# GATreg - Graph Attention Networks with Regularization

Mariam Ishtiaq
*AI Railroad Research Department*
*Korea Railroad Research Institute*
*Transportation System Engineering*
*University of Science and Technology*
Uiwang, Republic of Korea
mariam16@krri.re.kr

Jong-Un Won
*AI Railroad Research Department*
*Korea Railroad Research Institute*
*Transportation System Engineering*
*University of Science and Technology*
Uiwang, Republic of Korea
juwon@krri.re.kr

Sangchan Park
*Department of Technology and Society*
*The State University of New York Korea*
Incheon, Republic of Korea
sangchan.park@sunykorea.ac.kr

***Abstract*—To improve the applicability and performance of graph neural networks (GNNs); graph convolution networks (GCNs) and graph attention networks (GATs) have shown promising ways forward. However, lack of generalizability has been a major bottleneck for their widespread applications. To overcome this limitation of GNNs, we propose a regularization scheme for GAT, termed as GATreg. We use a novel loss function to achieve the optimal performance of GATreg. The proposed model has been analyzed using 3 benchmark datasets: Cora, Citeseer, and Pubmed. Our results show prominent improvements of classification accuracy, around 3%, compared to vanilla models and have the potential to be analyzed for further enhancements. Additionally, GATs can be explored to improve the reasoning ability of multi-modal large language models (LLMs), particularly to eliminate hallucinations. They can also be used for the quantum information theory-based analysis of GATs, for which this work analyzes recent literature.**

***Index Terms*—graph attention networks (GATs), graph convolutional networks (GCNs), graph regularization, multi-modal large language models, quantum information theory**

## I. Introduction

Graph Neural Networks (GNNs) operate on graph-structured data such as node features, edge features, and even global graph characteristics, making them versatile for different applications. GNNs leverage a message-passing framework, where nodes aggregate information from their neighbors to update their representations. This is particularly useful for semi-supervised graph node classification, which finds a wide range of applications in notable domains like social network analysis, biological networks, recommendation systems, and text classification. The results can pave the way for engineering state-of-the-art intelligent models using few-shot learning, fine-tuning large language models (LLMs), and exploring multi-modal models. Graph convolution networks (GCNs) and graph attention networks (GATs) are two prominent subclasses of GNNs aiming for generalization in graph learning. GCNs update node features by aggregating information from neighboring nodes using convolutions. GATs, on the other hand, employ an attention mechanism to graph neural networks, allowing nodes to attend differently to their neighbors when aggregating information.

Generalization in the realm of graph learning is crucial for ensuring that GNN models like GCNs and GATs can perform well on unseen data, especially in the context of unsupervised and semi-supervised learning. To address this limitation, this work provides two-fold contributions. 1) We study the information-theoretic perspective of graph regularization. By overcoming the issues of insufficient supervision and representation collapse, we enhance the performance of GCNs and GATs. 2) Our in-depth analysis opens new doors to analyze GATs for advanced domains like multi-modal LLMs and quantum information theory.

## II. Related Work

GNNs have a wide range of applications and have been rigorously studied with modifications to their architecture for performance boosts. [1] gives a recent comprehensive review of applications in the domains of data mining and computer vision.

Graph networks have been studied in the realm of quantum analog in [2]. The work proposes a baseline claim towards graphs for an entangled quantum system. In this realm, the quantization of edge data and adjacency matrix are potential open research problems.

GNNs suffer from graph information bottleneck (GIB), which implies a lack of regularization methods. Therefore, [3] proposes a GIB, specifically tailored for explainability, called GIBE. The paper studies the impact of regularization, sparsity, and masking while opening new research paradigms to explore the role of regularization, which motivates this work.

A similarity-based regularized softmax function has been proposed in [4]. The prediction results are regularized by a non-local total variation. However, this manual tuning approach needs to be studied for applicability in large complex data structures.

