EXPLORING OVER-SMOOTHING IN GRAPH ATTENTION NETWORKS FROM THE MARKOV CHAIN PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

The over-smoothing problem causing the depth limitation is an obstacle of developing deep graph neural network (GNN). Compared with Graph Convolutional Networks (GCN), over-smoothing in Graph Attention Network (GAT) has not drawed enough attention. In this work, we analyze the over-smoothing problem in GAT from the Markov chain perspective. First we establish a connection between GAT and a time-inhomogeneous random walk on the graph. Then we show that the GAT is not always over-smoothing using conclusions in the time-inhomogeneous Markov chain. Finally, we derive a sufficient condition for GAT to avoid over-smoothing based on our findings about the existence of the limiting distribution of the time-inhomogeneous Markov chain. We design experiments to verify our theoretical findings. Results show that our proposed sufficient condition can effectively improve over-smoothing problem in GAT and enhance the performance of the model.

1 INTRODUCTION

Graph neural networks (Kipf & Welling, 2017; Bruna et al., 2013; Defferrard et al., 2016; Veličković et al., 2018) have achieved great success in processing graph data which is rich in information about the relationships between objects. GAT (Veličković et al., 2018) is one of the most representative GNN models. It introduces the attention mechanism into GNN and inspires a class of attention-based GNN models (Abu-El-Haija et al., 2018; Zhang et al., 2018; Lee et al., 2018).

The deepening of the network has brought about changes in neural networks and caused a boom in deep learning. Unlike typical deep neural networks, in the training of graph neural networks, researchers have found that the performance of GNN decreases instead as the depth increases. There are several possible reasons for the depth limitations of GNN. Li et al. (2018) first attribute this anomaly to *over-smoothing*, a phenomenon in which the representations of different nodes tend to be consistent as the network deepens, leading to indistinguishable node representations. Many researchers have studied this problem and proposed some improvement methods (Li et al., 2018; Oono & Suzuki, 2020; Rong et al., 2020; Chen et al., 2020b;a; Zhao & Akoglu, 2019; Chiang et al., 2019). Over-smoothing has been observed not only in GNN, but also in other fields (Shi et al., 2022; Zhou et al., 2020). Studies of the over-smoothing problem in GNNs have also inspired these researches. However, the research of the over-smoothing problem has mostly focused on graph convolutional network. There is a lack of unique analysis of over-smoothing in GAT and corresponding improvement methods.

Noting the Markov property of the forward propagation process of GNNs and considering the node set as a state space, in this work, we connect GAT with a time-inhomogeneous random walk on the graph. Considering the nodes' representations as distributions on the state space, we interpret the over-smoothing in GAT as the convergence of the representation distribution to the limiting distribution. Using conclusions of Bowerman et al. (1977) and Huang et al. (1976) in the time-inhomogeneous Markov chain, we show that GAT does not necessarily suffer from over-smoothing in the Markovian sense. Further, we prove a necessary condition for the existence of the limiting distribution. Based on this conclusion, we derive a sufficient condition for GAT to avoid over-smoothing.

We verify our conclusions on the benchmark datasets. Based on the sufficient condition, we propose a regularization term which can be flexibly added to the training of the neural network. Results show

that our proposed sufficient condition can significantly improve the performance of GAT. In addition, the representation learned by different nodes is more inconsistent after adding the regularization term, which indicates that the over-smoothing in GAT is improved.

Contributions. In summary, our contributions are as follows:

- We establish a connection between GAT and a time-inhomogeneous random walk on the graph. Then we show that GAT is not always over-smoothing in the Markovian sense (Section 3).
- We study the existence of limiting distributions of the time-inhomogeneous Markov chain. And based on this, we give a sufficient condition for GAT to avoid over-smoothing (Section 4 and Theorem 6, 7).
- We propose a regularization term based on this sufficient condition, which can be simply and flexibly added to the training of GAT, and experimentally verify that our proposed condition can improve the model performance by solving the over-smoothing problem of GAT (Section 5).

Notation. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a connected non-bipartite graph, where $\mathcal{V} := \{1, 2, \ldots, N\}$ is the node set, $\mathcal{E}$ is the edge set, $N = |\mathcal{V}|$ is the number of nodes. If there are connected edges between nodes $u, v \in \mathcal{V}$, then denote by $(u, v) \in \mathcal{E}$. $\deg(u)$ denotes the degree of node $u \in \mathcal{V}$ and $\mathcal{N}(v)$ denotes the neighbors of node v. The corresponding adjacency matrix is $\boldsymbol{A}$ and the degree matrix is $\boldsymbol{D}$. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space, $(\mathcal{M}, \mathscr{M})$ be a finite state space. $\vec{\mathbf{x}} = \{\mathbf{x}_n, n \in \mathcal{T}\}$ is a stochastic process taking values in $\mathcal{M}$. $\mathcal{T}$ is a time parameter set. $P(i, j)$ denotes the element i, j of matrix $\boldsymbol{P}$.

2 Related work

In this section, we introduce graph attention network and the related work of over-smoothing.

2.1 Graph Attention Network

GAT (Veličković et al., 2018) establishes attention functions between nodes u and v with connected edges $(u, v) \in \mathcal{E}$

$$\alpha_{u,v}^{(l)} = \frac{\exp(\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_v^{(l-1)}))}{\sum_{k \in \mathcal{N}(u)} \exp(\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_k^{(l-1)}))} \tag{1}$$

where $\boldsymbol{h}_u^{(l)} \in \mathbb{R}^F$ is the embedding for node u at the layer l and

$$\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_v^{(l-1)}) := \text{LeakyReLU}(\mathbf{a}^{\mathrm{T}}[\boldsymbol{W}^{(l)}\boldsymbol{h}_u^{(l-1)} \| \boldsymbol{W}^{(l)}\boldsymbol{h}_v^{(l-1)}]),$$

where $\mathbf{a} \in \mathbb{R}^{2F}$ and $\boldsymbol{W}^{(l)}$ is the weight matrix. Then the GAT layer is defined as

$$\boldsymbol{h}_u^{(l)} := \sigma_{\boldsymbol{W}^{(l)}} \left(\sum_{v \in \mathcal{N}(u)} \alpha_{u,v}^{(l)} \boldsymbol{h}_v^{(l-1)} \right).$$

Written in matrix form

$$\boldsymbol{H}^{(l)} = \sigma_{\boldsymbol{W}^{(l)}}(\boldsymbol{P}_{\text{att}}^{(l)} \boldsymbol{H}^{(l-1)}),$$

where σ is the activation function, $\boldsymbol{P}_{\text{att}}^{(l)} \in \mathbb{R}^{N \times N}$ is the attention matrix satisfying $P_{\text{att}}^{(l)}(u, v) = \alpha_{u,v}^{(l)}$ if $v \in \mathcal{N}(u)$, otherwise $P_{\text{att}}^{(l)}(u, v) = 0$ and $\sum_{v=1}^{N} P_{\text{att}}^{(l)}(u, v) = 1$.

2.2 Over-smoothing

There is a phenomenon that the GNN has better experimental results in the shallow layer case, and instead do not work well in the deep layer case. The researchers find that this is due to the fact that during the GNN training process, the hidden layer representation of each node tend to converge to the same value as the number of layers increases. This phenomenon is called over-smoothing. This problem affects the deepening of GNN layers and limits the further development of GNN.

