Demystifying Oversmoothing in Attention-Based Graph Neural Networks

Xinyi Wu[1] **Amir Ajorlou**[1] **Zihui Wu**[2] **Ali Jadbabaie**[1]
[1]MIT [2] Caltech
{xinyiwu,ajorlou,jadbabai}@mit.edu zwu2@caltech.edu

Abstract

Oversmoothing in Graph Neural Networks (GNNs) refers to the phenomenon where increasing network depth leads to homogeneous node representations. While previous work has established that Graph Convolutional Networks (GCNs) exponentially lose expressive power, it remains controversial whether the graph attention mechanism can mitigate oversmoothing. In this work, we provide a definitive answer to this question, by viewing attention-based GNNs as nonlinear time-varying dynamical systems and incorporating tools and techniques from the theory of products of inhomogeneous matrices and the joint spectral radius. We establish that, contrary to popular belief, the graph attention mechanism cannot prevent oversmoothing and loses expressive power exponentially. The proposed framework extends the existing results on oversmoothing for symmetric GCNs to a significantly broader class of GNN models, including random walk GCNs, Graph Attention Networks (GATs) and (graph) transformers. In particular, our analysis accounts for asymmetric, state-dependent and time-varying aggregation operators and a wide range of common nonlinear activation functions, such as ReLU, LeakyReLU, GELU and SiLU.

1 Introduction

Graph neural networks (GNNs) have emerged as a powerful framework for learning with graph-structured data [1–7]. Most GNN models follow the *message-passing* paradigm [8], where the representation of each node is computed by recursively aggregating and transforming the representations of its neighboring nodes.

One notable drawback of repeated message-passing is *oversmoothing*, which refers to the phenomenon that stacking message-passing GNN layers makes node representations of the same connected component converge to the same vector [6, 9–14]. As a result, most GNNs used in practice remain relatively shallow and often only have few layers [6, 7, 15]. On the theory side, while previous works have shown that the symmetric Graph Convolution Networks (GCNs) with ReLU and LeakyReLU nonlinearities exponentially lose expressive power, analyzing the oversmoothing phenomenon in other types of GNNs is still an open question [10, 11]. In particular, the question of whether the graph attention mechanism can prevent oversmoothing has not been settled yet. Motivated by the capacity of graph attention to distinguish the importance of different edges in the graph, some works claim that oversmoothing is alleviated in Graph Attention Networks (GATs), heuristically crediting to GATs' ability to learn adaptive node-wise aggregation operators via the attention mechanism [16]. On the other hand, it has been empirically observed that similar to the case of GCNs, oversmoothing seems inevitable for attention-based GNNs such as GATs or (graph) transformers [14, 17].

In this work, we provide a definitive answer to this question — attention-based GNNs also lose expressive power exponentially, albeit potentially at a slower exponential rate compared to GCNs. Given that attention-based GNNs can be viewed as nonlinear time-varying dynamical systems, our analysis is built on the theory of products of inhomogeneous matrices [18, 19] and the concept of joint spectral radius [20], as these methods have been long proved effective in the analysis of

X. Wu et al., Demystifying Oversmoothing in Attention-Based Graph Neural Networks (Extended Abstract). Presented at the Second Learning on Graphs Conference (LoG 2023), Virtual Event, November 27–30, 2023.

time-inhomogeneous markov chains and ergodicity of dynamical systems [18, 19, 21]. Our approach generalizes the existing results on oversmoothing for symmetric GCNs to a significantly broader class of GNN models with asymmetric, state-dependent and time-varying aggregation operators and nonlinear activation functions under general conditions. In particular, our analysis accounts for asymmetric, state-dependent and time-varying aggregation operators and a wide range of common nonlinearities such as ReLU, LeakyReLU, and even non-monotone ones like GELU and SiLU.

2 Problem Setup

2.1 Graph attention mechanism and attention-based GNNs

Let $\mathcal{G}$ be a graph with N nodes. Given node representation vectors $X_i, X_j \in \mathbb{R}^d$, we use an attention function $\Psi : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}$ to compute a raw attention coefficient $e_{ij} = \Psi(W^\top X_i, W^\top X_j)$, $W \in \mathbb{R}^{d\times d'}$ that indicates the importance of node j's features to node i. Then the graph structure is injected into the mechanism by performing masked attention normalized using the softmax function, where for each node i, we only compute its attention to its neighbors: $P_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}_i}\exp(e_{ik})}$. The matrix P, where the ij^{th} entry is P_{ij}, is a row stochastic matrix. We refer to P as an *aggregation operator* in message-passing.

The update rule of a single graph attentional layer can be written as $X' = \sigma(PXW)$, where $\sigma(\cdot)$ is a pointwise nonlinearity function, and the aggregation operator P is a function of XW. As a result, the output of the t^{th} graph attentional layers can be written as $X^{(t+1)} = \sigma(P^{(t)}X^{(t)}W^{(t)})$, where $X^{(0)} = X \in \mathbb{R}^{N\times d}$ is the input node features, $W^{(t)} \in \mathbb{R}^{d'\times d'}$ for $t \in \mathbb{N}$ and $W^{(0)} \in \mathbb{R}^{d\times d'}$. For the rest of this work, without loss of generality, we assume that $d = d'$.

2.2 Measure of Oversmoothing

We establish our results on oversmoothing for attention-based GNNs using the following node similarity measure $\mu : \mathbb{R}^{N\times d} \to \mathbb{R}_{\geq 0}$, satisfying the criteria proposed in Rusch et al. [14]: $\mu(X) := \|X - \mathbf{1}\gamma_X\|_F$, where $\mathbf{1} \in \mathbb{R}^N$ is the all-one vector, $\gamma_X = \frac{\mathbf{1}^\top X}{N}$. In particular, $\mu(X) = 0$ if and only if all node representations converge to the same vector. Then oversmoothing with respect to μ is defined as the layer-wise exponential convergence of the node-similarity measure μ to zero, i.e. for $t \in \mathbb{N}$, with constants $C_1, C_2 > 0$,

$$\mu(X^{(t)}) \leq C_1 e^{-C_2 t}. \tag{1}$$

We note that our analysis directly applies to any Lipschitz node similarity measure, including the popular Dirichlet energy [11, 14].

