Moderate Message Passing Improves Calibration: A Universal Way to Mitigate Confidence Bias in Graph Neural Networks

Min Wang*, Hao Yang*, Jincai Huang, Qing Cheng†

[1]College of Systems Engineering, National University of Defense Technology
wangminwm@nudt.edu.cn, yanghao@nudt.edu.cn, huangjincai@nudt.edu.cn, sgggps@163.com

Abstract

Confidence calibration in Graph Neural Networks (GNNs) aims to align a model's predicted confidence with its actual accuracy. Recent studies have indicated that GNNs exhibit an under-confidence bias, which contrasts the over-confidence bias commonly observed in deep neural networks. However, our deeper investigation into this topic reveals that not all GNNs exhibit this behavior. Upon closer examination of message passing in GNNs, we found a clear link between message aggregation and confidence levels. Specifically, GNNs with extensive message aggregation, often seen in deep architectures or when leveraging large amounts of labeled data, tend to exhibit overconfidence. This overconfidence can be attributed to factors like over-learning and over-smoothing. Conversely, GNNs with fewer layers, known for their balanced message passing and superior node representation, may exhibit under-confidence. To counter these confidence biases, we introduce the Adaptive Unified Label Smoothing (AU-LS) technique. Our experiments show that AU-LS outperforms existing methods, addressing both over and under-confidence in various GNN scenarios.

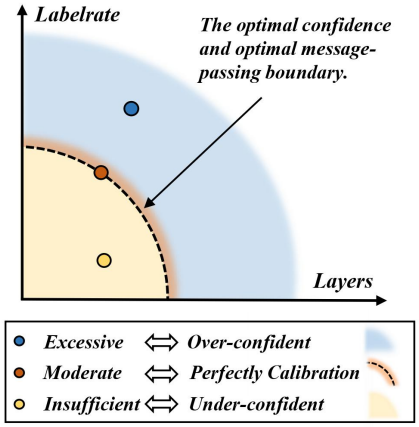


Figure 1: Schematic depicting the interplay among GNN layers, label rates, and message aggregation, and their impact on model confidence. The orange arc indicates the ideal confidence boundary.

Introduction

Graph Neural Networks (GNNs), designed for learning graph-structured data representations, have excelled in applications ranging from chemistry to social networks and transportation (Ruiz Puentes et al. 2022; Zhang et al. 2018; Hamilton, Ying, and Leskovec 2017; Zhou et al. 2020; Wang and Van Hoof 2022). Alongside their widespread use, there is a growing societal awareness regarding the imperative to ensure the reliability and trustworthiness of these systems (Yang et al. 2021; Wang et al. 2023; Li et al. 2023; Yang et al. 2023). A critical aspect in achieving trustworthy classifiers is the confidence calibration, aligning their confidence with the true likelihood of predictions, which is vital across various domains (Guo et al. 2017). This emphasis on confidence calibration is essential for upholding the trustworthiness of GNN-based classifiers in decision-making processes.

However, neural networks often face calibration challenges. Recent studies (Wang et al. 2021; Wang, Yang, and Cheng 2022) reveal a noteworthy divergence from conventional Deep Neural Networks (DNNs) behavior—GNNs tend to exhibit underconfidence instead of the overconfidence commonly observed in DNNs. This unique behavior prompts further investigation into refining GNN calibration techniques, especially in safety-critical applications where accurate confidence estimation is paramount.

In this work, we explore GNN calibration through the lens of message passing mechanisms, aiming to elucidate factors influencing GNN confidence calibration. Inspired by Shannon's principle, 'the essence of information is to eliminate uncertainty'(Burgin 2002), we adapt this concept to GNN calibration. Our novel insight, supported by mathematical proofs, reveals that GNN confidence is not consistently underconfident but correlates with the degree of message aggregation from neighboring nodes in the network. Message Aggregation in GNNs refers to the process by which a node gathers and combines information from its neighboring nodes, intending to reduce uncertainty and ideally increase model prediction confidence. Fig.1 visually illustrates the interplay between GNN layers, labelrates, and the extent of message aggregation, demonstrating how these

*These authors contributed equally.
†Corresponding author.

factors correlate with model confidence. We posit that insufficient aggregation, potentially resulting from shallow architectures or sparse connectivity, hampers the network's ability to assimilate contextual information, manifesting in underconfidence. Conversely, deep GNNs with high label rates risk excessive aggregation, which may obscure inter-cluster distinctions, resulting in overconfidence. Despite stagnant accuracy improvements, the loss function may guide the network to enhance output logit values. In summary, we argue that GNN calibration is a dual challenge: addressing under-confidence due to inadequate aggregation and rectifying over-confidence when aggregation is excessive.

Motivated by our analysis of the relationship between the amount of message aggregation and model confidence, we propose the **A**daptive **U**nified **L**abel **S**moothing (AU-LS), a method that can mitigate GNNs confidence bias and and calibrate predictions more reliably. AU-LS innovatively combines Label Smoothing, which adjusts the target outputs to prevent extreme predictions, and Negative Label Smoothing, which does the opposite, into a novel **U**nified **L**abel **S**moothing (ULS) framework. This ULS framework serves as the basis for calibrating the confidence of GNNs to align more consistently with their accuracy, irrespective of whether the models are under-confident or over-confident. Additionally, AU-LS introduces an adaptive component: it incorporates the output of a harmonic series parameterized by the current epoch value into the smoothing hyperparameter. This adaptively modulates calibration as training progresses, iteratively improving GNN's confidence. A significant advantage of AU-LS is its practicality—it can be readily adopted without intricate modifications to existing loss functions or training schemes, making it easy to implement within existing GNN frameworks.