Intuitively, Zhao & Akoglu (2019) proposes a normalization layer, Pairnorm, to avoid node representations from becoming too similar. Thus, the over-smoothing phenomenon is alleviated.

Another intuitive analysis of the over-smoothing is that as the network is stacked, the model forgets the initial input features and only updates the representations based on the structure of the graph data. It is natural to think that the problem of the model forgetting the initial features can be alleviated by reminding the network what its previous features are. Many methods have been proposed based on such intuitive analysis. The simplest one, Kipf & Welling (2017) propose to add residual connections to graph convolutional networks. The node representations of the hidden layer l are directly added to the node representations of the previous layer to remind the network not to forget the previous features. However, Chiang et al. (2019) argues that residual connectivity ignores the structure of the graph and should be considered to reflect more the influence of the weights of different neighboring nodes. So this work gives more weight to the representations from the previous layer in the message passing of each GCN layer by improving the graph convolution operator. Chen et al. (2020b); Li et al. (2019); Xu et al. (2018)also use this idea.

Oono & Suzuki (2020) connects the GCN with a dynamical system and interprets the over-smoothing problem as the convergence of the dynamical system to an invariant subspace. Rong et al. (2020) proposed DropEdge method based on the perspective of dynamical system. The idea of DropEdge is to randomly drop some edges in the original graph at each layer. This operation slows down the convergence of the dynamical system to the invariant subspace. Thus DropEdge method can alleviate the over-smoothing. Shi et al. (2022) also follows the idea of Oono & Suzuki (2020) and investigates the over-smoothing problem in BERT.

Most of the works on over-smoothing focus on GCN and ignore the discussion of GAT. Wang et al. (2019) first analyze the over-smoothing problem in GAT and improve GAT via margin-based constraints. However, we disagree with their conclusion that GAT will be over-smoothing. We discuss this in detail in Section 3.

3 Analysis of Over-smoothing in GAT

In this section, we analyze the over-smoothing problem in GAT from the Markov chain perspective. We show that forward propagation of GAT is a time-inhomogeneous random walk $\vec{\mathbf{v}}_{\text{att}}$ on the graph, and that over-smoothing is caused by the convergence of the representation distribution to the limiting distribution. Next, we show that GAT is not always oversmooth by analyzing that the limiting distribution of $\vec{\mathbf{v}}_{\text{att}}$ does not always exist.

3.1 Relationship between GAT and time-inhomogeneous random walk

We first connect GAT with a time-inhomogeneous random walk on the graph. The following defines the general random walk on the graph.

Definition 1 *Given a graph $\mathcal{G}$ and a starting node $u \in \mathcal{V}$, we select a neighbor $v \in \mathcal{N}(u)$ of it with positive probability $P^{(1)}(u,v) > 0$, and move to its neighbor. $P^{(1)}(u,v)$ satisfies $\sum_{v\in\mathcal{N}(u)} P^{(1)}(u,v) = 1$. Then we repeat this process. $P^{(t)}(u,v),\ t = 1,2,\ldots$ is not always the same. The random sequence of nodes selected this way is a* random walk on the graph.

Recalling the definition of GAT, we focus on its message passing $\boldsymbol{H}^{(l)} = \boldsymbol{P}_{\text{att}}^{(l)}\boldsymbol{H}^{(l-1)}$[1]. Since $P_{\text{att}}^{(l)}(u,v) \geq 0$ and $\sum_{v=1}^{N} P_{\text{att}}^{(l)}(u,v) = 1$, $\boldsymbol{P}_{\text{att}}^{(l)}$ is a one-step stochastic matrix of a random walk on the graph. Moreover, since $\boldsymbol{P}_{\text{att}}^{(l)},\ l = 1,2,\ldots$ is not the same, the forward propagation process of node representations in GAT is a time-inhomogeneous random walk on a graph, denoted as $\vec{\mathbf{v}}_{\text{att}}$. It has the state space $\mathcal{V}$ and the family of stochastic matrices $\left\{\boldsymbol{P}_{\text{att}}^{(1)}, \boldsymbol{P}_{\text{att}}^{(2)}, \ldots, \boldsymbol{P}_{\text{att}}^{(l)}, \ldots\right\}$. The inconsistency of the nodes message passing at each layer in GAT causes the time-inhomogeneousness of the corresponding chain $\vec{\mathbf{v}}_{\text{att}}$, which is an important difference between GAT and GNNs with consistent message passing such as GCN.

[1]Similar to Li et al. (2018); Wang et al. (2019), we omit the activation function. In fact, according to the spectral analysis of Wu et al. (2019) and the experimental results, the GNN omitted nonlinear activation function is not different from the GNN added activation function in terms of performance.

Following is the general definition of the limiting distribution. We use it to explain the over-smoothing problem.

Definition 2 *Let $\vec{\mathbf{x}} = \{\mathbf{x}_n, n \in \mathcal{T}\}$ be a time-inhomogeneous Markov chain on a finite state space $\mathcal{M}$, the initial distribution be μ_0 and the distribution of the chain $\vec{\mathbf{x}}$ at moment n be μ_n. π is the* limiting distribution *of the chain $\vec{\mathbf{x}}$, if $\mu_n \to \pi,\ n \to \infty$.*

The node representation $\boldsymbol{h} = \{\boldsymbol{h}_u, u \in \mathcal{V}\}$ is viewed as a discrete probability distribution over the node set $\mathcal{V}$. If $\vec{\mathbf{v}}_{\text{att}}$ has a limiting distribution π, then as the GAT propagates forward, the representation distribution converges to the limiting distribution. This causes the potential over-smoothing problem in GAT. However, the limiting distribution of the time-inhomogeneous Markov chain does not always exist. We next discuss specifically the possibility of GAT to avoid over-smoothing in Markovian sense.

3.2 GAT IS NOT ALWAYS OVER-SMOOTHING

In this subsection, we first show that the conclusion that the GAT will be over-smoothing cannot be proven.

The following theorem gives property of the family of stochastic matrices $\left\{\boldsymbol{P}_{\text{att}}^{(1)}, \boldsymbol{P}_{\text{att}}^{(2)}, \ldots, \boldsymbol{P}_{\text{att}}^{(l)}, \ldots\right\}$, shows the existence of stationary distribution of each graph attention matrix, and gives the explicit expression of stationary distribution. The proof is provided in Appendix A.

Theorem 3 *There exists a unique probability distribution $\pi^{(l)}$ on $\mathcal{V}$ satisfies*

$$\pi^{(l)} = \pi^{(l)} \boldsymbol{P}_{att}^{(l)}, \quad l = 1, 2, \ldots,$$

where $\pi^{(l)}(u) = \frac{\deg^{(l)}(u)}{\sum_{k \in \mathcal{V}} \deg^{(l)}(k)}$, $\deg^{(l)}(u) = \sum_{z \in \mathcal{N}(u)} \exp(\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_z^{(l-1)}))$.

Previously, Wang et al. (2019) discussed the over-smoothing problem in GAT and concluded that the GAT would over-smooth. Same as our work, they viewed the $\boldsymbol{P}_{\text{att}}^{(l)}$ at each layer as stochastic matrix of a random walk on the graph. However, they ignore the fact that the complete forward propagation process of GAT is essentially a time-inhomogeneous random walk on the graph. The core theorem stating that the GAT will over-smooth in their work is flawed. In its proof, the stationary distribution $\pi^{(l)}$ of the graph attention matrix $\boldsymbol{P}_{\text{att}}^{(l)}$ for each layer is consistent, i.e.

$$\pi^{(1)} = \pi^{(2)} = \cdots = \pi^{(l)} = \cdots .$$

However, since each layer $\phi^{(l)}$ is different, by Theorem 3,

$$\pi^{(1)} \neq \pi^{(2)} \neq \cdots \neq \pi^{(l)} \neq \cdots .$$

The conclusion that the GAT will be over-smoothing cannot be proven. Then we show that the GAT is not always over-smoothing.