2.3 Assumptions

We make the following assumptions (in fact, quite minimal) in deriving our results:

A1 The graph $\mathcal{G}$ is connected and has a self-loop at each node.
A2 The attention function $\Psi(\cdot,\cdot)$ is continuous.
A3 The sequence $\{\| \prod_{t=0}^{k} |W^{(t)}|\|_{\max}\}_{k=0}^{\infty}$ is bounded.
A4 The point-wise nonlinear activation function $\sigma(\cdot)$ satisfies $0 \leq \frac{\sigma(x)}{x} \leq 1$ for $x \neq 0$ and $\sigma(0) = 0$.

We note that all of these assumptions are either standard or quite general. The assumptions on the GNN architecture **A2** and **A4** can be easily verified for commonly used GNN designs. For example, the attention functions used in the GAT [7], GATv2 [22], and transformers [23] are specific cases that all satisfy **A2**. As for **A4**, one way to satisfy it is to have σ be 1-Lipschitz and $\sigma(x) \leq 0$ for $x < 0$ and $\sigma(x) \geq 0$ for $x > 0$. Then it is easy to verify that most of the commonly used nonlinear activation functions such as ReLU, LeakyReLU, GELU, SiLU, ELU, tanh all satisfy **A4**.

3 Main Results

3.1 Common connectivity structure among aggregation operators across different layers

Through writing the i^{th} column of $X^{(t+1)}$ as

$$X_{\cdot i}^{(t+1)} = \sum_{j_{t+1}=i,\,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} W_{j_k j_{k+1}}^{(k)}\right) D_{j_{t+1}}^{(t)} P^{(t)} \ldots D_{j_1}^{(0)} P^{(0)} X_{j_0}^{(0)}, \tag{2}$$

where $D_i^{(t)}$ is a diagonal matrix representing the effect of $\sigma(\cdot)$ to the i^{th} column of $P^{(t)}X^{(t)}W^{(t)}$, we can show the boundedness of the node representations' trajectories $X^{(t)}$ for all $t \geq 0$. We define $\mathcal{D}$ to be the set of all possible diagonal matrices $D_i^{(t)}$ satisfying **A4**: $\mathcal{D} := \{\text{diag}(\mathbf{d}) : \mathbf{d} \in \mathbb{R}^N, \mathbf{0} \leq_{\text{ew}} \mathbf{d} \leq_{\text{ew}} \mathbf{1}\}$. Then we establish the following key lemma suggesting that the graph attention mechanism cannot fundamentally change the connectivity pattern of the graph.

Lemma 1. *Under* **A2**-**A4**, *there exists* $\epsilon > 0$ *such that for all* $t \geq 0$ *and for any nodes* i, j *that are connected, we have* $P_{ij}^{(t)} \geq \epsilon$.

We define the family of row-stochastic matrices satisfying Lemma 1 below.

Definition 1. *Let* $\epsilon > 0$. *We define* $\mathcal{P}_{\mathcal{G},\epsilon}$ *to be the set of row-stochastic matrices satisfying the following conditions: 1.* $\epsilon \leq P_{ij} \leq 1$, *for* i, j *connected; 2.* $P_{ij} = 0$, *for* i, j *not connected.*

3.2 Ergodicity of infinite products of matrices

Ergodicity, in its most general form, deals with the long-term behavior of dynamical systems. The oversmoothing phenomenon in GNNs defined in the sense of (1) concerns the convergence of all rows of $X^{(t)}$ to a common vector at an exponential rate. To this end, we define ergodicity in our analysis as the convergence of infinite matrix products to a rank-one matrix with identical rows.

Definition 2 (Ergodicity)**.** *Let* $B \in \mathbb{R}^{(N-1)\times N}$ *be the orthogonal projection onto the space orthogonal to* $\text{span}\{\mathbf{1}\}$. *A sequence of matrices* $\{M^{(n)}\}_{n=0}^{\infty}$ *is ergodic if* $\lim_{t\to\infty} B\prod_{n=0}^{t} M^{(n)} = 0$.

We will take advantage of the following properties of the projection matrix B already established in Blondel et al. [21]:1. $B\mathbf{1} = 0$; 2. $\|Bx\|_2 = \|x\|_2$ for $x \in \mathbb{R}^N$ if $x^\top \mathbf{1} = 0$; 3. For $M \in \mathbb{R}^{N\times N}$, there exists a unique matrix $\tilde{M} \in \mathbb{R}^{(N-1)\times(N-1)}$ such that $BM = \tilde{M}B$.

Let $\mathcal{M}_{\mathcal{G},\epsilon} := \{DP : D \in \mathcal{D}, P \in \mathcal{P}_{\mathcal{G},\epsilon}\}$. Then any infinite product of matrices in $\mathcal{M}_{\mathcal{G},\epsilon}$ is ergodic.

Lemma 2. *Any sequence* $\{D^{(t)}P^{(t)}\}_{t=0}^{\infty}$ *in* $\mathcal{M}_{\mathcal{G},\epsilon}$ *is ergodic.*

3.3 Joint spectral radius

Finally, we make use of the concept of the joint spectral radius for a set of matrices [20] and employ it to deduce exponential convergence of node representations from our ergodicity result, Lemma 2.

Definition 3 (Joint Spectral Radius)**.** *For a collection of matrices* $\mathcal{M}$, *the joint spectral radius* $\text{JSR}(\mathcal{M})$ *is defined to be*

$$\text{JSR}(\mathcal{M}) = \limsup_{k\to\infty} \sup_{M_1,M_2,\ldots,M_k\in\mathcal{M}} \|M_1M_2\ldots M_k\|^{\frac{1}{k}},$$

and it is independent of the norm used.

To analyze the convergence rate of products of matrices in $\mathcal{M}_{\mathcal{G},\epsilon}$ to a rank-one matrix with identical rows, we investigate the dynamics induced by the matrices on the subspace orthogonal to $\text{span}\{\mathbf{1}\}$. More precisely, with the notion of ergodicity in Definition 2 and the goal of studying the convergence rate of a matrix product $BM_1M_2\ldots M_k$ where each $M_i \in \mathcal{M}_{\mathcal{G},\epsilon}$, we use the third property of the orthogonal projection B to write

$$BM_1M_2\ldots M_k = \tilde{M}_1\tilde{M}_2\ldots\tilde{M}_kB,$$

where each $\tilde{M}_i$ is the unique matrix in $\mathbb{R}^{(N-1)\times(N-1)}$ that satisfies $BM_i = \tilde{M}_iB$. To analyze products of such matrices $\tilde{M}_i$, let us define $\tilde{\mathcal{M}}_{\mathcal{G},\epsilon} := \{\tilde{M} : BM = \tilde{M}B, M \in \mathcal{M}_{\mathcal{G},\epsilon}\}$. We can use the ergodictiy result developed in Lemma 2 to show that the joint spectral radius of $\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}$ is strictly less than 1.