The contributions of our paper are summarized as follows:

- We reveal a novel correlation that the quantity of messages aggregated from neighboring nodes is linked to prediction confidence, challenging the notion that GNNs are consistently under-confident.
- We introduce AU-LS, an adaptive method designed to align GNNs' prediction confidence more closely with their actual accuracy, effectively addressing confidence bias issues. Extensive experiments show that AU-LS robustly calibrates GNNs under varying conditions, and outperforms existing SOTA calibration techniques.

Related Work

Message Passing Mechanism of GNNs

As an emerging machine learning model, GNNs are revolutionizing both scientific and industrial domains. In essence, GNNs operate by iteratively updating node representations through a Message Passing Mechanism (MP) (Gilmer et al. 2017). Given a graph $G = (V, E, X)$, with initial node representations $\boldsymbol{h}_v^{(0)}$ set as node attributes X_v, GNNs generally perform two main steps in each layer k:

Aggregation: Each node v aggregates message from its neighbors $\mathcal{N}_v$

$$\boldsymbol{m}_{\mathcal{N}(v)}^{(k)} = \text{Aggregate}^{(k)}\left(\left\{h_v^{(k-1)}, v \in \mathcal{N}(v)\right\}\right)$$

Update: Node v updates its representation based on its own previous representation and the aggregated messages

$$\boldsymbol{h}_v^{(k)} = \text{Update}^{(k)}\left(\boldsymbol{h}_v^{(k-1)}, \boldsymbol{m}_{\mathcal{N}(v)}^{(k)}\right)$$

Among the various GNN architectures inspired by this mechanism, GCNs (Kipf and Welling 2017) and GATs (Veličković et al. 2018) stand as notable examples, drawing inspiration from CNNs (Krizhevsky, Sutskever, and Hinton 2012) and self-attention mechanisms, respectively.

Confidence Calibration

Despite deep learning's focus on performance, predictive uncertainty is a lasting concern in the literature (Ghahramani 2015; Wang and Van Hoof 2020; Wang, Federici, and van Hoof 2022; Wang and van Hoof 2022; Shen et al. 2023). Recent studies, such as a significant work by (Guo et al. 2017), have paid attention to the confidence of predictions and highlighted the poor calibration of modern neural networks, emphasizing a critical mismatch between confidence and accuracy. Given the importance of obtaining a reliable neural network, confidence calibration has been studied in various contexts, to predict probability by estimating representative of true correctness likelihood. To address this miscalibration, methods generally fall into two categories: post-hoc calibration and regularization techniques.

Post-hoc. Post-hoc calibration methods, such as Temperature Scaling (Guo et al. 2017), adjust the model's logits using a calibration map learned from validation data. However, they may underperform under data distribution shifts (Ovadia et al. 2019). It learns a scalar learnable parameter $T > 0$, named as temperature, to smooth or sharpen the prediction probability logit vectors. In a k-class classifier, the confidence calibrated by Temperature Scaling is

$$\hat{q}_i = \max_k \sigma_{\text{SM}}\left(\mathbf{z}_i/T\right)^{(k)}$$

where $\mathbf{Z}_i$ is the logits vector for input $\mathbf{X}_i$ and σ_{SM} is the softmax function.

Regularization. In contrast to explicit confidence adjustment through post-hoc calibration, recent methods have explored implicit regularization techniques that affect the entropy of training labels. Focal Loss (Lin et al. 2017) minimizes the Kullback-Leibler (KL) divergence between target and predicted distributions, thus curbing model overconfidence (Mukhoti et al. 2020). Mixup (Zhang et al. 2017) employs input and label interpolation to prevent overfitting and reduce overconfidence. Label Smoothing (Szegedy et al. 2016) uses soft labels instead of one-hot labels to discourage extreme probability assignments and thereby improve calibration (Müller, Kornblith, and Hinton 2019).

Confidence Calibration of GNNs. Recently, the calibration of GNNs' confidence has emerged as a significant topic. Research has shown that GNNs are miscalibrated in supervised scenarios, and traditional calibration methods prove ineffective (Teixeira, Jalaian, and Ribeiro 2019). CaGCN (Wang et al. 2021) addresses this by introducing a topology-aware post-hoc calibration function, inspired by Temperature Scaling (Guo et al. 2017). This approach allows each node to

learn a unique temperature parameter T, effectively calibrating the GNNs. Despite these advancements, further exploration into GNNs' confidence calibration, particularly considering their distinct message passing mechanism, remains an open avenue for research.

Preliminaries

For a semi-supervised K-class classification task, we denote an undirected graph $G = (V, E)$, with V and E represent the set of nodes and edges respectively. Training dataset $\mathcal{D}(\mathcal{X}, \mathcal{Y}) = \left\{\left(\mathbf{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$ consists of N nodes, where $\mathbf{x}^{(i)} \in \mathcal{X}$ is the i^{th} node and $y^{(i)} \in \mathcal{Y} \subset \mathbb{R}^K$ corresponds to the ground-truth label of node i with K classes, provided as one-hot encoding. The GNN classifier $f_\theta : \mathcal{X} \to \mathbb{R}^K$ is learned to map an input to the label space, in which θ is a trainable parameter. $\mathbf{L}^{(i)} = f_\theta(\mathbf{x}^{(i)})$ denotes the logit vectors computed by f_θ on an input $\mathbf{x}^{(i)}$, where $\mathbf{L} = (l_k)_{1 \le k \le K}$ and l_k presents the probability of the k-th class.