Since over-smoothing in GAT is related to the limiting distribution of time-inhomogeneous random walk on the graph, we next focus on the limiting distribution of $\vec{V}_{\text{att}}$.

Compared to the time-homogeneous Markov chain, it is much more difficult to investigate the limiting distribution of the time-inhomogeneous chain. In order to study the convergence of the probability distribution on the state space, we introduce the Dobrushin contraction coefficient and the Dobrushin inequality. The proof is provided in Appendix A.

Lemma 4 *Let μ and ν be probability distributions on a finite state space $\mathcal{M}$ and $\boldsymbol{P}$ be a stochastic matrix, then*

$$\|\mu \boldsymbol{P} - \nu \boldsymbol{P}\|_1 \leq C(\boldsymbol{P})\ \|\mu - \nu\|_1,$$

where

$$C(\boldsymbol{P}) := \frac{1}{2} \sup_{i,j} \sum_{k \in \mathcal{M}} |P(i,k) - P(j,k)|$$

is called the Dobrushin contraction coefficient *of the stochastic matrix $\boldsymbol{P}$.*

Bowerman et al. (1977); Huang et al. (1976) discussed the limiting case that an arbitrary initial distribution transfer according to a time-inhomogeneous chain. The sufficient condition for the existence of the limiting distribution is summarized in the following lemma. The proof is provided in Appendix A.

Lemma 5 *Let $\vec{\mathbf{x}} = \{\mathbf{x}_n, n \in \mathcal{T}\}$ be a time-inhomogeneous Markov chain on a finite state space $\mathcal{M}$ with stochastic matrix $\boldsymbol{P}^{(n)}$. If the following $(1), (2)$ and $(3A)$ or $(3B)$ are satisfied*

(1) There exists a stationary distribution $\pi^{(n)}$ when $\boldsymbol{P}^{(n)}$ is treated as a stochastic matrix of a time-homogeneous chain;

(2) $\sum_n \|\pi^{(n)} - \pi^{(n+1)}\|_1 < \infty$;

(3A) (Isaacson-Madsen condition) For any probability distribution μ, ν on $\mathcal{M}$ and positive integer k

$$\|(\mu - \nu)\boldsymbol{P}^{(k)} \cdots \boldsymbol{P}^{(n)}\|_1 \to 0, \quad n \to \infty.$$

(3B) (Dobrushin condition) For any positive integers k

$$C(\boldsymbol{P}^{(k)} \cdots \boldsymbol{P}^{(n)}) \to 0, \quad n \to \infty.$$

Then there exists a probability measure π on $\mathcal{M}$ such that

(1) $\|\pi^{(n)} - \pi\|_1 \to 0, \quad n \to \infty$;

(2) Let the initial distribution be μ_0 and the distribution of the chain $\vec{\mathbf{x}}$ at step n be $\mu_n := \mu_{n-1}\boldsymbol{P}^{(n)}$, then for any initial distribution μ_0, we have

$$\|\mu_n - \pi\|_1 \to 0, \quad n \to \infty,$$

Returning to GAT, for time-inhomogeneous random walk $\vec{\mathbf{v}}_{\text{att}}$, its family of stochastic matrices $\{\boldsymbol{P}_{\text{att}}^{(l)}\}$ satisfies the condition (1) (Theorem 3). However, the series of positive terms $\sum_l \|\pi^{(l)} - \pi^{(l+1)}\|$ is possible to be divergent and the condition (2) of Lemma 5 can not be guaranteed. Moreover, according to the definition of $\{\boldsymbol{P}_{\text{att}}^{(l)}\}$ (Equation 1), neither the Isaacson-Madsen condition nor the Dobrushin condition can be guaranteed. So the time-inhomogeneous chain $\vec{\mathbf{v}}_{\text{att}}$ does not always have a limiting distribution. This indicates that GAT is not always over-smoothing.

4 SUFFICIENT CONDITION FOR GAT TO AVOID OVER-SMOOTHING

In this section, we propose and prove a necessary condition for the existence of limiting distribution for a time-inhomogeneous Markov chain. Then we apply this theoretical result to GAT and propose a sufficient condition to ensure that GAT can avoid over-smoothing.

In the study of Markov chains, researchers usually focus on the sufficient conditions for the existence of the limiting distribution. And the case when the limiting distribution does not exist has rarely been studied. We study the necessary conditions for the existence of the limit distribution in order to obtain sufficient conditions for its nonexistence.

The following theorem gives a necessary condition for the existence of the limit distribution of the time-inhomogeneous Markov chain. Although other necessary conditions exist, Theorem 6 is one of the most intuitive and simplest in form. The proof is provided in Appendix A.

Theorem 6 *Let $\vec{\mathbf{x}} = \{\mathbf{x}_n, n \in \mathcal{T}\}$ be a time-inhomogeneous Markov chain on a finite state space $\mathcal{M}$, and write its n-step stochastic matrix as $\boldsymbol{P}^{(n)}$, satisfying that, $\boldsymbol{P}^{(n)}$ is irreducible and aperiodic, there exists a unique stationary distribution $\pi^{(n)}$ as the time-homogeneous transition matrix, and $C(\boldsymbol{P}^{(n)}) < 1$. Let the initial distribution be μ_0 and the distribution of the chain $\vec{\mathbf{x}}$ at step n be $\mu_n := \mu_{n-1}\boldsymbol{P}^{(n)}$. Then*

$$\|\pi^{(n)} - \pi\| \to 0, \quad n \to \infty$$

is a necessary condition for existence of a probability distribution π on $\mathcal{M}$ such that $\|\mu_n - \pi\| \to 0, \ n \to \infty$.

We explain Theorem 6 intuitively. In the limit sense, transition of μ_{n-1} satisfies

$$\lim_{n\to\infty} \mu_{n-1}\boldsymbol{P}^{(n)} = \lim_{n\to\infty} \mu_n = \lim_{n\to\infty} \mu_{n-1} = \pi.$$

On the other hand, for all $n > 0$, $\pi^{(n)}$ is the unique solution of the equation

$$\mu = \mu\boldsymbol{P}^{(n)}.$$

Thus

$$\lim_{n\to\infty} \pi^{(n)} = \lim_{n\to\infty} \mu_{n-1} = \pi.$$

By Theorem 6, we give the following sufficient condition for GAT to avoid over-smoothing in Markovian sense. The proof is provided in Appendix A.

Theorem 7 *Let $\boldsymbol{h}_u^{(l)}$ be representation of node $u \in \mathcal{V}$ at the hidden layer l in GAT. The sufficient condition for GAT to avoid over-smoothing is that there exists $\delta > 0$ such that for any $l \geq 2$, satisfying*

$$\|\boldsymbol{h}_u^{(l-1)} - \boldsymbol{h}_u^{(l)}\|_1 > \delta. \tag{2}$$

When Equation 2 is satisfied, the time-inhomogeneous random walk $\vec{\mathbf{x}}_{\text{att}}$ corresponding to GAT does not have a limiting distribution, and thus GAT avoids potential over-smoothing problems in a Markovian sense. Theorem 7 has an intuitive meaning. The essence of over-smoothing is that the node representations converge with the propagation of the network. By Cauchy's convergence test, the condition exactly avoid representation $\boldsymbol{h}_u^{(l)}$ of the node u from converging as network deepens.