Lemma 3. *Let* $0 < \epsilon < 1$. *Under assumptions* **A1**-**A4**, $\text{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < 1$.

It follows from the definition of the joint spectral radius that if $\text{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < 1$, for any $\text{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < q < 1$, there exists a C for which

$$\|\tilde{M}_1\tilde{M}_2\ldots\tilde{M}_ky\| \leq Cq^k\|y\| \tag{3}$$

for all $y \in \mathbb{R}^{N-1}$ and $\tilde{M}_1, \tilde{M}_2, \ldots, \tilde{M}_k \in \tilde{\mathcal{M}}_{\mathcal{G},\epsilon}$.

3.4 Main Theorem

Applying (3) to the recursive expansion of $X_{\cdot i}^{(t+1)}$ in (2) using the 2-norm, we can prove the exponential convergence of $\mu(X^{(t)})$ to zero for the similarity measure $\mu(\cdot)$ defined in (1), which in turn implies the convergence of node representations to a common representation at an exponential rate. This completes the proof of the main result of this paper, which states that oversmoothing defined in (1) is unavoidable for attention-based GNNs.

Theorem 1 (Oversmoothing happens exponentially in attention-based GNNs)**.** *Under assumptions* **A1**-**A4**, $\mathrm{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < 1$ *and for any* q *satisfying* $\mathrm{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < q < 1$*, there exists* $C_1(q) > 0$ *such that*

$$\mu(X^{(t)}) \leq C_1 q^t \,, \forall t \geq 0 \,,$$

where $\mu(X) = \|X - \frac{\mathbf{1}\mathbf{1}^\top X}{N}\|_F$*. As a result, node representations* $X^{(t)}$ *exponentially converge to the same value as the model depth* $t \to \infty$*.*

3.5 Comparison with the GCN

Computing or approximating the joint spectral radius for a given set of matrices is known to be hard in general [24], yet it is straightforward to lower bound $\mathrm{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon})$.

Proposition 1. *Let* λ *be the second largest eigenvalue of* $D_{\mathrm{deg}}^{-1/2} A D_{\mathrm{deg}}^{-1/2}$*. Then under assumptions* **A1**-**A4**, *it holds that* $\lambda \leq \mathrm{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon})$.

A direct consequence of the above result is that the upper bound q on the convergence rate that we get for graph attention in Theorem 1 is at least as large as λ. On the other hand, previous work has already established that in the graph convolution case, the convergence rate of $\mu(X^{(t)})$ is $O(\lambda^t)$ [10, 11]. It is thus natural to expect attention-based GNNs to potentially have better expressive power at finite depth than GCNs, even though they both inevitably suffer from oversmoothing. This is also evident from the numerical experiments that we present in the next section.

4 Numerical Experiments

We validate our theoretical results on three real-world datasets: Cora, CiteSeer and PubMed [25] with two attention-based GNN architectures (GAT [7] and random walk GCN (constant attention function) [12, 13]) and five common nonlinearities. We ran each experiment 10 times. Figure 1 shows the evolution of $\mu(X^{(t)})$ in log-log scale on the largest connected component of each graph as we forward pass the input X into a trained model. The solid curve is the average over 10 runs and the band indicates one standard deviation around the average. We observe that oversmoothing happens exponentially in both GCNs and GATs with the rates varying depending on the choice of activation function. Notably, GCNs demonstrate faster rates of oversmoothing compared to GATs.

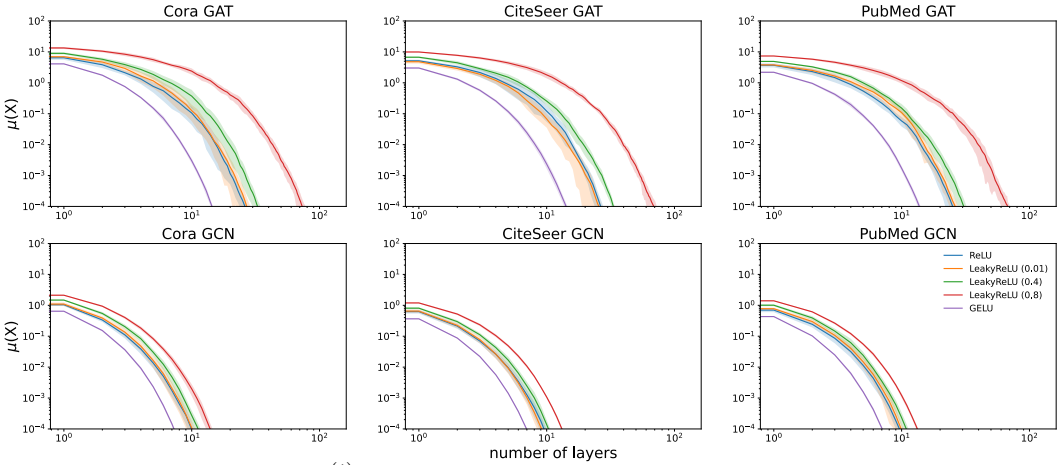

Figure 1: Evolution of $\mu(X^{(t)})$ (in log-log scale) on the largest connected component of three benchmark datasets: Cora, Citeseer, and PubMed.

Acknowledgements

This research has been supported in part by ARO MURI W911NF-19-0217, ONR N00014-20-1-2394, and the MIT-IBM Watson AI Lab.

References

[1] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *IJCNN*, 2005. 1

[2] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20:61–80, 2009.

[3] Joan Bruna, Wojciech Zaremba, Arthur D. Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *ICLR*, 2014.

[4] David Kristjanson Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy D. Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NeurIPS*, 2015.

[5] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*, 2016.