Regularization has also been studied to overcome non-convexity of objective functions [5]. Non-convex objective functions result in multiple local minima, thus preventing the algorithm from finding the global minimum, and hence, optimization. Therefore, a regularization term in linear mixing

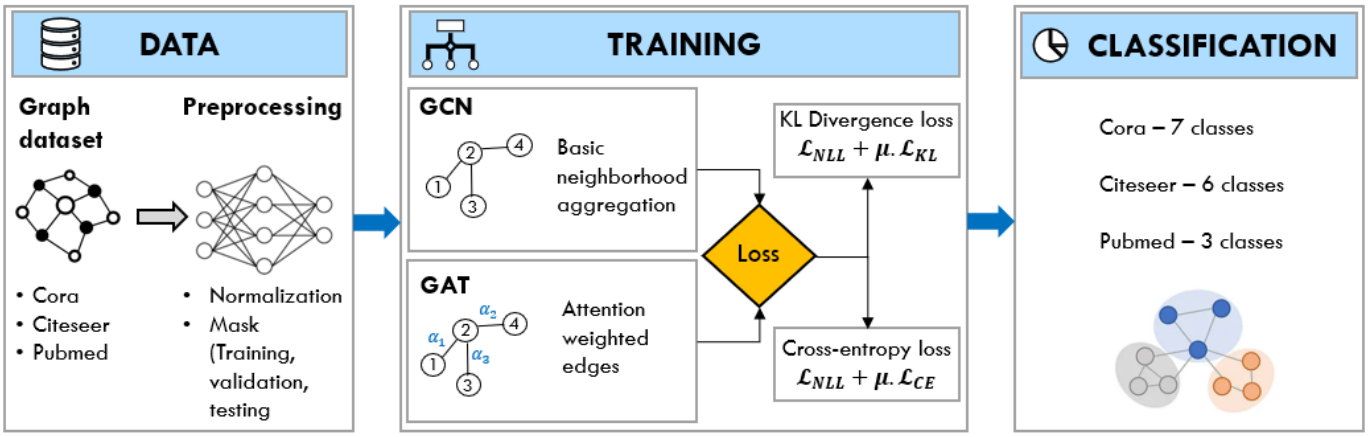


Fig. 1. GAT-reg pipeline with data preprocessing, training and classification stages.

models (LMMs) has been used as a sparsity regularization-based loss function.

## III. Methodology

In this work, we present an in-depth analysis on the information theory perspective of regularization, a schematic of which has been shown in Figure 1. The proposed regularization shcheme aims to avoid overfitting by keeping the training in line with the prior expectation, thus improving generalization.

### A. Dataset

We use three benchmark citation datasets: Cora, Citeseer, and Pubmed. In these datasets, the nodes refer to the scientific papers, and the edges refer to the citation relationships. Cora and Citeseer are built on the *bag-of-words (BoW)* representation, leading to high-dimensional sparse vectors since all words are treated equally. Pubmed dataset uses the *term frequency-inverse document frequency (TF-IDF)* approach, where each word is weighed by its frequency on a collection of documents in the dataset. The frequency $f(t,d)$, of a term $t$ in a document $d$ for BoW (Cora and Citeseer) and TF-IDF (Pubmed) can be estimated using equations (1) and (2) [6], respectively.

$$\text{BoW}(t,d) = f(t,d) \tag{1}$$

$$\begin{aligned}\text{TF-IDF}(t,d) &= \text{TF}(t,d) \times \text{IDF}(t,d) \\ &= \left(\frac{f(t,d)}{\sum_{t' \in d} f(t',d)}\right) \times \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)\end{aligned} \tag{2}$$

In (2), $\text{TF}(t,d)$ is the *term frequency* of $t$ in $d$, calculated as $\frac{f(t,d)}{\sum_{t' \in d} f(t',d)}$, and the denominator sums the frequencies of all terms $t'$ in document $d$. This normalizes the frequency of the term within the document, providing a relative measure. The IDF(t,d) term stands for the *inverse document frequency* of the term $t$, which measures how much information the term provides across all documents in the corpus. Here, $N$ is the total number of documents, and the number of documents containing the term $t$ is given by the denominator $|\{d \in D : t \in d\}|$.

Table I shows the statistics of the three datasets used for our experiments.