Perfect-calibrated Models: After softmax, the predicted probabilities are $\mathbf{s}^{(i)} = \left(\frac{e^{l_k}}{\sum_{j=1}^{K} e^{l_j}}\right)^{(i)}_{1 \le k \le K}$, where $\hat{p} = \max_k \frac{e^{l_k}}{\sum_{j=1}^{K} e^{l_j}}$ and $\hat{y} = \operatorname{argmax}_k \frac{e^{l_k}}{\sum_{j=1}^{K} e^{l_j}}$ represent the predicted confidence and the predicted label. For a perfectly calibrated GNN f_θ: $\mathbb{P}(\hat{y} = y \mid \hat{p} = p) = p, \forall p \in [0, 1]$.

Miscalibrated Models and Evaluation: In practice, GNNs rarely satisfy the condition for perfect calibration. To quantify such deviations, various metrics have been proposed to measure model calibration errors. An *over-confident* model predicts confidences that are systematically higher than the actual accuracy, while an *under-confident* model predicts confidences lower than its actual accuracy. One widely used evaluation metric is the Expected Calibration Error (ECE) (Naeini, Cooper, and Hauskrecht 2015), defined as: $\mathbb{E}\hat{p}[|\mathbb{P}(\hat{y} = y \mid \hat{p} = p) - p|]$. ECE is approximated as:

$$ECE \approx \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\operatorname{acc}(B_m) - \operatorname{conf}(B_m)\right|$$

where B_m is the set of samples whose confidence belongs to the m-th bin, with $|B_m|$ representing the number of these samples. The average accuracy and confidence of B_m are computed as $\operatorname{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$ and $\operatorname{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$, respectively.

In addition to ECE, Brier Score (Brier et al. 1950) and Negative Log Likelihood (NLL) (Friedman, Hastie, and Tibshirani 2001) are also used to evaluate model calibration.

Definition of Logit distances: Following the proposal by (Liu et al. 2022), we define the logit distances vector, $\mathbf{d}(\mathbf{L})$, as the measure of the distance between the predicted class and the rest of the classes:

$$\mathbf{d}(\mathbf{L}) = \left(\max_j (l_j) - l_k\right)_{1 \le k \le K}$$

where each element is non-negative and K is the number of classes.

Importantly, sharpening $\mathbf{d}(\mathbf{L})$ aims to increase the confidence of an *under-confident* model, while softening $\mathbf{d}(\mathbf{L})$ helps to reduce the *overconfidence* of a model.

Correlations between Confidence and Message Aggregation

In this section, we investigate the correlations between confidence and message aggregation from experimental observations and theoretical proofs.

Theoretical Proofs

We assume a simple GNN architecture using mean aggregation, and balanced data across classes to avoid biases in confidence predictions due to class imbalances.

Proposition 1. *Excessive message passing in deeper GNNs, especially with high label rates, blurs inter-cluster distinctions, resulting in overconfidence without accuracy improvements.*

Proof: Step 1: Assume node embeddings $h_v^{(l+1)}$ converge towards each other as l increases(Figure 2(a-i)), which can be expressed as:

$$\lim_{l \to \infty} d^{(l)}(u, v) = 0$$

Step 2: This convergence causes the classifier's decision boundary to blur, diminishing the distinction between categories.

Step 3: The Softmax classifier tends to concentrate probabilities more on specific categories during normalization. Consequently, the output probability distribution tends towards extremes. The high similarity in node representations contributes to low accuracy, thus resulting in overconfidence:

$$\mathbb{P}\left(\hat{y} = y \mid \hat{p}_v^{(L)}\right) < \hat{p}_v^{(L)}$$

Proposition 2. *Insufficient message passing, due to shallow architectures or limited node connectivity, leads to under-confident predictions by hindering the network's assimilation of broader structural information.*

Proof: Step 1: For shallow GNNs, node v's representation $h_v^{(L)}$ primarily incorporates local neighbor information.

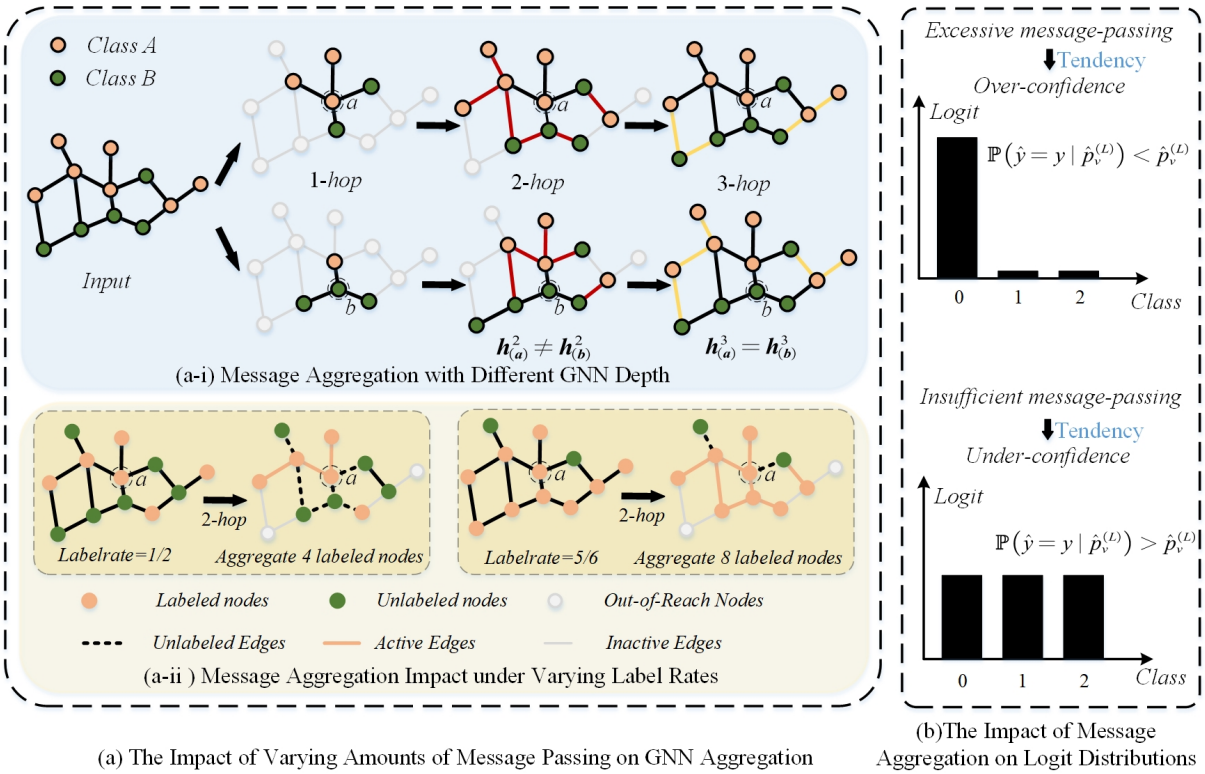
Step 2: This limited local context results in a less distinct probability distribution and under-confident predictions.