Since Theorem 6 generally holds for all time-inhomogeneous Markov chains, GNN related to a time-inhomogeneous Markov chain such as GEN (Li et al., 2020) can obtain the sufficient conditions similar to Theorem 7 to avoid over-smoothing in Markovian sense. The conclusion we obtained about the existence of the limiting distribution is general. The analysis of GEN is provided in Appendix B.

Considering that this sufficient condition is task-agnostic, we can formulate this condition to a regularization term and add it to the loss function determined by its original task. Formally, assume the original loss function is $L_\theta(x)$, the total loss function can be rewritten as:

$$\hat{L}_\theta(x) = L_\theta(x) + \text{RT}_\theta(x) \tag{3}$$

where θ is the parameter of neural networks and $\text{RT}_\theta(x)$ is the regularization term determined by the sufficient condition in Theorem 7.

5 EXPERIMENTS

In this section, we experimentally verify the correctness of our theoretical results. We rewrite the sufficient condition for GAT to avoid oversmoothing in the Theorem 7 as a regularization term. It can be flexibly added to the training of the network. The experimental results show that our proposed condition can effectively avoid the over-smoothing problem and improve the performance of GAT. We also conduct experiments on GEN-SoftMax (Li et al., 2020) (Section 5.4) and we leave the theoretical details of GEN in Appendix B.

5.1 SETUP

In this section we briefly introduce the experimental settings. See Appendix C for more specific settings. We verify our conclusions while keeping the other hyperparameters the same[2] (network structure, learning rate, dropout, epoch, etc.).

Variant of sufficient condition. Notice that the sufficient condition in the Theorem 7 is in the form of inequality, which is not conducive to experiments. In the concrete implementation, let $\boldsymbol{h}_u^{(l)}$ be representation of node $u \in \mathcal{V}$ at the hidden layer l. We normalize the distance of the node

[2]Since we do not aim to refresh State of the Arts, these are not necessarily the optimal hyperparameters.

Table 1: Results of GAT

#layers	model	datasets			
		Cora	Citeseer	Pubmed	ogbn-arxiv
3	GAT	0.7773(±0.0054)	0.6643(±0.0063)	0.7616(±0.0115)	0.7117(±0.0023)
	GAT-RT	0.7884(±0.0157)	0.6678(±0.0157)	0.7673(±0.0064)	0.7115(±0.0025)
4	GAT	0.7602(±0.0166)	0.6541(±0.0076)	0.7534(±0.0114)	0.7144(±0.0015)
	GAT-RT	0.7872(±0.0127)	0.6692(±0.0072)	0.7659(±0.0123)	0.7104(±0.0022)
5	GAT	0.4821(±0.3021)	0.3472(±0.2582)	0.7653(±0.0072)	0.7061(±0.0082)
	GAT-RT	0.7648(±0.0077)	0.6208(±0.0380)	0.7684(±0.0063)	0.7063(±0.0040)
6	GAT	0.2774(±0.2542)	0.2474(±0.1947)	0.7468(±0.0084)	0.6396(±0.1031)
	GAT-RT	0.6454(±0.2508)	0.2706(±0.1884)	0.7664(±0.0092)	0.6709(±0.0344)
7	GAT	0.1672(±0.0780)	0.1768(±0.0064)	0.7468(±0.0045)	0.4307(±0.1720)
	GAT-RT	0.3244(±0.2465)	0.1915(±0.0200)	0.7596(±0.0107)	0.5653(±0.1122)
8	GAT	0.0958(±0.0059)	0.1902(±0.0598)	0.7076(±0.0112)	-
	GAT-RT	0.1678(±0.0756)	0.1864(±0.0229)	0.7618(±0.0114)	-

representations between two adjacent layers and then let it approximate to a given hyperparameter threshold $T \in (0, 1)$, i.e., for the GNN model with n layers, we obtain a regularization term:

$$\mathrm{RT}_\theta(x) = \left(\frac{1}{n} \sum_{l=1}^{n} (\| \ \mathrm{Sigmoid}(\boldsymbol{h}_u^{(l-1)}) - \mathrm{Sigmoid}(\boldsymbol{h}_u^{(l)}) \ \|) - T \right)^2 . \tag{4}$$

Since there must exist $\delta > 0$ that satisfies $T > \delta$, Equation 2 can be satisfied if this term is perfectly minimized. For a detailed choice of the threshold T, we put it in Appendix C.

Datasets. In terms of datasets, we follow the datasets used in the original work of GAT (Veličković et al., 2018) as well as the OGB benchmark. We use four standard benchmark datasets - ogbn-arxiv (Hu et al., 2020), Cora, Citeseer, and Pubmed (Sen et al., 2008), covering the basic transductive learning tasks.

Implementation details. For the specific implementation, we refer to the open-source code of vanilla GAT, and models with different layers share the same settings: We use the Adam SGD optimizer (Kingma & Ba, 2015) with learning rate 0.01, the hidden dimension is 64, each GAT layer has 8 heads and the amount of training epoch is 500. All experiments are conducted on a single Nvidia Tesla v100.

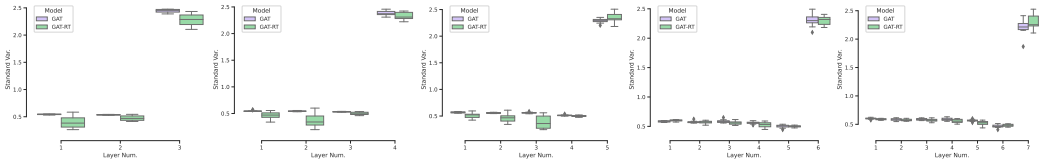
5.2 Results of GAT

For record simplicity, we denote the GAT after adding the regularization term to the training as GAT-RT. To keep statistical confidence, we repeat all experiments 10 times and record the mean value and standard deviation. Results shown in Table 1 demonstrate that almost on each dataset and number of layers, GAT-RT will obtain an improvement in the performance. Specifically, on Cora and Citeseer, GAT's performance begins to decrease drastically when layer numbers surpass 6 and 5 but GAT-RT can relieve this trend in some way. On Pubmed, vanilla GAT's performance has a gradual decline. The performance of vanilla GAT decrease 6% when the layer number is 8. However, GAT-RT's performance keeps competitive for all layer numbers. For ogbn-arxiv, GAT-RT performs as competitive as GAT when the layer number is small but outperforms GAT by a big margin when the layer number is large. Specifically when the layer number is 6 and 7, the performance improves roughly by 3% and 13% respectivly.

5.3 Verification of avoiding over-smoothing

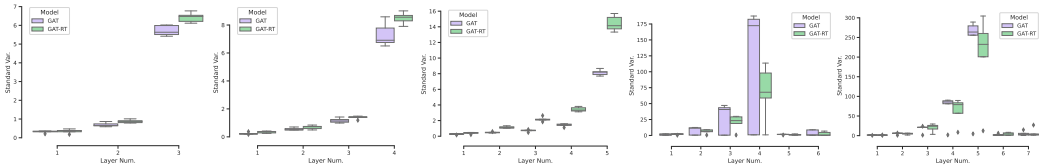
In this subsection, we further experimentally show that the sufficient conditions in Section 4 not only improve the performance of the model but also do avoid the over-smoothing problem.