TABLE I
Statistics of Cora, Citeseer, and Pubmed Datasets

| Dataset | NumNodes | NumEdges | NumFeats | NumClasses |
|---|---|---|---|---|
| Cora | 2708 | 10556 | 1433 | 7 |
| Citeseer | 3327 | 9104 | 3703 | 6 |
| Pubmed | 19717 | 88648 | 500 | 3 |

We use the Plantoid standard split for our experiment using training, validation, and testing splits in the ratio 70:20:10. The Plantoid split is a benchmark that uses fixed class distribution to ensure dataset balance and a consistent comparison across experiments. We also use masks to assign nodes to appropriate subsets so the model can focus on appropriate data during training, validation, and testing.

## IV. Model Generalization

We analyze three models: GCN, GAT, and regularized graph attention networks (GATreg). GCNs [7] aggregate feature information from the neighbors of a node in a layer-wise convolutional manner. A single layer of a GCN can be given by (3):

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \tag{3}$$

where $\mathbf{H}^{(l)}$ is the feature matrix at layer $l$, $\mathbf{A}$ is the adjacency matrix of the graph, $\mathbf{D}$ is the degree matrix, $\mathbf{W}^{(l)}$ is the weight matrix for current layer $l+1$, and $\sigma$ is the nonlinear activation function. GCN is inherently limited to this fixed aggregation function to learn neighboring node features. To enhance the generalizability and include the weightage of different neighbors and their contributions to the node representation, GAT [8] can be expressed as (4):

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_k]\right)\right)} \tag{4}$$

where $\alpha_{ij}$ is the attention coefficient between nodes $i$ and $j$, $\mathbf{h}_i$ and $\mathbf{h}_j$ are the feature vectors for the nodes, $\mathbf{W}$ is the weight matrix, and $\mathbf{a}$ is the attention vector.

*a) Loss Function:* During training, the cross entropy loss (5) compares the predicted class probabilities (obtained from the softmax function) with the actual class labels, while Kullback-Leibler (KL) divergence (6) measures the difference between the predicted distribution $P$ (after the propagation through the network) and a target distribution $Q$ (true or smoothed).

$$\text{CE} = -\sum_{i=1}^{C} y_i \log(p_i) \tag{5}$$

$$\text{KL}(P\|Q) = \sum_{i=1}^{C} p_i \log\left(\frac{p_i}{q_i}\right) \tag{6}$$

We integrate these losses separately to study their impact on each model using the 3 selected benchmark datasets. Loss minimization is vital to improving a model's generalizability, convergence, and robustness, and we aim to achieve this using regularization.

*b) Regularization: GAT-reg:* Different regularization approaches have been used in literature, like [9] adopting a spatial regularization technique by minimizing the distance between latent representations. We define a regularization based loss function $\mathcal{L}$ composed of the negative log likelihood loss $\mathcal{L}_{\text{NLL}}$ and a regularization term based on CE loss $\mathcal{L}_{\text{CE}}$ from (5), and KL divergence $\mathcal{L}_{\text{KL}}$ from (6), given in equations (7) and (8), respectively.

$$\mathcal{L}_{\text{reg-KL}} = \mathcal{L}_{\text{NLL}} + \mu \cdot \mathcal{L}_{\text{KL}} \tag{7}$$

$$\mathcal{L}_{\text{reg-CE}} = \mathcal{L}_{\text{NLL}} + \mu \cdot \mathcal{L}_{\text{CE}} \tag{8}$$

To encourage the model to make an accurate prediction, we use a negative log-likelihood loss to measure the discrepancy between the predicted probabilities of the model and the true labels for the training data. CE or KL divergence are used as regularizers, keeping the model's output distribution aligned. The hyperparameter $\mu$ prevents overfitting by penalizing large coefficients.

*c) Experiment:* The hardware and environment specification for our experiment is given in Table II.

TABLE II
SYSTEM AND ENVIRONMENT SPECIFICATION

| Component | Specification |
|---|---|
| Operating system | Windows 10 |
| Processor | 11th Gen Intel® Core™ i7-11800H @ 2.30 GHz |
| RAM | 64.0 GB |
| GPU | NVIDIA GeForce RTX 3070 |
| CUDA version | 11.3 |
| Pytorch Geometric version | 2.3.1 |

The hyperparameters used for our experiments are provided in Table III. The set of hyperparameters was chosen based on the results of multiple runs using the selected 3 datasets to obtain consistently better performance using the GATreg baseline. However, they can be further fine-tuned using more diverse datasets to perform better. For instance, $\mu$=0 implies vanilla GAT and GCN models, while a variation like $\mu$=0.5 implies a higher impact of the regularized loss term.