$$P\left(y_v \mid h_v^{(L)}\right) \approx \frac{1}{C}$$

Experimental Observations

We experimentally analyze the influence of GNN depth and label rate in the training dataset on model confidence, considering various architectures and datasets.

Influence of deeper GNN architecture: In a semi-supervised classification task, we extend a GCN to a 3-layer architecture and assess its calibration on the Cora, Citeseer, and Pubmed datasets with a label rate $L/C = 20$. Contrary to the conclusion by CaGCN (Wang et al. 2021), we observe that in some cases, a 3-layer GCN exhibits over-confidence, as demonstrated by a lower average accuracy compared to the average confidence (see Figure 3).



(a) The Impact of Varying Amounts of Message Passing on GNN Aggregation

(b)The Impact of Message Aggregation on Logit Distributions

Figure 2: Visual Representation of Message Aggregation and Its Consequences in GNNs. (a) illustrates how varying GNN depth (a-i) and Labelrate (a-ii) impact message aggregation. (b) contrasts the logit distributions of different classes under conditions of excessive and insufficient message passing.

Influence of labelrate of training dataset: Using a 2-layer GCN model, we vary label rates ($L/C = 20, 30, 40, 50, 60$) on Cora, Citeseer, and Pubmed datasets (Sen et al. 2008). Figure 4 illustrates a positive correlation between label rate and model confidence.

Based on the above observations and proofs, we can draw the following preliminary conclusions: In GNNs, message aggregation appears to be associated with the balance between confidence and accuracy. Excessive aggregation may lead to a phenomenon referred to as 'over-smoothing,' and potentially 'overfitting,' resulting in overconfident predictions. Conversely, insufficient aggregation may induce underconfident predictions. These observations suggest the existence of a critical threshold for message aggregation, potentially representing a level of aggregation that optimizes predictive confidence. This threshold appears to be influenced by factors such as the depth of the GNN architecture and the label rate of the training dataset.

Methods

Based on the relationship between model confidence and the amount of message aggregation and the discovery that GNNs not only suffer under-confidence, but also over-confidence, we propose our method Adaptive Unified Label Smoothing (AU-LS) to adaptively calibrate GNNs' confidence to be more aligned with their accuracy.

Label Smoothing

Label Smoothing (Szegedy et al. 2016) is a regularization strategy, which replaces hard target distribution with a weighted average of the original hard target distribution and the uniform distribution. Specifically, the original one-hot training labels $\mathbf{y} \in \{0,1\}^K$ become soft labels $\mathbf{y}^\epsilon = (y_k^\epsilon)_{1\le k\le K}$, where $y_k^\epsilon = y_k(1-\epsilon) + \frac{\epsilon}{K}$ and parameter $\epsilon \in [0,1]$ controls the degree of the smoothing effect. The optimization objective of LS is to minimize the Cross-Entropy between the soften labels y_k^ϵ and model predictions p_k, which is computed as:

$$\mathcal{L}_{\text{LS}}(\mathbf{y}, \mathbf{p}) = -\sum_k y_k^\epsilon \log s_k \tag{1}$$

In addition to improving the performance of classifiers, recent evidence (Lukasik et al. 2020; Müller, Kornblith, and Hinton 2019) suggests that LS(Szegedy et al. 2016) can alleviate models' over-confidence by preventing the network from assigning the full probability to a single class.

Negative Label Smoothing: Contrary to Label Smoothing where $\epsilon \in [0,1]$, in Negative Label Smoothing, we allow ϵ to be in the range $[-1,0]$. For Negative Label Smoothing (NLS), the soft labels are computed as: $y_k^\epsilon = y_k(1-\epsilon) + \frac{\epsilon}{K}$.