[6] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 1

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. 1, 2, 4, 13

[8] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017. 1

[9] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*, 2018. 1

[10] Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *ICLR*, 2020. 1, 4

[11] Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. In *ICML Graph Representation Learning and Beyond (GRL+) Workshop*, 2020. 1, 2, 4

[12] Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over)smoothing. In *NeurIPS*, 2022. 4

[13] Xinyi Wu, Zhengdao Chen, William Wang, and Ali Jadbabaie. A non-asymptotic analysis of oversmoothing in graph neural networks. In *ICLR*, 2023. 4

[14] T.Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *ArXiv*, abs/2303.10993, 2023. 1, 2

[15] 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*, 32:4–24, 2019. 1

[16] Yimeng Min, Frederik Wenkel, and Guy Wolf. Scattering gcn: Overcoming oversmoothness in graph convolutional networks. In *NeurIPS*, 2020. 1

[17] Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen M. S. Lee, and James Tin-Yau Kwok. Revisiting over-smoothing in bert from the perspective of graph. In *ICLR*, 2022. 1

[18] Darald J. Hartfiel. *Nonhomogeneous Matrix Products*. 2002. 1, 2, 8

[19] Eugene Seneta. *Non-negative Matrices and Markov Chains*. 2008. 1, 2, 8

[20] Gian-Carlo Rota and W. Gilbert Strang. A note on the joint spectral radius. 1960. 1, 3

[21] Vincent D. Blondel, Julien M. Hendrickx, Alexander Olshevsky, and John N. Tsitsiklis. Convergence in multiagent coordination, consensus, and flocking. *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 2996–3000, 2005. 2, 3

[22] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *ICLR*, 2022. 2

[23] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[24] John N. Tsitsiklis and Vincent D. Blondel. The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate. *Mathematics of Control, Signals and Systems*, 10:31–40, 1997. 4

[25] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *ICML*, 2016. 4

[26] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. 2008. 8

[27] Jacques Theys. Joint spectral radius: theory and approximations. *Ph. D. dissertation*, 2005. 11

[28] Peter D. Lax. *Functional Analysis*. 2002. 13

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach 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 *NeurIPS*, 2019. 13

[30] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 13

A Basic Facts about Matrix Norms

In this section, we list some basic facts about matrix norms that will be helpful in comprehending the subsequent proofs.

A.1 Matrix norms induced by vector norms

Suppose a vector norm $\|\cdot\|_\alpha$ on $\mathbb{R}^n$ and a vector norm $\|\cdot\|_\beta$ on $\mathbb{R}^m$ are given. Any matrix $M \in \mathbb{R}^{m\times n}$ induces a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$ with respect to the standard basis, and one defines the corresponding *induced norm* or *operator norm* by

$$\|M\|_{\alpha,\beta} = \sup\left\{ \frac{\|Mv\|_\beta}{\|v\|_\alpha}, v \in \mathbb{R}^n, v \neq \mathbf{0} \right\}.$$

If the p-norm for vectors ($1 \leq p \leq \infty$) is used for both spaces $\mathbb{R}^n$ and $\mathbb{R}^m$, then the corresponding operator norm is

$$\|M\|_p = \sup_{v\neq\mathbf{0}} \frac{\|Mv\|_p}{\|v\|_p}.$$

The matrix 1-norm and ∞-norm can be computed by

$$\|M\|_1 = \max_{1\leq j} \sum_{i=1}^{m} |M_{ij}|,$$

that is, the maximum absolute column sum of the matrix M;

$$\|M\|_\infty = \max_{1\leq m} \sum_{j=1}^{n} |M_{ij}|,$$

that is, the maximum absolute row sum of the matrix M.

Remark. In the special case of $p = 2$, the induced matrix norm $\|\cdot\|_2$ is called the *spectral norm*, and is equal to the largest singular value of the matrix.

For square matrices, we note that the name "spectral norm" does not imply the quantity is directly related to the spectrum of a matrix, unless the matrix is symmetric.

Example. We give the following example of a stochastic matrix P, whose spectral radius is 1, but its spectral norm is greater than 1.

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.25 & 0.75 \end{bmatrix} \qquad \|P\|_2 \approx 1.0188$$

A.2 Matrix (p, q)-norms

The Frobenius norm of a matrix $M \in \mathbb{R}^{m\times n}$ is defined as

$$\|M\|_F = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{m} |M_{ij}|^2}\,,$$

and it belongs to a family of entry-wise matrix norms: for $1 \le p, q \le \infty$, the matrix (p, q)-norm is defined as

$$\|M\|_{p,q} = \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{m} |M_{ij}|^p\right)^{q/p}\right)^{1/q}.$$

The special case $p = q = 2$ is the Frobenius norm $\|\cdot\|_F$, and $p = q = \infty$ yields the max norm $\|\cdot\|_{\max}$.

A.3 Equivalence of norms

For any two matrix norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, we have that for all matrices $M \in \mathbb{R}^{m\times n}$,

$$r\|M\|_\alpha \le \|M\|_\beta \le s\|M\|_\alpha$$

for some positive numbers r and s. In particular, the following inequality holds for the 2-norm $\|\cdot\|_2$ and the ∞-norm $\|\cdot\|_\infty$:

$$\frac{1}{\sqrt{n}}\|M\|_\infty \le \|M\|_2 \le \sqrt{m}\|M\|_\infty\,.$$

B Proof of Lemma 1

We can use the formulation in (2) to show the boundedness of the node representations' trajectories $X^{(t)}$ for all $t \in \mathbb{N}_{\ge 0}$, which in turn implies the boundedness of the input to graph attention in each layer, $X^{(t)}W^{(t)}$.

Lemma 4. *Under assumptions* **A3**-**A4***, there exists $C > 0$ such that $\|X^{(t)}\|_{\max} \le C$ for all $t \in \mathbb{N}_{\ge 0}$.*

Proof of Lemma 4. According to the formulation (2):

$$X_{\cdot i}^{(t+1)} = \sum_{j_{t+1}=i,\,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} W_{j_k j_{k+1}}^{(k)}\right) D_{j_{t+1}}^{(t)} P^{(t)} \ldots D_{j_1}^{(0)} P^{(0)} X_{j_0}^{(0)}\,,$$

we thus obtain that

$$\begin{aligned}
\|X_{\cdot i}^{(t+1)}\|_\infty &= \left\| \sum_{j_{t+1}=i\,,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} W_{j_k j_{k+1}}^{(k)}\right) D_{j_{t+1}}^{(t)} P^{(t)} \ldots D_{j_1}^{(0)} P^{(0)} X_{j_0}^{(0)} \right\|_\infty \\
&\le \sum_{j_{t+1}=i\,,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} \left|W_{j_k j_{k+1}}^{(k)}\right|\right) \left\| D_{j_{t+1}}^{(t)} P^{(t)} \ldots D_{j_1}^{(0)} P^{(0)} \right\|_\infty \left\| X_{j_0}^{(0)} \right\|_\infty \\
&\le \sum_{j_{t+1}=i\,,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} \left|W_{j_k j_{k+1}}^{(k)}\right|\right) \left\| X_{j_0}^{(0)} \right\|_\infty \\
&\le C_0 \left(\sum_{j_{t+1}=i\,,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} \left|W_{j_k j_{k+1}}^{(k)}\right|\right) \right) \\
&= C_0 \|(|W^{(0)}|\ldots|W^{(t)}|)_{\cdot i}\|_1\,,
\end{aligned}$$

where C_0 equals the maximal entry in $|X^{(0)}|$.

