_{c_i c_j}}, \tag{28}$$

and the bound gets tighter for larger degrees. The RHS of Eq. (28) is of the order $\Omega\left(1\right)$. That is, asymptotically, Eq. (27) contributes vanishingly to Eq. (25) when compared to Eq. (28). It then follows from Eqs. (25), (26), and (28) that asymptotically:

$$\mathbb{E}\left[\left[\hat{\mathbf{A}}^2\right]_{ij}\right] \lessapprox D^{-1}_{c_i c_i} [\mathbf{B}]_{c_i,:} \mathbf{D}^{-1} \boldsymbol{\pi} \delta_{ij} + \frac{\left(D_{c_i c_i} D_{c_j c_j}\right)^{-\frac{1}{2}} \left[\mathbf{B}\left\{\mathbf{D}^{-1} - \mathbf{D}^{-2}\left(\mathbf{I}_k - e^{-\mathbf{D}}\right)\right\} \mathbf{\Pi} \mathbf{B}\right]_{c_i c_j}}{n} (1 - \delta_{ij}),$$

which is a tighter bound for larger class-wise mean degrees. Substituting in Eq. (6) yields the desired expression for $h^2\left(\hat{\mathbf{A}}_{\text{sym}}\right)$. □

D.2 Supplementary

In this section we state some technical results and provide their proofs.

Proposition 1 (Mean value inequality; Rudin [21])**.** *Let $f : [a, b] \to \mathbb{R}^N$ be a continuous vector-valued function that is differentiable on $(a, b) \subset \mathbb{R}$ then $\exists c \in (a, b)$ such that:*

$$\left\|f(b) - f(a)\right\| \leq (b - a) \left\|f'(c)\right\|.$$

Proposition 2 (Mean value inequality for matrices)**.** *Let $f : Z \to \mathbb{R}^N$ be a continuous vector-valued function that is differentiable on a convex subset $Z \subset \mathbb{R}^{M \times L}$ then for $\mathbf{A} \in Z, \mathbf{B} \in Z$:*

$$\left\|f(\mathbf{B}) - f(\mathbf{A})\right\| \leq \left\|\mathbf{B} - \mathbf{A}\right\| \sup_{\mathbf{C} \in Z} \left\|\nabla f(\mathbf{C})\right\|. \tag{29}$$

Proof. Define $g : [0,1] \to \mathbb{R}^n$ as $g(x) = f(x\mathbf{A} + (1-x)\mathbf{B})$ which will be continuous on $[0,1]$ and differentiable on $(0,1)$ due to the continuous differentiability of f at every point $x\mathbf{A} + (1-x)\mathbf{B} \in Z$. We note that $f'(x) = \left\{\mathrm{Tr}\left([\nabla f(x\mathbf{A} + (1-x)\mathbf{B})]_i(\mathbf{B} - \mathbf{A})^T\right)\right\}_{i=1}^N$ where $\nabla f(\cdot)$ is the $N \times M \times L$ Jacobian tensor of f. Then applying the mean value inequality from Proposition 1 to g we obtain that $\exists c \in (0,1)$ such that

$$\|g(1) - g(0)\| \leq \|g'(c)\| \implies \|f(\mathbf{B}) - f(\mathbf{A})\| \leq \left\|\left\{\mathrm{Tr}\left([\nabla f(c\mathbf{A} + (1-c)\mathbf{B})]_i(\mathbf{B} - \mathbf{A})^T\right)\right\}_{i=1}^N\right\|.$$

By the Cauchy–Schwarz inequality $\left[\mathrm{Tr}\left([\nabla f(c\mathbf{A} + (1-c)\mathbf{B})]_i(\mathbf{B} - \mathbf{A})^T\right)\right]^2 \leq \|[\nabla f(c\mathbf{A} + (1-c)\mathbf{B})]_i\|^2 \|\mathbf{B} - \mathbf{A}\|^2$ which when substituted above yields

$$\|f(\mathbf{B}) - f(\mathbf{A})\| \leq \|\mathbf{B} - \mathbf{A}\| \sqrt{\sum_{i=1}^N \|[\nabla f(c\mathbf{A} + (1-c)\mathbf{B})]_i\|^2} = \|\mathbf{B} - \mathbf{A}\| \|\nabla f(c\mathbf{A} + (1-c)\mathbf{B})\|.$$