Proposed Method: AU-LS

To take into account the two miscalibration cases of GNN including over and under confidence, we unify Negative La-

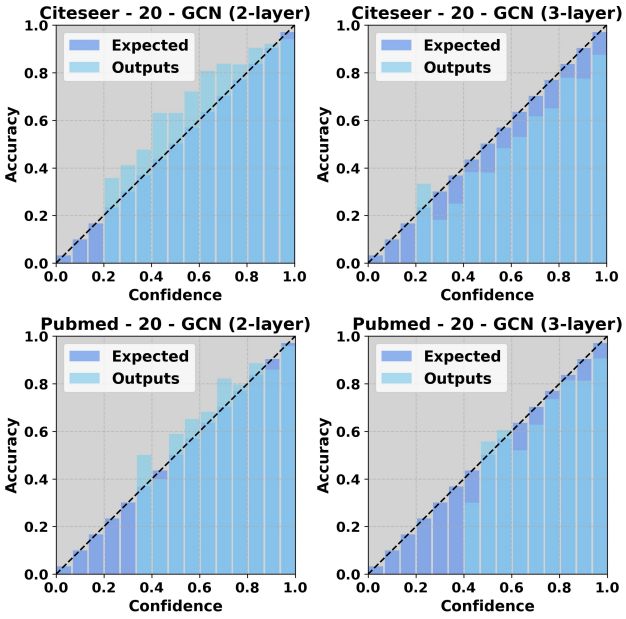


Figure 3: Reliability diagrams ($M = 15$) for GCN, with label rate $L/C = 20$. (Left: 2-layer GCN on Citeseer and Pubmed; Right: 3-layer GCN on Citeseer and Pubmed.

bel Smoothing and Positive Label Smoothing into the framework of **U**nified **L**abel **S**moothing (ULS).

Unified Label Smoothing: In the ULS framework, $\epsilon \in [-1, 1]$ and soft labels are as follows:

$$y_k^{\epsilon} = y_k \left(1 - \epsilon\right) + \frac{\epsilon}{K} \tag{2}$$

where $\epsilon \in \begin{cases} [0, 1], \ if \text{ overconfident} \\ [-1, 0], \ if \text{ underconfident} \end{cases}$

Adaptive Unified Label Smoothing: Considering that the degree of miscalibration of the model will continue to decrease with the training process, an adaptive parameter τ is regularized to smooth parameter ϵ, $\tau = 1/(1+$ current epoch), representing the output of a harmonic series relative to the current epoch in the training process. Using a harmonic series to determine τ, gradually decreases the effect of the smooth parameter ϵ in the loss, which can accommodate the GNN's tendency to become more and more calibrated in the training process. ϵ controls the direction and degree of label smoothing: positive values for handling overconfidence, and negative values for underconfidence. γ is introduced to adaptively adjust ϵ as training progresses, which allows for finer control over the smoothing effect. The loss function of AU-LS can be formulated as follows:

$$\mathcal{L}_{\text{AU-LS}}(\mathbf{y}, \mathbf{p}) = -\sum_k y_k^{\epsilon+\gamma\tau} \log s_k \tag{3}$$

where $\gamma \in \begin{cases} [0, 1], \ if \ \epsilon \in [-1, 0] \\ [-1, 0], \ if \ \epsilon \in [0, 1] \end{cases}$

Here, soften labels $y_k^{\epsilon+\gamma\tau} = y_k(1 - (\epsilon + \gamma\tau)) + \frac{\epsilon+\gamma\tau}{K}$ replace the hard labels and parameter $\epsilon + \gamma\tau$ adaptively controls the degree of the smoothing effect.

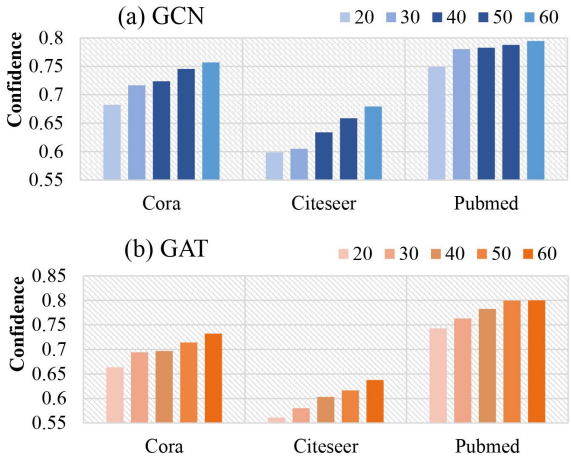


Figure 4: Confidence of 2-layer GCN and GAT with different labelrates ($L/C = 20, 30, 40, 50, 60$) of training set on different datasets (Cora, Citeseer and Pubmed).

To better understand the mechanism of AU-LS, we decompose the loss function in Eq.(3) into a standard cross-entropy term and a Kullback-Leibler (KL) divergence between uniform distribution $\mathbf{u} = \frac{1}{K}$ and the softmax prediction $\mathbf{s}$:

$$\mathcal{L}_{\text{AU-LS}}(\mathbf{y}, \mathbf{p}) \overset{c}{=} \mathcal{L}_{\text{CE}} + \frac{\epsilon + \gamma\tau}{1 - (\epsilon + \gamma\tau)} \mathcal{D}_{\text{KL}}(\mathbf{u} \| \mathbf{s}) \tag{4}$$

where $\overset{c}{=}$ stands for equality up to additive and/or nonnegative multiplicative constants.