Since the neural network is a black-box model, we cannot explicitly compute the stationary distribution of the graph neural network when it is over-smoothed. Therefore we measure the degree of over-smoothing by calculating the standard deviation of each node's representation at each layer. A lower value implies more severe over-smoothing.



(a) 3-layer (b) 4-layer (c) 5-layer (d) 6-layer (e) 7-layer

Figure 1: Measurement of over-smoothing of GAT on ogbn-arxiv.



(a) 3-layer (b) 4-layer (c) 5-layer (d) 6-layer (e) 7-layer

Figure 2: Measurement of over-smoothing of GAT on Cora.

Results shown in Fig. 1-4 demonstrate that the node representations obtained from GAT-RT are more diverse than those from GAT, which means the alleviation of over-smoothing. It's also interesting that there is an accordance between the performance and over-smoothing, for example on Cora dataset, the performance would have a huge decrease when the number of layers is larger than 5, Fig. 2 shows the over-smoothing phenomenon is severe at the same time. Also on Pubmed dataset, the performance is relatively stable and the corresponding Fig. 4 shows that the model trained on this dataset suffers from over-smoothing lightly. These results enlighten us that over-smoothing may be caused by various objective reasons, e.g. the property of the dataset, and GAT-RT can relieve this negative effect to some extent.

5.4 Extension to GEN

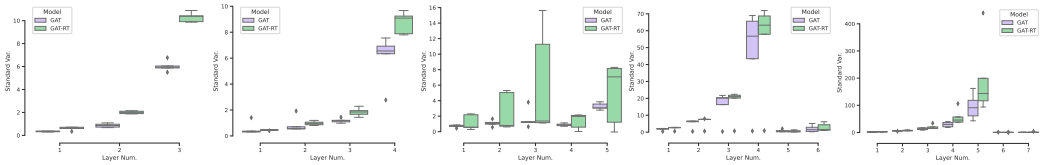
Because GEN (Li et al., 2020) shares the same time-inhomogeneousness property compared with GAT, we can obtain the similar sufficient condition (Theorem 8) using Theorem 6. See Appendix B for a detailed analysis. We conduct experiments of GEN on OGB (Hu et al., 2020) dataset. The detailed experimental setup is shown in Appendix C. In Table. 2, results show that there is a significant improvement in each dataset and each layer compared with the original model when adding our proposed regularization term. Due to the various tricks during the implementation of GEN such as residual connections, which may relieve over-smoothing, the difference in the degree of over-smoothing between finite layers GEN and GEN-RT is not significant enough. So we don't demonstrate the degree of over-smoothing here.

Table 2: GEN's performance on OGB datasets

datasets	model	#layers			
		7	14	28	56
ogbn-arxiv	GEN	0.7140(±0.0003)	0.7198(±0.0007)	0.7192(± 0.0016)	-
	GEN-RT	0.7181(±0.0006)	0.7204(±0.0014)	0.7220(±0.0008)	-
ogbg-molhiv	GEN	0.7858(±0.0117)	0.7757(±0.0019)	0.7641(±0.0058)	0.7696(±0.0075)
	GEN-RT	0.7872(±0.0083)	0.7838(±0.0024)	0.7835(±0.0010)	0.7795(±0.0027)
ogbg-ppa	GEN	0.7554(±0.0073)	0.7631(±0.0065)	0.7712(±0.0071)	-
	GEN-RT	0.7600(±0.0062)	0.7700(±0.0081)	0.7800(±0.0037)	-

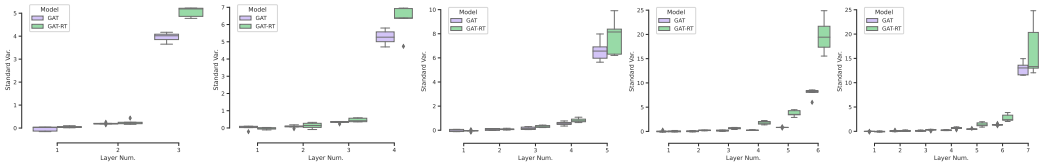
6 Conclusion

In this work, we analyze the over-smoothing problem in GAT from a Markov chain perspective. First we relate GAT to a time-inhomogeneous random walk on the graph. By analyzing the limiting dis-



(a) 3-layer (b) 4-layer (c) 5-layer (d) 6-layer (e) 7-layer

Figure 3: Measurement of over-smoothing of GAT on Citeseer.



(a) 3-layer (b) 4-layer (c) 5-layer (d) 6-layer (e) 7-layer

Figure 4: Measurement of over-smoothing of GAT on Pubmed.

tribution of this random walk, we show that it is possible for GAT to avoid potential over-smoothing. Then, we study the limiting distribution of the general time-inhomogeneous Markov chain, and propose a necessary condition for the existence of the limiting distribution. Based on this result, we derive a sufficient condition for GAT to avoid over-smoothing in the Markovian sense. Our results can also be generalized to other GNN models related to time-inhomogeneous Markov chains. Finally, in our experiments we design a regularization term which can be flexibly added to the training. Results on the benchmark datasets show that our theoretical analysis is correct.

REFERENCES

Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alexander A Alemi. Watch your step: Learning node embeddings via graph attention. *Advances in neural information processing systems*, 31, 2018.

Bruce Bowerman, HT David, and Dean Isaacson. The convergence of cesaro averages for certain nonstationary markov chains. *Stochastic processes and their applications*, 5(3):221–230, 1977.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3438–3445, 2020a.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, pp. 1725–1735, 2020b.

Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 257–266, 2019.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

Cheng-Chi Huang, Dean Isaacson, and B Vinograde. The rate of convergence of certain nonhomogeneous markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35(2): 141–146, 1976.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

John Boaz Lee, Ryan Rossi, and Xiangnan Kong. Graph classification using structural attention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1666–1674, 2018.

Guohao Li, Matthias Müller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *International Conference on Computer Vision*, pp. 9266–9275, 2019.

Guohao Li, Chenxin Xiong, Ali Thabet, and Bernard Ghanem. Deepergcn: All you need to train deeper gcns. *arXiv preprint arXiv:2006.07739*, 2020.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI conference on artificial intelligence*, 2018.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*, 2020.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen MS Lee, and James Kwok. Revisiting over-smoothing in bert from the perspective of graph. In *International Conference on Learning Representations*, 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Guangtao Wang, Rex Ying, Jing Huang, and Jure Leskovec. Improving graph attention networks with large margin-based constraints. *arXiv preprint arXiv:1910.11945*, 2019.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International Conference on Machine Learning*, 2019.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pp. 5453–5462, 2018.

Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit Yan Yeung. Gaan: Gated attention networks for learning on large and spatiotemporal graphs. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018.

Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2019.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. Bert loses patience: Fast and robust inference with early exit. *Advances in Neural Information Processing Systems*, 33, 2020.

A PROOFS

Theorem 3 *There exists a unique probability distribution $\pi^{(l)}$ on $\mathcal{V}$ satisfies*

$$\pi^{(l)} = \pi^{(l)} \boldsymbol{P}_{att}^{(l)}, \quad l = 1, 2, \ldots,$$

where $\pi^{(l)}(u) = \frac{\deg^{(l)}(u)}{\sum_{k\in\mathcal{V}} \deg^{(l)}(k)}$, $\deg^{(l)}(u) = \sum_{z\in\mathcal{N}(u)} \exp(\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_z^{(l-1)}))$.