The assumption **A3** implies that there exists $C' > 0$ such that for all $t \in \mathbb{N}_{\geq 0}$ and $i \in [d]$,

$$\|(|W^{(0)}|...|W^{(t)}|)_{\cdot i}\|_1 \leq C'N\,.$$

Hence there exists $C'' > 0$ such that for all $t \in \mathbb{N}_{\geq 0}$ and $i \in [d]$, we have

$$\|X^{(t)}_{\cdot i}\|_\infty \leq C''\,,$$

proving the existence of $C > 0$ such that $\|X^{(t)}\|_{\max} \leq C$ for all $t \in \mathbb{N}_{\geq 0}$. □

For a continuous $\Psi(\cdot,\cdot)$[1], the following lemma is a direct consequence of Lemma 4, suggesting that the graph attention mechanism cannot fundamentally change the connectivity pattern of the graph.

C Proof of Lemma 2

C.1 Ergodicity of infinite products of matrices in $\mathcal{P}_{\mathcal{G},\epsilon}$

In this section, we make use the existing results on the ergodicity of infinite products of inhomogeneous stochastic matrices [18, 19] to show that any sequence of matrices in $\mathcal{P}_{\mathcal{G},\epsilon}$ is ergodic.

Lemma 5. *Fix $\epsilon > 0$. Consider a sequence of matrices $\{P^{(t)}\}_{t=0}^{\infty}$ in $\mathcal{P}_{\mathcal{G},\epsilon}$. That is, $P^{(t)} \in \mathcal{P}_{\mathcal{G},\epsilon}$ for all $t \in \mathbb{N}_{\geq 0}$. Then $\{P^{(t)}\}_{t=0}^{\infty}$ is ergodic.*

Proof of Lemma 5. The following sufficient condition guarantees the ergodicity of the infinite products of row-stochastic matrices.

Lemma 6 (Corollary 5.1 [18])**.** *Consider a sequence of row-stochastic matrices $\{S^{(t)}\}_{t=0}^{\infty}$. Let a_t and b_t be the smallest and largest entries in $S^{(t)}$, respectively. If $\sum_{t=0}^{\infty} \frac{a_t}{b_t} = \infty$, then $\{S^{(t)}\}_{t=0}^{\infty}$ is ergodic.*

In order to make use of the above result, we first show that long products of $P^{(t)}$'s from $\mathcal{P}_{\mathcal{G},\epsilon}$ will eventually become strictly positive. For $t_0 \leq t_1$, we denote

$$P^{(t_1:t_0)} = P^{(t_1)} \dots P^{(t_0)}\,.$$

Lemma 7. *Under the assumption* **A1***, there exist $T \in \mathbb{N}$ and $c > 0$ such that for all $t_0 \geq 0$,*

$$c \leq P_{ij}^{(t_0+T:t_0)} \leq 1\,, \forall 1 \leq i,j \leq N\,.$$

Proof of Lemma 7. Fix any $T \in \mathbb{N}_{\geq 0}$. Since $\|P^{(t)}\|_\infty \leq 1$ for any $P^{(t)} \in \mathcal{P}_{\mathcal{G},\epsilon}$, it follows that $\|P^{(t_0+T:t_0)}\|_\infty \leq 1$ and hence $P_{ij}^{(t_0+T:t_0)} \leq 1$, for all $1 \leq i,j \leq N$.

To show the lower bound, without loss of generality, we will show that there exist $T \in \mathbb{N}$ and $c > 0$ such that

$$P_{ij}^{(T:0)} \geq c\,, \forall 1 \leq i,j \leq N\,.$$

Since each $P^{(t)}$ has the same connectivity pattern as the original graph $\mathcal{G}$, it follows from the assumption **A1** that there exists $T \in \mathbb{N}$ such that $P^{(T:0)}$ is a positive matrix, following a similar argument as the one for Proposition 1.7 in [26]: For each pair of nodes i, j, since we assume that the graph $\mathcal{G}$ is connected, there exists $r(i,j)$ such that $P_{ij}^{(r(i,j):0)} > 0$. on the other hand, since we also assume each node has a self-loop, $P_{ii}^{(t:0)} > 0$ for all $t \geq 0$ and hence for $t \geq r(i,j)$,

$$P_{ij}^{(t:0)} \geq P_{ii}^{(t-r(i,j))} P_{ij}^{(r(i,j):0)} > 0\,.$$

[1] More generally, for $\Psi(\cdot,\cdot)$ that outputs bounded attention scores for bounded inputs.

For $t \geq t(i) := \max_{j \in \mathcal{G}} r(i,j)$, we have $P_{ij}^{(t:0)} > 0$ for all node j in $\mathcal{G}$. Finally, if $t \geq T := \max_{i \in \mathcal{G}} t(i)$, then $P_{ij}^{(t:0)} > 0$ for all pairs of nodes i, j in $\mathcal{G}$. Notice that $P_{ij}^{(T:0)}$ is a weighted sum of walks of length T between nodes i and j, and hence $P_{ij}^{(T:0)} > 0$ if and only if there exists a walk of length T between nodes i and j. Since for all $t \in \mathbb{N}_{\geq 0}$, $P_{ij}^{(t)} \geq \epsilon$ if $(i,j) \in E(\mathcal{G})$, we conclude that $P_{ij}^{(T:0)} \geq \epsilon^T := c$. □

Given the sequence $\{P^{(t)}\}_{t=0}^{\infty}$, we use $T \in \mathbb{N}$ from Lemma 7 and define

$$\bar{P}^{(k)} := P^{((k+1)T:kT)} .$$

Then $\{P^{(t)}\}_{t=0}^{\infty}$ is ergodic if and only if $\{\bar{P}^{(k)}\}_{k=0}^{\infty}$ is ergodic. Notice that by Lemma 7, for all $k \in \mathbb{N}_{\geq 0}$, there exists $c > 0$ such that $c \leq \bar{P}_{ij}^{(k)} \leq 1 \,, \forall 1 \leq i,j \leq N$. Then Lemma 5 is a direct consequence of Lemma 6. □