TABLE III
HYPERPARAMETERS AND THEIR DESCRIPTIONS

| Hyperparameters | Description |
|---|---|
| **Epochs** | 1000 |
| **Learning Rate** | 0.01 |
| **Dataset** | Cora, Citeseer, Pubmed |
| **Hidden Channels (GCN)** | 16 (used in the GCN model) |
| **Hidden Channels (GAT)** | 8 (used in the GAT model) |
| **Heads** | 8 (number of attention heads in GAT) |
| **Dropout Rate** | 0.6 (dropout used in GAT layers) |
| **Weight Decay** | 5e-4 (for Adam optimizer) |
| **mu** ($\mu$) | 0.5 |

## V. RESULTS

Table IV shows the classification accuracy results obtained for our experiments, with the best results in bold. The better performance of GCN compared to GAT can be attributed to the smaller graph structures and relatively simpler relationships between nodes in Cora, Citeseer, and Pubmed, where the uniform message passing of GCN is effective. The attention mechanism of GAT introduces complexity in the model, which is ineffective in less complex datasets. Our regularized GAT model allows an adaptive focus on relevant connections, which is useful for datasets with non-uniform nodes, such as semi-supervised scenarios. Moreover, the regularization prohibits overfitting, thus enhancing generalization to unseen data.

TABLE IV
PERFORMANCE ANALYSIS OF GRAPH CONVOLUTIONAL NETWORK (GCN), GRAPH ATTENTION NETWORK (GAT), AND REGULARIZED GRAPH ATTENTION NETWORK (GAT-REG). THE RESULTS OF EACH MODEL HAVE BEEN REPORTED USING KL DIVERGENCE (KL) AND CROSS-ENTROPY (CE) LOSS ANALYSIS, USING CLASSIFICATION ACCURACY ± STANDARD DEVIATION.

| Model | Datasets | | |
|---|---|---|---|
| | **Cora** | **Citeseer** | **Pubmed** |
| **RGCN** [10] | 0.819 ± 0.011 | 0.741 ± 0.016 | **0.792 ± 0.021** |
| **PREGGAT** [11] | 0.8297 ± 0.0119 | 0.7000 ± 0.0189 | 0.7639 ± 0.0146 |
| **LSGAT** [12] | 0.791 ± 0.013 | 0.681 ± 0.009 | 0.786 ± 0.012 |
| **Our Results** | | | |
| **GCN-KL** | 0.8014 ± 0.0037 | 0.7224 ± 0.0042 | **0.7906 ± 0.0036** |
| **GCN-CE** | 0.8105 ± 0.0058 | 0.7078 ± 0.0037 | 0.7896 ± 0.0046 |
| **GAT-KL** | 0.7998 ± 0.0115 | 0.7036 ± 0.0172 | 0.7677 ± 0.0085 |
| **GAT-CE** | 0.8140 ± 0.0045 | 0.7058 ± 0.0084 | 0.7663 ± 0.0055 |
| **GATreg-CE**(7) | **0.8407 ± 0.00436** | **0.7243 ± 0.00372** | 0.7812 ± 0.00708 |
| **GATreg-KL**(8) | 0.8246 ± 0.00639 | 0.7172 ± 0.01187 | 0.7811 ± 0.00789 |

We also compare our work to other regularization techniques. A graph similarity regularized softmax for GNNs

(RGNN) has been proposed in [10] for semi-supervised node classification. A propagation regularization scheme for GAT (PREGGAT) [11] uses a non-trivial variant of graph Laplacian regularization. The efficacy of regularization in boosting the performance of GNNs has been evaluated. An aggregation-based regulation scheme, layer-wise self-adaptive GAT (LS-GAT), has been proposed in [12]. LSGAT allows for the integration of GAT into existing GNN models without architecture changes. Compared to these state-of-the-art models, our regularization scheme is outperforming in terms of classification accuracy. A representative radar diagram of our results has been shown in Figure 2.