Since $c\mathbf{A} + (1-c)\mathbf{B} \in Z$ taking a supremum over Z gives us the RHS of Eq. (29). □

Proposition 3 (Expectation of transformation of Poisson distributed random variable)**.** *Let* $X \sim \mathrm{Poisson}(\lambda)$ *be a Poisson distributed random variable with rate parameter* $\lambda > 0$*, then:*

$$\mathbb{E}\left[\frac{1}{X+1}\right] = \frac{1 - e^{-\lambda}}{\lambda}, \tag{30a}$$

$$\mathbb{E}\left[\frac{1}{X+2}\right] = \frac{\lambda - 1 + e^{-\lambda}}{\lambda^2}, \tag{30b}$$

$$\sqrt{\frac{1}{\lambda} - \frac{1}{2\lambda^2}} < \mathbb{E}\left[\frac{1}{\sqrt{X+1}}\right] < \frac{1}{\sqrt{\lambda}}. \tag{30c}$$

Proof. Consider the LHS of Eq. (30a):

$$\mathbb{E}\left[\frac{1}{X+1}\right] = \sum_{k=0}^{\infty} \frac{\mathbb{P}(X = k)}{k+1} = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k+1)!} = \frac{e^{-\lambda}}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!} = \frac{1 - e^{-\lambda}}{\lambda},$$

where we use the fact that X is Poisson distributed and the series expansion of the exponential.

Similarly, consider the LHS of Eq. (30b):

$$\begin{aligned}\mathbb{E}\left[\frac{1}{X+2}\right] &= \sum_{k=0}^{\infty} \frac{\mathbb{P}(X = k)}{k+2} = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k(k+1)}{(k+2)!} = e^{-\lambda}\frac{d}{d\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+2)!} \\ &= e^{-\lambda}\frac{d}{d\lambda}\frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+2}}{(k+2)!} = e^{-\lambda}\frac{d}{d\lambda}\frac{e^{\lambda} - 1 - \lambda}{\lambda} = \frac{\lambda - 1 + e^{-\lambda}}{\lambda^2}.\end{aligned}$$

Next, consider the upper bound in Eq. (30c). Due to concavity of the square root, Jensen's inequality yields:

$$\mathbb{E}\left[\frac{1}{\sqrt{X+1}}\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{X+1}\right]} = \sqrt{\frac{1 - e^{-\lambda}}{\lambda}} < \frac{1}{\sqrt{\lambda}},$$

for $\lambda > 0$, and using Eq. (30a).

Finally, consider another random variable Y independent and identically distributed (i.i.d.) as X, i.e. with the rate parameter λ. Then the AM–GM inequality for $X + 1$ and $Y + 1$ implies:

$$\sqrt{(X+1)(Y+1)} \leq \frac{X + Y + 2}{2} \implies \mathbb{E}\left[\frac{1}{\sqrt{(X+1)(Y+1)}}\right] \geq 2\mathbb{E}\left[\frac{1}{X + Y + 2}\right].$$

Since X and Y are i.i.d. Poisson, $X + 1 \perp\!\!\!\perp Y + 1$ and $X + Y \sim \mathrm{Poisson}(\lambda)$, which when used above alongside Eq. (30b) yields:

$$\mathbb{E}\left[\frac{1}{\sqrt{X+1}}\right]\mathbb{E}\left[\frac{1}{\sqrt{Y+1}}\right] \geq \frac{2\lambda - 1 + e^{-2\lambda}}{2\lambda^2} \implies \mathbb{E}\left[\frac{1}{\sqrt{X+1}}\right]^2 > \frac{1}{\lambda} - \frac{1}{2\lambda^2},$$

for $\lambda > 0$, which yields the lower bound in Eq. (30c). □