Proof Since $\mathcal{G}$ is a connected graph and the stochastic matrix $\boldsymbol{P}_{\text{att}}^{(l)}$ satisfies $P_{\text{att}}^{(l)}(u, v) > 0,\ (u, v) \in \mathcal{E}$, for any node $u, z \in \mathcal{V}$, there exists $n \in \mathbb{Z}^+$ that satisfies

$$\left(P_{\text{att}}^{(l)}\right)^n (u, z) > 0.$$

Thus $\boldsymbol{P}_{\text{att}}^{(l)}$ is irreducible. We consider the period of any $u \in \mathcal{V}$. Then since $\mathcal{G}$ is a non-bipartite graph, period of u is 1. Thus $\boldsymbol{P}_{\text{att}}^{(l)}$ is aperiodic. Then there exists a unique stationary distribution $\pi^{(l)}$ of $\boldsymbol{P}_{\text{att}}^{(l)}$.

For all $v \in \mathcal{V}$, since

$$\begin{aligned}
\sum_{u\in\mathcal{V}} \pi(u)^{(l)} P_{\text{att}}^{(l)}(u, v) &= \sum_{(u,v)\in\mathcal{E}} \frac{\deg^{(l)}(u)}{\sum_{k\in\mathcal{V}} \deg^{(l)}(k)} \frac{\exp(\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_v^{(l-1)}))}{\sum_{z\in\mathcal{N}(u)} \exp(\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_z^{(l-1)}))} \\
&= \sum_{(u,v)\in\mathcal{E}} \frac{\exp(\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_v^{(l-1)}))}{\sum_{k\in\mathcal{V}} \deg^{(l)}(k)} \\
&= \frac{\deg^{(l)}(v)}{\sum_{k\in\mathcal{V}} \deg^{(l)}(k)} \\
&= \pi^{(l)}(v),
\end{aligned}$$

$\pi^{(l)}$ is a stationary distribution of $\boldsymbol{P}_{\text{att}}^{(l)}$. ■

Lemma 4 *Let μ and ν be probability distributions on a finite state space $\mathcal{M}$ and $\boldsymbol{P}$ be a stochastic matrix, then*

$$\|\mu\boldsymbol{P} - \nu\boldsymbol{P}\|_1 \leq C(\boldsymbol{P})\ \|\mu - \nu\|_1,$$

where

$$C(\boldsymbol{P}) := \frac{1}{2}\ \sup_{i,j}\ \sum_{k\in\mathcal{M}}\ |P(i,k) - P(j,k)|$$

is called the Dobrushin contraction coefficient *of the stochastic matrix $\boldsymbol{P}$.*

Proof Denote the positive and negative parts of a as a^+ and a^- respectively . Then denote

$$\rho_k^+ = (\mu(k) - \nu(k))^+ / \frac{1}{2}\|\mu - \nu\|_1, \quad \rho_k^- = (\mu(k) - \nu(k))^- / \frac{1}{2}\|\mu - \nu\_1|.$$

Since

$$\sum_k [(\mu(k) - \nu(k))^+ - (\mu(k) - \nu(k))^-] = \sum_k (\mu(k) - \nu(k)) = 0,$$

$$\sum_k [(\mu(k) - \nu(k))^+ + (\mu(k) - \nu(k))^-] = \sum_k |\mu(k) - \nu(k)| = \|\mu - \nu\|_1,$$

then

$$\sum_k \rho_k^+ = \sum_k \rho_k^- = \frac{2}{\|\mu - \nu\|_1} \sum_k (\mu(k) - \nu(k))^+ = 1$$

and

$$\mu(k) - \nu(k) = (\mu(k) - \nu(k))^+ - (\mu(k) - \nu(k))^- = \frac{1}{2}(\rho_k^+ - \rho_k^-)\|\mu - \nu\|_1.$$

Thus

$$\begin{aligned}
\|\mu\boldsymbol{P} - \nu\boldsymbol{P}\|_1 &= \sum_k \left| \sum_i P(i,k) \cdot (\mu(i) - \nu(i)) \right| \\
&= \frac{1}{2}\|\mu - \nu\| \sum_k \left| \sum_i P(i,k) \cdot (\rho_i^+ - \rho_i^-) \right| \\
&= \frac{1}{2}\|\mu - \nu\| \sum_k \left| \sum_i \sum_j (P(i,k) - P(j,k)) \cdot (\rho_i^+ \rho_j^-) \right| \\
&\leq \|\mu - \nu\| \sum_i \sum_j (\rho_i^+ \rho_j^-) \sum_k \left(\frac{1}{2}|P(i,k) - P(j,k)| \right) \\
&\leq C(\boldsymbol{P})\|\mu - \nu\|_1.
\end{aligned}$$

■

Lemma 5 *Let $\vec{\mathbf{x}} = \{\mathbf{x}_n, n \in \mathcal{T}\}$ be a time-inhomogeneous Markov chain on a finite state space $\mathcal{M}$ with stochastic matrix $\boldsymbol{P}^{(n)}$. If the following $(1), (2)$ and $(3A)$ or $(3B)$ are satisfied*

- *(1) There exists a stationary distribution $\pi^{(n)}$ when $\boldsymbol{P}^{(n)}$ is treated as a stochastic matrix of a time-homogeneous chain;*
- *(2) $\sum_n \|\pi^{(n)} - \pi^{(n+1)}\|_1 < \infty$;*
- *(3A) (Isaacson-Madsen condition) For any probability distribution μ, ν on $\mathcal{M}$ and positive integer k*

$$\|(\mu - \nu)\boldsymbol{P}^{(k)} \cdots \boldsymbol{P}^{(n)}\|_1 \to 0, \quad n \to \infty.$$

- *(3B) (Dobrushin condition) For any positive integers k*

$$C(\boldsymbol{P}^{(k)} \cdots \boldsymbol{P}^{(n)}) \to 0, \quad n \to \infty.$$

Then there exists a probability measure π on $\mathcal{M}$ such that

- *(1) $\|\pi^{(n)} - \pi\|_1 \to 0, \quad n \to \infty$;*

(2) Let the initial distribution be μ_0 and the distribution of the chain $\vec{\mathbf{x}}$ at step n be $\mu_n := \mu_{n-1}\boldsymbol{P}^{(n)}$, then for any initial distribution μ_0, we have

$$\|\mu_n - \pi\|_1 \to 0, \quad n \to \infty,$$

Proof Let $\mathscr{L}(\mathcal{M})$ be the set of all real-valued functions on $\mathcal{M}$. For $f,\ g \in \mathscr{L}(\mathcal{M})$, define metric

$$d(f,\ g) := \|f - g\|_1.$$

Then $\mathscr{L}(\mathcal{M})$ is the complete metric space.

First of all, from condition (2), using the triangle inequality, $\{\pi^{(n)}\}$ is the Cauchy column in $\mathscr{L}(\mathcal{M})$, so conclusion (1) is correct.