Proposition 3. *The upper bound of the linear penalty for constraint* $\mathbf{d}(\mathbf{L}) = \mathbf{0}$ *is* $\mathcal{D}_{KL}(\mathbf{u}\|\mathbf{s})$*, and meanwhile the lower bound is the value after subtracting constant from* $\mathcal{D}_{KL}(\mathbf{u}\|\mathbf{s})$.

$$\mathcal{D}_{\text{KL}}(\mathbf{u} \| \mathbf{s}) - \log(K) \overset{c}{\leq} \frac{1}{K} \sum_k \mathbf{d}(\mathbf{L}) \overset{c}{\leq} \mathcal{D}_{\text{KL}}(\mathbf{u} \| \mathbf{s})$$

where $\overset{c}{\leq}$ stands for inequality up to an additive constant. Prop.3 means that $\mathcal{D}_{\text{KL}}(\mathbf{u}\|\mathbf{s})$ can be seen as a soft-penalty optimizer, which implicitly tackle hard equality constraint $\mathbf{d}(\mathbf{L}) = \mathbf{0}$. The proof of Prop.3 is provided in Appendix A. The reasons for miscalibration of GNNs can be mitigated are as follows:

- **Overconfidence:** $\epsilon \in [0, 1]$ and $\gamma \in [-1, 0]$ lead $\frac{\epsilon+\gamma\tau}{1-(\epsilon+\gamma\tau)}$ to be positive, thus imposing positive linear penalty $\mathbf{d}(\mathbf{L}) = \mathbf{0}$ to the loss function, which encourages equal logits, reducing overconfidence.
- **Underconfidence:** $\epsilon \in [-1, 0]$ and $\gamma \in [0, 1]$ lead $\frac{\epsilon+\gamma\tau}{1-(\epsilon+\gamma\tau)}$ to be negative, thus imposing negative linear penalty $\mathbf{d}(\mathbf{L}) = \mathbf{0}$ to the loss function, which discourages equal logits, boosting model confidence in predictions.

Experiments

To evaluate the calibration performance of our proposed method AU-LS, including calibrating overconfidence and

Dataset	L/C	GCN (2-layer)					GAT (2-layer)				
		Uncal.	TS	MS	CaGCN	Ours	Uncal.	TS	MS	CaGCN	Ours
Cora	20	0.1347	0.0488	0.0414	0.0401	**0.0356**	0.1558	0.0717	0.0544	0.0450	**0.0297**
	40	0.1134	0.0417	0.0372	0.0407	**0.0353**	0.1340	0.0485	0.0491	0.0365	**0.0291**
	60	0.0937	0.0355	0.0364	0.0376	**0.0299**	0.1201	0.0393	0.0411	0.0313	**0.0285**
Citeseer	20	0.1248	0.0641	0.0644	0.0595	**0.0435**	0.1534	0.0916	0.0633	0.0572	**0.0491**
	40	0.0957	0.0601	0.0538	0.0545	**0.0438**	0.1252	0.0797	0.0590	0.0532	**0.0494**
	60	0.0806	0.0559	0.0521	0.0546	**0.0448**	0.1090	0.0648	0.0519	0.0525	**0.0489**
Pubmed	20	0.0586	0.0541	0.0476	0.0405	**0.0376**	0.0835	0.0656	0.0501	**0.0356**	0.0471
	40	0.0444	0.0446	0.0436	0.0402	**0.0366**	0.0869	0.0658	0.0539	0.0308	**0.0255**
	60	0.0445	0.0367	0.0318	0.0311	**0.0289**	0.0993	0.0669	0.0483	0.0308	**0.0265**
CoraFull	20	0.1986	0.1013	–	0.0776	**0.0696**	0.2119	0.1101	–	0.0788	**0.0712**
	40	0.2321	0.1117	–	0.0701	**0.0565**	0.2438	0.1133	–	0.0738	**0.0679**
	60	0.2337	0.0981	–	0.0768	**0.0713**	0.2497	0.1133	–	0.0849	**0.0841**

Table 1: ECE (M=20) for various calibration methods on 2-layer GCN and GAT across datasets with label rates $L/C = 20, 40, 60$. Bold results indicate best performance; (-) denotes failure to improve model reliability.

underconfidence, we design to use commonly-used classifiers with different training set labelrates and classifiers with different number of layers to compare the results of AU-LS with that of existing SOTA methods.

Datasets and Baselines

Calibration methods are compared on four widely used citation network datasets: Cora, Citeseer, Pubmed (Sen et al. 2008), and CoraFull (Bojchevski and Günnemann 2018). The nodes in these datasets represent paper IDs, and edges represent citation links between papers. Brief statistics of these datasets are summarized in Appendix B. Note that the number of training set nodes is obtained by multiplying the number of classes with a quantity denoted as labelrate, which represents the number of labeled nodes per class.

GCNs (Kipf and Welling 2017) and GATs (Veličković et al. 2018), which are the most representative GNNs models, are chosen as classifiers in our experiments. To create scenarios where some models are underconfident and others are overconfident, we vary the labelrates ($L/C = 20, 40, 60$) of the training set and change the depth of the base classifiers. Specifically, we experiment with 2-layer and 3-layer configurations. For a fair comparison, we choose three calibration methods as baselines, including two classic post-hoc calibration methods: temperature scaling and matrix scaling (Guo et al. 2017), which adjust the model's outputs to improve calibration, and CaGCN (Wang et al. 2021), which is a method specifically designed for calibrating GNNs.

Implementation Details

Base models GCN and GAT are tuned to optimal performance by following parameters used in (Kipf and Welling 2017) and (Veličković et al. 2018). In the calibration experiments of GCN (2-layer), to be fair, we train a GCN with the same parameters as CaGCN (Wang et al. 2021), that is, the hidden layer dimension is set to 16, the number of hidden units is set to 64, learning rate (lr) is set to 0.01, weight decay of CoraFull is set to 0.03, weight decay of other three datasets is set to $5e-3$,and dropout rate is set to 0.5. Similarly, in the calibration experiments of GAT (2-layer), we employ the parameters values reported in CaGCN (Wang et al. 2021). More precisely, the number of hidden units is 8, learning rate (lr) is 0.0005, dropout rate is 0.6. We choose ECE as evaluation metric and set M as 20. The same parameter setting as 2-layer are used in 3-layer experiments. For different datasets and labelrate L/C, we apply various ϵ and γ to AU-LS and tune them to get optimal performance.