C.2 Notations and auxiliary results

Consider a sequence $\{D^{(t)}P^{(t)}\}_{t=0}^{\infty}$ in $\mathcal{M}_{\mathcal{G},\epsilon}$. For $t_0 \leq t_1$, define

$$Q_{t_0,t_1} := D^{(t_1)}P^{(t_1)}...D^{(t_0)}P^{(t_0)}$$

and

$$\delta_t = \|D^{(t)} - I_N\|_\infty ,$$

where I_N denotes the $N \times N$ identity matrix. It is also useful to define

$$\begin{aligned}\hat{Q}_{t_0,t_1} :=& P^{(t_1)}Q_{t_0,t_1-1} \\ :=& P^{(t_1)}D^{(t_1-1)}P^{(t_1-1)}...D^{(t_0)}P^{(t_0)} .\end{aligned}$$

We start by proving the following key lemma, which states that long products of matrices in $\mathcal{M}_{\mathcal{G},\epsilon}$ eventually become a contraction in ∞-norm.

Lemma 8. *There exist* $0 < c < 1$ *and* $T \in \mathbb{N}$ *such that for all* $t_0 \leq t_1$,

$$\|\hat{Q}_{t_0,t_1+T}\|_\infty \leq (1 - c\delta_{t_1})\|\hat{Q}_{t_0,t_1}\|_\infty .$$

Proof of Lemma 8. First observe that for every $T \geq 0$,

$$\begin{aligned}\|\hat{Q}_{t_0,t_1+T}\|_\infty &\leq \|P^{(t_1+T)}D^{(t_1+T-1)}P^{(t_1+T-1)}...D^{(t_1+1)}P^{(t_1+1)}D^{(t_1)}\|_\infty \|\hat{Q}_{t_0,t_1}\|_\infty \\ &\leq \|P^{(t_1+T)}P^{(t_1+T-1)}...P^{(t_1+1)}D^{(t_1)}\|_\infty \|\hat{Q}_{t_0,t_1}\|_\infty ,\end{aligned}$$

where the second inequality is based on the following element-wise inequality:

$$P^{(t_1+T)}P^{(t_1+T-1)}...D^{(t_1+1)}P^{(t_1+1)} \leq_{\text{ew}} P^{(t_1+T)}P^{(t_1+T-1)}...P^{(t_1+1)} .$$

By Lemma 7, there exist $T \in \mathbb{N}$ and $0 < c < 1$ such that

$$(P^{(t_1+T)}...P^{(t_1+1)})_{ij} \geq c, \, \forall 1 \leq i,j \leq N .$$

Since the matrix product $P^{(t_1+T)}P^{(t_1+T-1)}...P^{(t_1+1)}$ is row-stochastic, multiplying it with the diagonal matrix $D^{(t_1)}$ from right decreases the row sums by at least $c(1 - D_{\min}^{(t_1)}) = c\delta_{t_1}$, where $D_{\min}^{(t_1)}$ here denotes the smallest diagonal entry of the diagonal matrix $D^{(t_1)}$. Hence,

$$\|P^{(t_1+T)}P^{(t_1+T-1)}...P^{(t_1+1)}D^{(t_1)}\|_\infty \leq 1 - c\delta_{t_1} .$$

□

Now define $\beta_k := \prod_{t=0}^{k}(1 - c\delta_t)$ and let $\beta := \lim_{k\to\infty} \beta_k$. Note that β is well-defined because the partial product is non-increasing and bounded from below. Then we present the following result, which is stated as Lemma 9 in the main paper and from which the ergodicity of any sequence in $\mathcal{M}_{\mathcal{G},\epsilon}$ is an immediate result.

Lemma 9. *Let* $\beta_k := \prod_{t=0}^{k}(1 - c\delta_t)$ *and* $\beta := \lim_{k\to\infty} \beta_k$.

1. *If* $\beta = 0$, *then* $\lim_{k\to\infty} Q_{0,k} = 0$;
2. *If* $\beta > 0$, *then* $\lim_{k\to\infty} BQ_{0,k} = 0$.

Proof of Lemma 9. We will prove the two cases separately.

[Case $\beta = 0$**].** We will show that $\beta = 0$ implies $\lim_{k\to\infty} \|\hat{Q}_{0,k}\|_\infty = 0$, and as a result, $\lim_{k\to\infty} \|Q_{0,k}\|_\infty = 0$. For $0 \le j \le T-1$, let us define

$$\beta^j := \prod_{k=0}^{\infty}(1 - \delta_{j+kT}) .$$

Then by Lemma 8, we get that

$$\lim_{k\to\infty} \|\hat{Q}_{0,kT}\|_\infty \le \beta^j \|\hat{Q}_{0,j}\|_\infty .$$

By construction, $\beta = \Pi_{j=0}^{T-1}\beta^j$. Hence, if $\beta = 0$ then $\beta^{j_0} = 0$ for some $0 \le j_0 \le T-1$, which yields $\lim_{k\to\infty} \|\hat{Q}_{0,k}\|_\infty = 0$. Consequently, $\lim_{k\to\infty} \|Q_{0,k}\|_\infty = 0$ implies that $\lim_{k\to\infty} Q_{0,k} = 0$.

[Case $\beta > 0$**].** First observe that if $\beta > 0$, then $\forall 0 < \eta < 1$, there exist $m \in \mathbb{N}_{\ge 0}$ such that

$$\prod_{t=m}^{\infty}(1 - c\delta_t) > 1 - \eta . \tag{4}$$

Using $1 - x \le e^{-x}$ for all $x \in \mathbb{R}$, we deduce

$$\prod_{t=m}^{\infty} e^{-c\delta_t} > 1 - \eta .$$

It also follows from (4) that $1 - c\delta_t > 1 - \eta$, or equivalently $\delta_t < \frac{\eta}{c}$ for $t \ge m$. Choosing $\eta < \frac{c}{2}$ thus ensures that $\delta_t < \frac{1}{2}$ for $t \ge m$. Putting this together with the fact that, there exists[2] $b > 0$ such that $1 - x \ge e^{-bx}$ for all $x \in [0, \frac{1}{2}]$, we obtain

$$\prod_{t=m}^{\infty}(1 - \delta_t) \ge \prod_{t=m}^{\infty} e^{-b\delta_t} > (1 - \eta)^{\frac{b}{c}} := 1 - \eta' . \tag{5}$$