To prove conclusion (2). Using the triangle inequality and the Dobrushin's inequality, we get

$$\begin{aligned}\|\mu_0\boldsymbol{P}^{(1)}\cdots\boldsymbol{P}^{(n)} - \pi\|_1 &\le \|(\mu_0\boldsymbol{P}^{(1)}\cdots\boldsymbol{P}^{(k-1)} - \pi)\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)}\|_1 + \|\pi\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)} - \pi\|_1 \\ &\le 2C(\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)}) + \|\pi\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)} - \pi\|_1.\end{aligned} \tag{A1}$$

The first term of the equation (A1) tends to 0 when the condition (3A) or (3B) is satisfied. Using the stationary distribution condition $\pi^{(n)} = \pi^{(n)}\boldsymbol{P}^{(n)}$ for the second term, we deduce that for $k \ge N$ we have

$$\begin{aligned}\|\pi\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)} - \pi\|_1 &= \|(\pi - \pi^{(k)})\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)} + (\pi^{(k)}\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)} - \pi)\|_1 \\ &= \|(\pi - \pi^{(k)})\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)} + (\pi^{(k)}\boldsymbol{P}^{(k+1)}\cdots\boldsymbol{P}^{(n)} - \pi)\|_1 \\ &= \|(\pi - \pi^{(k)})\boldsymbol{P}^{(k)}\cdots\boldsymbol{P}^{(n)} + \sum_{j=1}^{n-k}(\pi^{(k+j-1)} - \pi^{(k+j)})\boldsymbol{P}^{(k+j)}\cdots\boldsymbol{P}^{(n)} + (\pi^{(n)} - \pi)\|_1 \\ &\ge \sup_{n\ge N}\|(\pi - \pi^{(n)})\|_1 + \sum_{n\ge N}\|(\pi^n - \pi^{(n+1)})\|_1 + \sup_{n\ge N}\|(\pi^{(n)} - \pi)\|_1 \to 0.\end{aligned}$$

So conclusion (2) is correct. ■

Theorem 6 *Let $\vec{\mathbf{x}} = \{\mathbf{x}_n, n \in \mathcal{T}\}$ be a time-inhomogeneous Markov chain on a finite state space $\mathcal{M}$, and write its n-step stochastic matrix as $\boldsymbol{P}^{(n)}$, satisfying that, $\boldsymbol{P}^{(n)}$ is irreducible and aperiodic, there exists a unique stationary distribution $\pi^{(n)}$ as the time-homogeneous transition matrix, and $C(\boldsymbol{P}^{(n)}) < 1$. Let the initial distribution be μ_0 and the distribution of the chain $\vec{\mathbf{x}}$ at step n be $\mu_n := \mu_{n-1}\boldsymbol{P}^{(n)}$. Then*

$$\|\pi^{(n)} - \pi\|_1 \to 0, \quad n \to \infty$$

is a necessary condition for existence of a probability distribution π on $\mathcal{M}$ such that $\|\mu_n - \pi\| \to 0,\ n \to \infty$.

Proof Suppose there exists a probability measure π on $\mathcal{M}$ such that $\|\mu_n - \pi\|_1 \to 0, n \to \infty$. The following conclusion

$$\|\pi^{(n)} - \mu_{n-1}\|_1 \to 0, \quad n \to \infty$$

is proved by contradiction. If for any $N \in \mathbb{N}^+$, there exists $\delta > 0$, when $n > N$, all have

$$\|\pi^{(n)} - \mu_{n-1}\|_1 > \delta.$$

Then by the triangle inequality and the Dobrushin inequality (Lemma4)

$$\begin{aligned}\|\mu_n - \mu_{n-1}\|_1 &= \|(\pi^{(n)} - \mu_{n-1}) - (\pi^{(n)} - \mu_n)\|_1 \\ &\ge \|\pi^{(n)} - \mu_{n-1}\|_1 - \|\pi^{(n)} - \mu_n\|_1 \\ &= \|\pi^{(n)} - \mu_{n-1}\|_1 - \|\pi^{(n)}\boldsymbol{P}^{(n)} - \mu_{n-1}\boldsymbol{P}^{(n)}\|_1 \\ &\ge \|\pi^{(n)} - \mu_{n-1}\|_1 - C(\boldsymbol{P}^{(n)})\|\pi^{(n)} - \mu_{n-1}\|_1 \\ &= (1 - C(\boldsymbol{P}^{(n)}))\|\pi^{(n)} - \mu_{n-1}\|_1 \\ &> (1 - C(\boldsymbol{P}^{(n)}))\delta.\end{aligned}$$

And since $C(\boldsymbol{P}^{(n)}) < 1$, then for any $N \in \mathbb{N}^+$, there exists $(1 - C(\boldsymbol{P}^{(n)}))\delta > 0$, for all $n > N$,

$$\|\mu_n - \mu_{n-1}\|_1 > (1 - C(\boldsymbol{P}^{(n)}))\delta.$$

By Cauchy's convergence test, it is contradictory to

$$\|\mu_n - \pi\|_1 \to 0, \quad n \to \infty.$$

Thus for any $\epsilon > 0$, there exists $N_1 \in \mathbb{N}^+$, and when $n > N_1$,

$$\|\pi^{(n)} - \mu_{n-1}\|_1 < \frac{\epsilon}{2}.$$

Since $\|\mu_n - \pi\|_1 \to 0,\ n \to \infty$, there exists $N_2 \in \mathbb{N}^+$, for all $n > N_2$,

$$\|\mu_{n-1} - \pi\|_1 < \frac{\epsilon}{2}.$$

Taking $N = \max\{N_1, N_2\}$, when $n > N$, we have

$$\begin{aligned}\|\pi^{(n)} - \pi\|_1 &= \|(\pi^{(n)} - \mu_{n-1}) + (\mu_{n-1} - \pi)\|_1 \\ &\leq \|\pi^{(n)} - \mu_{n-1}\|_1 + \|\mu_{n-1} - \pi\|_1 \\ &< \epsilon.\end{aligned}$$

Then

$$\|\pi^{(n)} - \pi\|_1 \to 0, \quad n \to \infty.$$

■

Theorem 7 *Let $\boldsymbol{h}_u^{(l)}$ be representation of node $u \in \mathcal{V}$ at the hidden layer l in GAT. The sufficient condition for GAT to avoid over-smoothing is that there exists $\delta > 0$ such that for any $l \geq 2$, satisfying*

$$\|\boldsymbol{h}_u^{(l-1)} - \boldsymbol{h}_u^{(l)}\|_1 > \delta.$$

Proof By Theorem 3, for the GAT operator $\boldsymbol{P}_{\text{att}}^{(l)}$,

$$\pi^{(l)}(u) = \frac{\deg^{(l)}(u)}{\sum_{v \in \mathcal{V}} \deg^{(l)}(v)} \quad \forall u \in \mathcal{V},$$

where $\deg^{(l)}(u) := \sum_{k \in \mathcal{N}(u)} \exp(\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_k^{(l-1)}))$ is the weighted degree of u, where

$$\phi^{(l)}(\boldsymbol{h}_u^{(l-1)}, \boldsymbol{h}_k^{(l-1)}) := \text{LeakyReLU}(\mathbf{a}^{\mathrm{T}}[\boldsymbol{W}^{(l)}\boldsymbol{h}_u^{(l-1)} \| \boldsymbol{W}^{(l)}\boldsymbol{h}_k^{(l-1)}]).$$