Results

Experimental results for calibration performance are in Table 1 and Table 2, showing ECE of various-layer GCN and GAT models across datasets and label rates. Uncal, TS and MS here is the abbreviation of uncalibration, Temperature Scaling and Matrix Scaling respectively.

GNNs with 2-layer: As can be seen from Table 1, compared with other calibration methods, the proposed AU-LS achieves the lowest ECE in almost all cases, indicating the effectiveness of AU-LS in calibrating underconfidence.

GNNs with 3-layer: In Table 2, we observe: (1) 3-layer GNNs are better calibrated than 2-layer ones, which can be verified mutually with the analysis of the correlations between message aggregation and confidence. Specifically, moderate message aggregation contributes to a well-calibrated model. (2) TS, MS and CaGCN fail to calibrate GCN and GAT with 3 layers in almost all datasets and labelrates. It is useless to learn a fixed transformation parameter or matrix for a well-calibrated model, because the confidence of predicted samples are balanced, with some samples are overconfident whereas the others are underconfident. This is a common problem with post-hoc calibration methods, but our AU-LS does not suffer this problem. (3) Compared to existing methods that do not achieve ideal performance, our proposed method AU-LS pushes the model towards better calibration on the basis that the base model is already well calibrated (low ECE).

AU-LS enhances classification accuracy, a feat not matched by existing post-hoc methods, detailed in Ap-

Dataset	L/C	GCN (3-layer)					GAT (3-layer)				
		Uncal.	TS	MS	CaGCN	Ours	Uncal.	TS	MS	CaGCN	Ours
Cora	20	0.0477	0.1059	0.0707	0.0521	**0.0297**	0.0478	0.0688	0.0537	0.0664	**0.0395**
	40	0.0369	0.0958	0.0553	0.0574	**0.0252**	0.0442	0.0470	0.0375	0.0485	**0.0267**
	60	0.0402	0.0822	0.0543	0.0393	**0.0389**	0.0386	0.0829	0.0495	0.0490	**0.0347**
Citeseer	20	0.0667	0.1992	0.0791	0.0848	**0.0499**	0.0798	0.0891	0.1021	0.0912	**0.0579**
	40	0.0613	0.1280	0.0661	0.0539	**0.0379**	0.0799	0.0717	0.0713	0.1042	**0.0647**
	60	0.0598	0.1200	0.0454	0.0636	**0.0383**	0.0744	0.0765	0.0599	0.0875	**0.0567**
Pubmed	20	0.0576	0.0828	0.0425	0.0389	**0.0373**	0.0593	0.0760	0.0401	0.0486	**0.0372**
	40	0.0456	0.0957	0.0437	0.0459	**0.0241**	0.0404	0.0613	0.0504	0.0360	**0.0341**
	60	0.0379	0.0768	0.0459	0.0335	**0.0292**	0.0371	0.0522	0.0450	0.0446	**0.0260**
Corafull	20	0.0767	0.2219	–	0.1221	**0.0577**	0.0932	0.1322	–	0.1132	**0.0894**
	40	0.0592	0.1288	–	0.0606	**0.0545**	0.0735	0.0979	–	0.0902	**0.0708**
	60	0.0568	0.1338	–	0.0709	**0.0378**	0.0777	0.0956	–	0.1028	**0.0690**

Table 2: ECE (M=20) for various calibration methods on 3-layer GCN and GAT across datasets with label rates $L/C = 20, 40, 60$. Bold results indicate best performance; (-) denotes failure to improve model reliability.

Dataset	L/C	LS		ULS		AU-LS	
		Module		Module		Module	
		Unified ☒	Adaptive ☒	Unified ☒	Adaptive ☑	Unified ☑	Adaptive ☑
		GCN-2	GCN-3	GCN-2	GCN-3	GCN-2	GCN-3
Cora	20	✗	✗	0.0715	0.0449	**0.0356**	**0.0297**
	40	✗	✗	0.0604	0.0262	**0.0353**	**0.0252**
	60	✗	✗	0.0524	0.0448	**0.0299**	**0.0389**
Citeseer	20	✗	0.0718	0.0488	0.0772	**0.0435**	**0.0499**
	40	✗	0.0643	0.0497	0.0617	**0.0438**	**0.0379**
	60	✗	0.0476	0.0495	0.0491	**0.0448**	**0.0383**
Pubmed	20	✗	0.0402	0.0443	0.0381	**0.0376**	**0.0373**
	40	✗	0.0598	0.0426	0.0596	**0.0366**	**0.0241**
	60	✗	0.0496	0.0311	0.0509	**0.0289**	**0.0292**
CoraFull	20	✗	0.0958	0.0725	0.0936	**0.0696**	**0.0577**
	40	✗	0.0772	0.0749	0.0743	**0.0565**	**0.0545**
	60	✗	0.0587	0.0802	0.0606	**0.0713**	**0.0378**

Table 3: Ablation Study: ECE (M=20) for LS, ULS, and AU-LS on GCN and GAT across datasets with label rates. ☒ indicates no module; ☑ indicates module included.

pendix C1. Additionally, CaGCN's post-hoc training results in longer compute times compared to AU-LS.

In summary, extensive experiments not only prove that AU-LS can deal with the issue of under-confidence and over-confidence, and achieve remarkable calibration performances, but also corroborate with our observation that the confidence of the GNN model is positively related to the amount of message aggregation.