Define the product of row-stochastic matrices $P^{(M:m)} := P^{(M)} \dots P^{(m)}$. It is easy to verify the following element-wise inequality:

$$\left(\prod_{t=m}^{M}(1 - c\delta_t)\right) P^{(M:m)} \le_{\text{ew}} Q_{m,M} \le_{\text{ew}} P^{(M:m)} ,$$

which together with (5) leads to

$$(1 - \eta')P^{(M:m)} \le_{\text{ew}} Q_{m,M} \le_{\text{ew}} P^{(M:m)} . \tag{6}$$

Therefore,

$$\begin{aligned}
\|BQ_{m,M}\|_\infty &= \|B(Q_{m,M} - P^{(M:m)}) + BP^{(M:m)}\|_\infty \\
&\le \|B(Q_{m,M} - P^{(M:m)})\|_\infty + \|BP^{(M:m)}\|_\infty \\
&= \|B(Q_{m,M} - P^{(M:m)})\|_\infty \\
&\le \|B\|_\infty \|Q_{m,M} - P^{(M:m)}\|_\infty \\
&\le \eta' \|B\|_\infty \\
&\le \eta' \sqrt{N} ,
\end{aligned}$$

[2] Choose, e.g., $b = 2\log 2$.

where the last inequality is due to the fact that $\|B\|_2 = 1$. By definition, $Q_{0,M} = Q_{m,M}Q_{0,m-1}$, and hence

$$\|BQ_{0,M}\|_\infty \leq \|BQ_{m,M}\|_\infty \|Q_{0,m-1}\|_\infty \leq \|BQ_{m,M}\|_\infty \leq \eta'\sqrt{N}\,. \tag{7}$$

The above inequality (7) holds when taking $M \to \infty$. Then taking $\eta \to 0$ implies $\eta' \to 0$ and together with (7), we conclude that

$$\lim_{M\to\infty} \|BQ_{0,M}\|_\infty = 0\,,$$

and therefore,

$$\lim_{M\to\infty} BQ_{0,M} = 0\,.$$

□

C.3 Proof of Lemma 2

Notice that both cases $\beta = 0$ and $\beta > 0$ in Lemma 9 imply the ergodicity of $\{D^{(t)}P^{(t)}\}_{t=0}^{\infty}$. Hence the statement is a direct corollary of Lemma 9.

D Proof of Lemma 3

In order to show that $\mathrm{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < 1$, we start by making the following observation.

Lemma 10. *A sequence $\{M^{(n)}\}_{n=0}^{\infty}$ is ergodic if and only if $\prod_{n=0}^{t} \tilde{M}^{(n)}$ converges to the zero matrix.*

Proof of Lemma 10. For any $t \in \mathbb{N}_{\geq 0}$, it follows from the third property of the orthogonal projection B (see, Page 6 of the main paper) that

$$B\prod_{n=0}^{t} M^{(n)} = \prod_{n=0}^{t} \tilde{M}^{(n)} B\,.$$

Hence

$$\begin{aligned}
\{M^{(n)}\}_{n=0}^{\infty} \text{ is ergodic} &\iff \lim_{t\to\infty} B\prod_{n=0}^{t} M^{(n)} = 0 \\
&\iff \lim_{t\to\infty} \prod_{n=0}^{t} \tilde{M}^{(n)} B = 0 \\
&\iff \lim_{t\to\infty} \prod_{n=0}^{t} \tilde{M}^{(n)} = 0\,.
\end{aligned}$$

□

Next, we utilize the following result, as a means to ensure a joint spectral radius strictly less than 1 for a bounded set of matrices.

Lemma 11 (Proposition 3.2 in [27])**.** *For any bounded set of matrices $\mathcal{M}$, $\mathrm{JSR}(\mathcal{M}) < 1$ if and only if for any sequence $\{M^{(n)}\}_{n=0}^{\infty}$ in $\mathcal{M}$, $\prod_{n=0}^{t} M^{(n)}$ converges to the zero matrix.*

Here, “bounded" means that there exists an upper bound on the norms of the matrices in the set. Note that $\mathcal{M}_{\mathcal{G},\epsilon}$ is bounded because $\|DP\|_\infty \leq 1$, $DP \in \mathcal{M}_{\mathcal{G},\epsilon}$. To show that $\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}$ is also bounded, let $\tilde{M} \in \tilde{\mathcal{M}}_{\mathcal{G},\epsilon}$, then by definition, we have

$$\tilde{M}B = BM, M \in \mathcal{M}_{\mathcal{G},\epsilon} \Rightarrow \tilde{M} = BMB^T\,,$$

since $BB^T = I_{N-1}$. As a result,

$$\|\tilde{M}\|_2 = \|BMB^T\|_2 \leq \|M\|_2 \leq \sqrt{N}\,,$$

where the first inequality is due to $\|B\|_2 = \|B^\top\|_2 = 1$, and the second ineuality follows from $\|M\|_\infty \leq 1$.

Combining Lemma 2, Lemma 10 and Lemma 11, we conclude that $\mathrm{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < 1$.

E Proof of Theorem 1

Recall the formulation of $X_{\cdot i}^{(t+1)}$ in (2):

$$X_{\cdot i}^{(t+1)} = \sigma(P^{(t)}(X^{(t)}W^{(t)})_{\cdot i}) = \sum_{j_{t+1}=i,\,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} W_{j_k j_{k+1}}^{(k)}\right) D_{j_{t+1}}^{(t)} P^{(t)} \ldots D_{j_1}^{(0)} P^{(0)} X_{j_0}^{(0)} .$$

