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 $1 \leq p, q \leq \infty$, 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 Proposition 1

It is straightforward to check that $\|X - \mathbf{1}\gamma_X\|_F$ satisfies the two axioms of a node similarity measure:

1. $\|X - \mathbf{1}\gamma_X\|_F = 0 \Longleftrightarrow X = \mathbf{1}\gamma_X \Longleftrightarrow X_i = \gamma_X$ for all node i.
2. Let $\gamma_X = \frac{\mathbf{1}^\top X}{N}$ and $\gamma_Y = \frac{\mathbf{1}^\top Y}{N}$, then $\gamma_X + \gamma_Y = \frac{\mathbf{1}^\top (X+Y)}{N} = \gamma_{X+Y}$. So

$$\begin{aligned}\mu(X+Y) &= \|(X+Y) - \mathbf{1}(\gamma_X+\gamma_Y)\|_F = \|X - \mathbf{1}\gamma_X + Y - \mathbf{1}\gamma_Y\|_F\\ &\le \|X-\mathbf{1}\gamma_X\|_F + \|Y - \mathbf{1}\gamma_Y\|_F\\ &= \mu(X)+\mu(Y)\,.\end{aligned}$$

C Proof of Lemma 1

According to the formulation (6):

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

we thus obtain that

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

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

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

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

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

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

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

D Proof of Lemma 2

Lemma 2 is a direct corollary of Lemma 1 and the assumption that $\Psi(\cdot,\cdot)$ assigns bounded attention scores to bounded inputs.

E Proof of Lemma 3

E.1 Auxiliary results

We make use of the following sufficient condition for the ergodicity of the infinite products of row-stochastic matrices.

Lemma 7 (Corollary 5.1 [2])**.** *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 $t_0 \le t_1$, we denote

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

Lemma 8. *Under the assumption* **A1***, there exist $T\in\mathbb{N}$ and $c>0$ such that for all $t_0\ge 0$,*

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

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

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

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

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 [4]: 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 $t\ge 0$ and hence for $t\ge r(i,j)$,

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

For $t\ge t(i):=\max_{j\in\mathcal{G}} r(i,j)$, we have $P_{ij}^{(t:0)}>0$ for all node j in $\mathcal{G}$. Finally, if $t\ge T:=\max_{i\in\mathcal{G}} t(i)$, 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 $t\in\mathbb{N}_{\ge 0}$, $P_{ij}^{(t)}\ge\epsilon$ if $(i,j)\in E(\mathcal{G})$, we conclude that $P_{ij}^{(T:0)}\ge\epsilon^T := c$. $\square$

E.2 Proof of Lemma 3

Given the sequence $\{P^{(t)}\}_{t=0}^{\infty}$, we use $T\in\mathbb{N}$ from Lemma 8 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 8, for all $k\in\mathbb{N}_{\ge 0}$, there exists $c>0$ such that $c\le \bar{P}_{ij}^{(k)}\le 1\,, \forall 1\le i,j\le N$. Then Lemma 3 is a direct consequence of Lemma 7.

F Proof of Lemma 5

F.1 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 9. *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. 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 8, 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 $c(1 - D^{(t_1)}_{\min}) = c\delta_{t_1}$, where $D^{(t_1)}_{\min}$ here denotes the smallest diagonal entry of the diagonal matrix $D^{(t_1)}$. Hence,

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

□

F.2 Proof of Lemma 4

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 4 in the main paper and from which the ergodicity of any sequence in $\mathcal{M}_{\mathcal{G},\epsilon}$ is an immediate result.

Lemma 4. 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. 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 9, 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

$$\prod_{t=m}^{\infty} (1 - c\delta_t) > 1 - \eta . \tag{8}$$

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 (8) 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

$$\prod_{t=m}^{\infty} (1 - \delta_t) \ge \prod_{t=m}^{\infty} e^{-b\delta_t} > (1 - \eta)^{\frac{b}{c}} := 1 - \eta' . \tag{9}$$

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 (9) leads to

$$(1 - \eta') P^{(M:m)} \le_{\text{ew}} Q_{m,M} \le_{\text{ew}} P^{(M:m)} . \tag{10}$$

Therefore,

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

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

$$\|BQ_{0,M}\|_\infty \le \|BQ_{m,M}\|_\infty \|Q_{0,m-1}\|_\infty \le \|BQ_{m,M}\|_\infty \le \eta' \sqrt{N} . \tag{11}$$

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

The above inequality (11) holds when taking $M \to \infty$. Then taking $\eta \to 0$ implies $\eta' \to 0$ and together with (11), we conclude that

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

and therefore,

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

□

F.3 Proof of Lemma 5

Notice that both cases $\beta = 0$ and $\beta > 0$ in Lemma 4 imply the ergodicity of $\{D^{(t)}P^{(t)}\}_{t=0}^\infty$. Hence the statement is a direct corollary of Lemma 4.

G Proof of Lemma 6

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. 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

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

□

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 [6])**.** *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 5, Lemma 10 and Lemma 11, we conclude that $\mathrm{JSR}(\tilde{\mathcal{M}}_{\mathcal{G},\epsilon}) < 1$.

H Proof of Theorem 1

Recall the formulation of $X_{\cdot i}^{(t+1)}$ in (6):

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

Then it follows that

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

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 (7), 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

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

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 $t \in \mathbb{N}_{\geq 0}$,

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

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

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

I Proof of Proposition 2

Since $D_{\deg}^{-1}A$ is similar to $D_{\deg}^{-1/2}AD_{\deg}^{-1/2}$, they have the same spectrum. For $D_{\deg}^{-1}A$, 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 $D_{\deg}^{-1}A \in \mathcal{P}_{\mathcal{G},\epsilon}$.

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*

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

Proof. 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 [3]. Then comparing with the definition of the joint spectral radius, we can immediately conclude the statement. □

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

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

as desired.

J Numerical Experiments

Here we provide more details on the numerical experiments presented in Section 5. All models were implemented with PyTorch [5] and PyTorch Geometric [1] .

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

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

  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.

References

[1] Matthias Fey and Jan E. Lenssen. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[2] Darald J. Hartfiel. *Nonhomogeneous Matrix Products*. 2002.

[3] Peter D. Lax. *Functional Analysis*. 2002.

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

[5] 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.

[6] Jacques Theys. Joint Spectral Radius: theory and approximations. *Ph. D. dissertation*, 2005.

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *ICLR*, 2018.