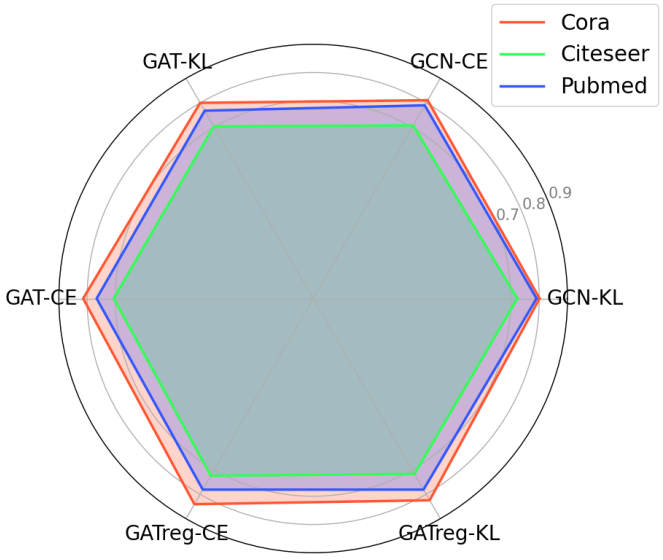


Fig. 2. Comparative analysis of classification accuracy across Cora, Citeseer, and Pubmed datasets. We compare the performance of GCN, GAT, and GATreg models for cross-entropy and KL divergence loss.

Figure 3 shows that GATreg-CE consistently achieves a higher test accuracy over 10 runs on the Cora dataset. It is also evident that the regularization significantly enhances the model's performance compared to other models.

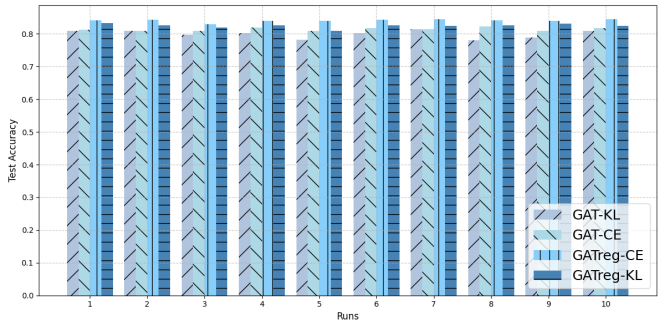


Fig. 3. Test Accuracy over 10 Runs (1000 epochs each) for Cora dataset.

From Figure 4, we observed that regularized models result in better overall performance. Additionally, GATreg-CE stands out with the highest average accuracy, indicating its effectiveness across multiple runs on the Citeseer dataset. The optimal performance of GATreg-CE across Cora and Citeseer datasets suggests that the proposed regularization techniques are particularly suited to the datasets' characteristics, leading to better generalization.

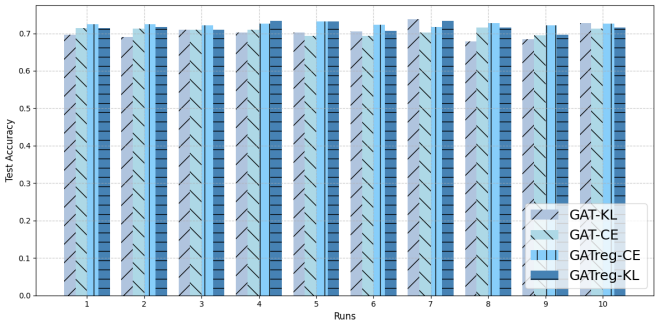


Fig. 4. Test Accuracy over 10 Runs (1000 epochs each) for Citeseer dataset.

While GCN shows the best performance for the Pubmed dataset, Figure 2, the accuracy variation is less pronounced compared to Cora and Citeseer. Earlier literature has improved the performance of proposed GAT models over Pubmed dataset using hyperparameter tuning [13], where a large number of attention heads has led to better results. However, the standard set of hyperparameters in Table 3 has been achieved through empirical optimization of GATreg using the selected datasets. GATreg-CE, although not showing the best results, can be viewed as a suitable contender across diverse datasets. The performance enhancement induced by the regularization is significantly evident from Figure 5, on Pubmed dataset.