Since $\mathcal{G}$ is connected, non-bipartite graph,

$$C(\boldsymbol{P}_{\text{att}}^{(l)}) < 1.$$

By Theorem 6, the sufficient condition for that there is no probability measure π on $\mathcal{M}$ such that

$$\|\mu_n - \pi\|_1 \to 0, \quad n \to \infty$$

is

$$\|\pi^{(n)} - \pi\|_1 \nrightarrow 0, \quad n \to \infty.$$

By the Cauchy's convergence test, it is equivalent to the existence of $\delta_\pi > 0$ such that for any $l \geq 1$, satisfying

$$\|\pi^{(l)} - \pi^{(l+1)}\|_1 > \delta_\pi.$$

Let $D^{(l)} := \sum_{u \in \mathcal{V}} \deg^{(l)}(u)$, $D_{\text{min}} = \min\{D^{(l)}, D^{(l+1)}\}$,

$$\begin{aligned}\|\pi^{(l)} - \pi^{(l+1)}\|_1 &= \sum_{u \in \mathcal{V}} \left| \frac{\deg^{(l)}(u)}{D^{(l)}} - \frac{\deg^{(l+1)}(u)}{D^{(l+1)}} \right| \\ &> \left| \frac{\deg^{(l)}(u)}{D^{(l)}} - \frac{\deg^{(l+1)}(u)}{D^{(l+1)}} \right| \\ &> \frac{1}{D_{\text{min}}} \left| \deg^{(l)}(u) - \deg^{(l+1)}(u) \right|\end{aligned}$$

and

$$\deg^{(l)}(u) := \sum_{k \in \mathcal{N}(u)} \exp(\text{LeakyReLU}(\mathbf{a}^{\mathrm{T}}[\boldsymbol{W}^{(l)}\boldsymbol{h}_u^{(l-1)} \| \boldsymbol{W}^{(l)}\boldsymbol{h}_k^{(l-1)}])).$$

Then if there exists $\delta > 0$ such that for any $l \geq 2$, satisfying

$$\|\boldsymbol{h}_u^{(l-1)} - \boldsymbol{h}_u^{(l)}\|_1 > \delta,$$

there must exist $\delta_\pi > 0$ such that for any $l \geq 1$, satisfying

$$\|\pi^{(l)} - \pi^{(l+1)}\|_1 > \delta_\pi.$$

■

B ANALYSIS OF GEN-SOFTMAX

In this appendix, we analyze Generalized Aggregation Networks (GEN-SoftMax) (Li et al., 2020) proposed for training deeper GNNs, which can be related to a time-inhomogeneous Markov chain.

In order to train deeper GNN models, Li et al. (2020) proposes a new message passing method between nodes u and v

$$\lambda_{u,v}^{(l)} := \frac{\exp(\beta \mathbf{m}_{u,v}^{(l-1)})}{\sum_{k \in \mathcal{N}(u)} \exp(\beta \boldsymbol{m}_{u,k}^{(l-1)})},$$

where β is inverse temperature and

$$\mathbf{m}_{u,v}^{(l)} := \text{ReLU}(\boldsymbol{h}_v + \mathbb{I}(\boldsymbol{h}_{(u,v)}^{(l)}) \cdot \boldsymbol{h}_{(u,v)}^{(l)}) + \epsilon \qquad v \in \mathcal{N}(u),$$

where $\mathbb{I}(\,\cdot\,)$ is an indicator function being 1 when edge representations exist otherwise 0, ϵ is a small positive constant chosen as 10^{-7}. Then the definition of message passing in GEN-SoftMax is

$$\boldsymbol{h}_u^{(l)} := \sum_{v \in \mathcal{N}(u)} \lambda_{u,v}^{(l)} \boldsymbol{h}_v^{(l-1)}.$$

Write in matrix form

$$\boldsymbol{H}^{(l)} = \boldsymbol{P}_{\text{GEN}}^{(l)} \boldsymbol{H}^{(l-1)},$$

where $\boldsymbol{P}_{\text{GEN}}^{(l)} \in \mathbb{R}^{N \times N}$ satisfies $P_{\text{GEN}}^{(l)}(u,v) = \lambda_{u,v}^{(l)}$ if $v \in \mathcal{N}(u)$ otherwise $P_{\text{GEN}}^{(l)}(u,v) = 0$, and $\sum_{v=1}^{N} P_{\text{GEN}}^{(l)}(u,v) = 1$.

Similar to the discussion of GAT in Section 3, we can relate GEN-SoftMax to a time-inhomogeneous Markov chain $\vec{\mathbf{v}}_{\text{GEN}}$ with a family of transition matrices of

$$\left\{ \boldsymbol{P}_{\text{GEN}}^{(1)}, \boldsymbol{P}_{\text{GEN}}^{(2)}, \ldots, \boldsymbol{P}_{\text{GEN}}^{(l)}, \ldots \right\}.$$

According to the discussion in Section 3, GEN-SoftMax is not always over-smoothing.

Similar to Theorem 7, we have the following sufficient condition to ensure that GEN-SoftMax avoids over-smoothing.

Theorem 8 *Let $\boldsymbol{h}_u^{(l)}$ be representation of node $u \in \mathcal{V}$ at the hidden layer l in GEN-SoftMax, then a sufficient condition for GEN-SoftMax to avoid over-smoothing in the Markovian sense is that there exists $\delta > 0$ such that for any $l \geq 1$, satisfying*

$$\|\boldsymbol{h}_u^{(l)} - \boldsymbol{h}_u^{(l+1)}\|_1 > \delta.$$

C EXPERIMENTAL DETAILS

In this appendix, we add more details on the experiments. Table 3 shows the basic information of each dataset used in our experiments.

Table 4 demonstrates the configuration of GNN models, actually, we keep the same setting in the corresponding paper, the only difference is we add the extra proposed regularization term in the optimization objective.

In Table 5, we show the detailed selection of threshold T in Equation 4.

Table 3: Summary of the statistics and data split of datasets.

Dataset	(Avg.) Nodes	(Avg.) Edges	Features	Class	Train(#/%)	Val.(#/%)	Test(#/%)
Cora	2708(1 graph)	5429	1433	7	140	500	1000
Citeseer	3327(1 graph)	4732	3703	6	120	500	1000
Pubmed	19717(1 graph)	44338	500	3	60	500	1000
ogbn-arvix	169,343(1 graph)	1,166,243	128	40	0.54	0.18	0.28
ogbg-molhiv	25.5(41,127 graph)	27.5	9	2	0.8	0.1	0.1
ogbg-ppa	243.4(158,100 graph)	2,266.1	7	37	0.49	0.29	0.22

Table 4: Training configuration

Model	Dataset	Hidden.	LR.	Dropout	Epoch	Block	GCN Agg.	β
GAT	Cora	64	1e-2	0.5	500	-	-	-
	Citeseer	64	1e-2	0.5	500	-	-	-
	Pubmed	64	1e-2	0.5	500	-	-	-
GEN	ogbn-arvix	256	1e-4	0.2	300	Res+	softmax_sg	1e-1
	ogbg-molhiv	128	1e-3	0.5	500	Res+	softmax	1
	ogbg-ppa	128	1e-2	0.5	200	Res+	softmax_sg	1e-2

Table 5: Selection of threshold T on different layer numbers

datasets	model	#layers					
		3	4	5	6	7	8
Cora	GAT-RT	1	0.5	0.6	0.8	1	0.8
Citeseer	GAT-RT	0.3	0.3	1	0.4	0.2	0.1
Pubmed	GAT-RT	0.3	0.2	0.2	0.8	0.5	0.4
		7	14	28	56		
ogbn-arxiv	GEN-RT	0.3	0.1	(0.1)0.3	-		
ogbg-molhiv	GEN-RT	0.7	0.1	0.5	1		
ogbg-ppa	GEN-RT	0.1	0.3	0.1	-		