Then it follows that

$$\begin{aligned}
\|BX_{\cdot i}^{(t+1)}\|_2 &= \left\| \sum_{j_{t+1}=i,\,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} W_{j_k j_{k+1}}^{(k)}\right) B D_{j_{t+1}}^{(t)} P^{(t)} \ldots D_{j_1}^{(0)} P^{(0)} X_{j_0}^{(0)} \right\|_2 \\
&\leq \sum_{j_{t+1}=i,\,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} \left|W_{j_k j_{k+1}}^{(k)}\right|\right) \left\| B D_{j_{t+1}}^{(t)} P^{(t)} \ldots D_{j_1}^{(0)} P^{(0)} X_{j_0}^{(0)} \right\|_2 \\
&= \sum_{j_{t+1}=i,\,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} \left|W_{j_k j_{k+1}}^{(k)}\right|\right) \left\| \tilde{D}_{j_{t+1}}^{(t)} \tilde{P}^{(t)} \ldots \tilde{D}_{j_1}^{(0)} \tilde{P}^{(0)} B X_{j_0}^{(0)} \right\|_2 \\
&\leq \sum_{j_{t+1}=i,\,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} \left|W_{j_k j_{k+1}}^{(k)}\right|\right) C q^{t+1} \left\| B X_{j_0}^{(0)} \right\|_2 \\
&\leq C' q^{t+1} \left(\sum_{j_{t+1}=i,\,(j_t,\ldots,j_0)\in[d]^{t+1}} \left(\prod_{k=0}^{t} \left|W_{j_k j_{k+1}}^{(k)}\right|\right) \right) \\
&= C' q^{t+1} \|(|W^{(0)}|\ldots|W^{(t)}|)_{\cdot i}\|_1 ,
\end{aligned}$$

where $C' = C \max_{j\in[d]} \|BX_j^{(0)}\|_2$ and $\|\cdot\|_1$ denotes the 1-norm. Specifically, the first inequality follows from the triangle inequality, and the second inequality is due to the property of the joint spectral radius in (3), where $\mathrm{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < q < 1$.

Since $\|Bx\|_2 = \|x\|_2$ if $x^\top \mathbf{1} = 0$ for $x \in \mathbb{R}^N$, we also have that if $X^\top \mathbf{1} = 0$ for $X \in \mathbb{R}^{N\times d}$, then

$$\|BX\|_F = \|X\|_F ,$$

using which we obtain that

$$\begin{aligned}
\mu(X^{(t+1)}) = \|X^{(t+1)} - \mathbf{1}\gamma_{X^{(t+1)}}\|_F = \|BX^{(t+1)}\|_F &= \sqrt{\sum_{i=1}^{d} \|BX_{\cdot i}^{(t+1)}\|_2^2} \\
&\leq C' q^{t+1} \sqrt{\sum_{i=1}^{d} \|(|W^{(0)}|\ldots|W^{(t)}|)_{\cdot i}\|_1^2} \\
&\leq C' q^{t+1} \sqrt{\left(\sum_{i=1}^{d} \|(|W^{(0)}|\ldots|W^{(t)}|)_{\cdot i}\|_1\right)^2} \\
&= C' q^{t+1} \|(|W^{(0)}|\ldots|W^{(t)}|\|_{1,1} ,
\end{aligned}$$

where $\|\cdot\|_{1,1}$ denotes the matrix $(1,1)$-norm (recall from Section A.2 that for a matrix $M \in \mathbb{R}^{m\times n}$, we have $\|M\|_{1,1} = \sum_{i=1}^m \sum_{j=1}^n |M_{ij}|$). The assumption **A3** implies that there exists C'' such that for all $t \in \mathbb{N}_{\geq 0}$,

$$\|(|W^{(0)}|\ldots|W^{(t)}|)\|_{1,1} \leq C'' d^2 .$$

Thus we conclude that there exists C_1 such that for all $t \in \mathbb{N}_{\geq 0}$,

$$\mu(X^{(t)}) \leq C_1 q^t .$$

F Proof of Proposition 1

Since $D_{\text{deg}}^{-1}A$ is similar to $D_{\text{deg}}^{-1/2}AD_{\text{deg}}^{-1/2}$, they have the same spectrum. For $D_{\text{deg}}^{-1}A$, the smallest nonzero entry has value $1/d_{\max}$, where $d_{\max}$ is the maximum node degree in $\mathcal{G}$. On the other hand, it follows from the definition of $\mathcal{P}_{\mathcal{G},\epsilon}$ that

$$\epsilon d_{\max} \leq 1\,.$$

Therefore, $\epsilon \leq 1/d_{\max}$ and thus $D_{\text{deg}}^{-1}A \in \mathcal{P}_{\mathcal{G},\epsilon}$.

We proceed by proving the following result.

Lemma 12. *For any M in $\mathcal{M}$, the spectral radius of M denoted by $\rho(M)$, satisfies*

$$\rho(M) \leq \text{JSR}(\mathcal{M})\,.$$

Proof of Lemma 12. Gelfand's formula states that $\rho(M) = \lim_{k\to\infty} \|M^k\|^{\frac{1}{k}}$, where the quantity is independent of the norm used [28]. Then comparing with the definition of the joint spectral radius, we can immediately conclude the statement. □

Let $B(D_{\text{deg}}^{-1}A) = \tilde{P}B$. By definition, $\tilde{P} \in \tilde{\mathcal{M}}_{\mathcal{G},\epsilon}$ since $D_{\text{deg}}^{-1}A \in \mathcal{P}_{\mathcal{G},\epsilon}$ as shown before the lemma. Moreover, the spectrum of $\tilde{P}$ is the spectrum of $D_{\text{deg}}^{-1}A$ after reducing the multiplicity of eigenvalue 1 by one. Under the assumption **A1**, the eigenvalue 1 of $D_{\text{deg}}^{-1}A$ has multiplicity 1, and hence $\rho(\tilde{P}) = \lambda$, where λ is the second largest eigenvalue of $D_{\text{deg}}^{-1}A$. Putting this together with Lemma 12, we conclude that

$$\lambda \leq \text{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon})$$

as desired.

G Numerical Experiments

Here we provide more details on the numerical experiments. All models were implemented with PyTorch [29] and PyTorch Geometric [30].

Datasets. We used `torch_geometric.datasets.planetoid` provided in PyTorch Geometric for all the three datasets: Cora, CiteSeer, and PubMed with their default training and test splits.

Model details.

- For GAT, we consider the architecture proposed in Veličković et al. [7] with each attentional layer sharing the parameter a in $\text{LeakyReLU}(a^\top[W^\top X_i || W^\top X_j]), a \in \mathbb{R}^{2d'}$ to compute the attention scores.
- For GCN, we consider the standard random walk graph convolution $D_{\text{deg}}^{-1}A$. That is, the update rule of each graph convolutional layer can be written as

  $$X' = D_{\text{deg}}^{-1}AXW\,,$$

  where X and X' are the input and output node representations, respectively, and W is the shared learnable weight matrix in the layer.

Compute. We trained all of our models on a Telsa V100 GPU.

Training details. In all experiments, we used the Adam optimizer using a learning rate of 0.00001 and 0.0005 weight decay and trained for 1000 epoch.