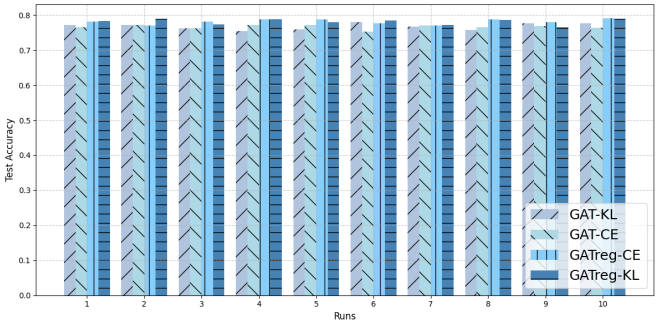


Fig. 5. Test Accuracy over 10 Runs (1000 epochs each) for Pubmed dataset.

## VI. Conclusion and Future Work Direction

To increase the applicability of GATs for diverse applications, we analyzed the performance of graph-based networks like GCNs and GATs across three benchmark datasets: Cora, Citeseer and Pubmed. While GCN has better results compared to GAT, our proposed regularization architecture, GATreg, significantly improved the performance of GAT model. This highlights the significance of the proposed regularization scheme in increasing the GNN's robustness.

This regularization can also be studied to enhance LLM's capabilities in multimodal networks. In the future, we also aim to improve this regularization scheme and study its relevance and analogy in the domain of quantum graph networks.

## ACKNOWLEDGMENT

This research was supported by a grant from the R&D Program: 'Development of Rail-Specific Digital Resource Technology based on an AI-Enabled Rail Support Platform,' grant number PK2401C1, of the Korea Railroad Research Institute (KRRI).

## REFERENCES

[1] C. Chen, Y. Wu, Q. Dai, H.-Y. Zhou, M. Xu, S. Yang, X. Han, and Y. Yu, "A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] F. Razavinia and G. Haghighatdoost, "A route to quantum computing through the theory of quantum graphs." [Online]. Available: http://arxiv.org/abs/2404.13773

[3] J. Fang, G. Zhang, K. Wang, W. Du, Y. Duan, Y. Wu, R. Zimmermann, X. Chu, and Y. Liang, "On regularization for explaining graph neural networks: An information theory perspective," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[4] Y. Yang, J. Liu, and W. Wan, "Graph similarity regularized softmax for semi-supervised node classification." [Online]. Available: http://arxiv.org/abs/2409.13544

[5] W. He, H. Zhang, and L. Zhang, "Sparsity-regularized robust non-negative matrix factorization for hyperspectral unmixing," vol. 9, no. 9, pp. 4267–4279, conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7403908

[6] C. Wu, "Graph representation learning: from kernel to neural networks," 2021, artificial Intelligence [cs.AI], Institut Polytechnique de Paris, English. NNT: 2021IPPAX135. tel-03662478.

[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *CoRR*, vol. abs/1609.02907, 2016. [Online]. Available: http://arxiv.org/abs/1609.02907

[8] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks." [Online]. Available: http://arxiv.org/abs/1710.10903

[9] M. Agarwal, P. Das, and U. Bhatia, "Spatially regularized graph attention autoencoder framework for detecting rainfall extremes." [Online]. Available: http://arxiv.org/abs/2411.07753

[10] Y. Yang, J. Liu, and W. Wan, "Graph similarity regularized softmax for semi-supervised node classification." [Online]. Available: http://arxiv.org/abs/2409.13544

[11] H. Yang, K. Ma, and J. Cheng, "Rethinking graph regularization for graph neural networks," *CoRR*, vol. abs/2009.02027, 2020. [Online]. Available: https://arxiv.org/abs/2009.02027

[12] G. Su, H. Wang, Y. Zhang, W. Zhang, and X. Lin, "Simple and deep graph attention networks," *Knowledge-Based Systems*, vol. 293, p. 111649, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705124002843

[13] W. Gu, F. Gao, X. Lou, and J. Zhang, "Link prediction via graph attention network." [Online]. Available: http://arxiv.org/abs/1910.04807