Ablation Study

AU-LS, comprising a self-adaptive module for dynamic smooth parameter adjustment and a unified label smoothing framework for aligning confidence with accuracy, is assessed for its calibration efficacy in various configurations on 2-layer and 3-layer GCNs across datasets and label rates, as shown in Table 3.(✗) represents that the method can not calibrate the base model to be more reliable in the case of existing datasets and labelrates (further results on GAT with 2-layer are shown in Appendix C2). Due to the underconfidence of GCN with 2-layer and some cases in GCN with 3-layer, Label smoothing (LS) with positive ϵ will aggravate the lack of confidence. A tendency can be observed that the calibration performance becomes better with the addition of modules, that is, AU-LS achieve better calibration performance compared to LS and ULS, thus demonstrating every module we add to AU-LS is reasonable and helpful for model calibration.

Conclusion

In this paper, from the perspective of message passing mechanism, we bring an novel point to the community that not all GNNs are under-confident and there exists a correlation between the amount of message aggregated from neighbors and confidence. Specifically, a model aggregates more message than the threshold may suffer over-confidence, whereas insufficient message aggregation leads to under-confidence. This discovery will benefit to more effective implementation of GNNs' calibration and bring inspiration to follow-up research. To mitigate the confidence bias suffered by GNNs, we propose the method AU-LS to adaptively calibrate GNNs' confidence to be more aligned with their accuracy of the predictions. Extensive experiments demonstrate the effectiveness of AU-LS in calibrating GNNs under different conditions, including over-confidence and under-confidence. Furthermore, compared with existing state-of-the-art calibrated methods, AU-LS achieves more significant calibration performance, revealing the superiority of AU-LS.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.62306326.

References

Bojchevski, A.; and Günnemann, S. 2018. Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking. In *International Conference on Learning Representations*, 1–13.

Brier, G. W.; et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3.

Burgin, M. 2002. The essence of information: Paradoxes, contradictions, and solutions. In *Electronic Conference on Foundations of Information Science: The nature of information: Conceptions, misconceptions, and paradoxes (FIS 2002). Retrieved September*, volume 13, 2013. Citeseer.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2001. The elements of statistical learning, volume 1 Springer series in statistics Springer.

Ghahramani, Z. 2015. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553): 452–459.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9): 1–46.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, B.; Ben Ayed, I.; Galdran, A.; and Dolz, J. 2022. The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 80–88.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, 6448–6458. PMLR.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; and Dokania, P. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33: 15288–15299.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Ruiz Puentes, P.; Rueda-Gensini, L.; Valderrama, N.; Hernández, I.; González, C.; Daza, L.; Muñoz-Camargo, C.; Cruz, J. C.; and Arbeláez, P. 2022. Predicting target–ligand interactions with graph convolutional networks for interpretable pharmaceutical discovery. *Scientific reports*, 12(1): 1–17.

Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI magazine*, 29(3): 93–93.

Shen, J.; Zhen, X.; Worring, M.; et al. 2023. Episodic Multi-Task Learning with Heterogeneous Neural Processes. *arXiv preprint arXiv:2310.18713*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Teixeira, L.; Jalaian, B.; and Ribeiro, B. 2019. Are graph neural networks miscalibrated? *arXiv preprint arXiv:1905.02296*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph attention networks. In *ICLR*.

Wang, M.; Yang, H.; and Cheng, Q. 2022. GCL: Graph Calibration Loss for Trustworthy Graph Neural Network. In *Proceedings of the 30th ACM International Conference on Multimedia*, 988–996.

Wang, Q.; Federici, M.; and van Hoof, H. 2022. Bridge the Inference Gaps of Neural Processes via Expectation Maximization. In *The Eleventh International Conference on Learning Representations*.

Wang, Q.; Lv, Y.; Feng, Y.; Xie, Z.; and Huang, J. 2023. A Simple Yet Effective Strategy to Robustify the Meta Learning Paradigm. *arXiv preprint arXiv:2310.00708*.

Wang, Q.; and Van Hoof, H. 2020. Doubly stochastic variational inference for neural processes with hierarchical latent variables. In *International Conference on Machine Learning*, 10018–10028. PMLR.

Wang, Q.; and van Hoof, H. 2022. Learning expressive meta-representations with mixture of expert neural processes. *Advances in neural information processing systems*, 35: 26242–26255.

Wang, Q.; and Van Hoof, H. 2022. Model-based meta reinforcement learning using graph structured surrogate models and amortized policy search. In *International Conference on Machine Learning*, 23055–23077. PMLR.

Wang, X.; Liu, H.; Shi, C.; and Yang, C. 2021. Be confident! towards trustworthy graph neural networks via confidence calibration. *Advances in Neural Information Processing Systems*, 34: 23768–23779.

Yang, H.; Wang, M.; Yu, Z.; and Zhou, Y. 2023. A Simple Stochastic Neural Network for Improving Adversarial Robustness. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2297–2302. IEEE.

Yang, H.; Wang, M.; Zhou, Y.; and Yang, Y. 2021. Towards Stochastic Neural Network via Feature Distribution Calibration. In *2021 IEEE International Conference on Data Mining (ICDM)*, 1445–1450. IEEE.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, J.; Shi, X.; Xie, J.; Ma, H.; King, I.; and Yeung, D. Y. 2018. GaAN: Gated Attention Networks for Learning on Large and Spatiotemporal Graphs. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*.

Zhou, F.; Yang, Q.; Zhong, T.; Chen, D.; and Zhang, N. 2020. Variational graph neural networks for road traffic prediction in intelligent transportation systems. *IEEE Transactions on Industrial Informatics*, 17(4): 2802–2812.