Topology-Imbalance Learning for Semi-Supervised Node Classification

Deli Chen\textsuperscript{1,2}, Yankai Lin\textsuperscript{1}, Guangxiang Zhao\textsuperscript{2}, Xuancheng Ren\textsuperscript{2}, Peng Li\textsuperscript{1}, Jie Zhou\textsuperscript{1}, Xu Sun\textsuperscript{2}

\textsuperscript{1}Pattern Recognition Center, WeChat AI, Tencent Inc., China
\textsuperscript{2}MOE Key Lab of Computational Linguistics, School of EECS, Peking University
\{delichen, yankailin, patrickpli, withtomzhou\}@tencent.com
\{zhaoguangxiang,renxc,xusun\}@pku.edu.cn

Abstract

The class imbalance problem, as an important issue in learning node representations, has drawn increasing attention from the community. Although the imbalance considered by existing studies roots from the unequal quantity of labeled examples in different classes (quantity imbalance), we argue that graph data expose a unique source of imbalance from the asymmetric topological properties of the labeled nodes, i.e., labeled nodes are not equal in terms of their structural role in the graph (topology imbalance). In this work, we first probe the previously unknown topology-imbalance issue, including its characteristics, causes, and threats to semi-supervised node classification learning. We then provide a unified view to jointly analyzing the quantity- and topology-imbalance issues by considering the node influence shift phenomenon with the Label Propagation algorithm. In light of our analysis, we devise an influence conflict detection-based metric Totoro to measure the degree of graph topology imbalance and propose a model-agnostic method ReNode to address the topology-imbalance issue by re-weighting the influence of labeled nodes adaptively based on their relative positions to class boundaries. Systematic experiments demonstrate the effectiveness and generalizability of our method in relieving topology-imbalance issue and promoting semi-supervised node classification. The further analysis unveils varied sensitivity of different graph neural networks (GNNs) to topology imbalance, which may serve as a new perspective in evaluating GNN architectures.

1 Introduction

Graph is a widely-used data structure\textsuperscript{51}, where the nodes are connected to each other through natural or handcrafted edges. Similar to other data structures, the representation learning for node classification faces the challenge of quantity-imbalance issue, where the labeling size varies among classes and the decision boundaries of trained classifiers are mainly decided by the majority classes\textsuperscript{46}. There have been a series of studies\textsuperscript{45}\textsuperscript{11}\textsuperscript{49} handling the Quantity-Imbalance Node Representation Learning (short as QINL). However, different with other data structures, graph-structured data suffers from another aspect of the imbalance problem: the imbalance caused by the asymmetric and uneven topology of labeled nodes, where the decision boundaries are driven by the labeled nodes close to the topological class boundaries (left of Figure[1]) thus interfering with the model learning.

Present Work. For the first time, we recognize the **Topology-Imbalance Node Representation Learning** (short as TINL) as a graph-specific imbalance learning topic, which mainly focus on the

\footnote{\textsuperscript{1}The code is available at \url{https://github.com/victorchen96/ReNode}}
Figure 1: Schematic diagram of the topology-imbalance issue in node representation learning. The color and the hue denote the type and the intensity of each node's received influence from the labeled nodes, respectively. The left shows that nodes close to the boundary have the risk of information conflict and nodes far away from labeled nodes have the risk of information insufficient. The right shows that our method can decrease the training weights of labeled nodes (R1) close to the class boundary and increase the weights of labeled nodes (B and R2) close to the class centers, thus relieving the topology-imbalance issue.

Decision boundaries shift phenomena driven by the topology imbalance in graph and is an essential component for node imbalance learning. Comparing with the well-explored QINL that studies the imbalance caused by the numbers of labeled nodes, TINL explores the imbalance caused by the positions of labeled nodes and owns the following characteristics:

- **Ubiquity**: Due to the complex connections of the graph nodes, the topology structure of nodes in different categories is naturally asymmetric, which makes TINL an essential characteristic in node representation learning. Hence, it is difficult to construct a completely symmetric labeling set even with an abundant annotation budget.

- **Perniciousness**: The influence from labeled nodes decays with the topology distance [3]. The asymmetric topology of labeled nodes in different classes and the uneven distribution of labeled nodes in the same class will cause the influence conflict and influence insufficient problems (left of Figure 1) respectively, resulting in a shift of decision boundaries.

- **Orthogonality**: Quantity-imbalance studies [49, 8, 5] usually treat the labeled nodes of the same class as a whole and devise solutions based on the total numbers of each class, while TINL explores the influence of the unique position of each labeled node on decision boundaries. Thus, TINL is independent of QINL in terms of the object of study.

Exploring TINL is of great importance for node representation learning due to its ubiquity and perniciousness. However, the methods [17, 22] for quantity imbalance can be hardly applied to TINL because of the orthogonality. To remedy the topology-imbalance issue, thus promoting the node classification, we propose a model-agnostic training framework ReNode to re-weight the labeled nodes according to their positions. We devise the conflict detection-based Topology Relative Location (Totoro) metric to leverage the interaction among labeled nodes across the whole graph to locate their structural positions. Based on the Totoro metric, we further increase the training weights of nodes with small conflict that are highly likely to be close to topological class centers to make them play a more pivotal role during training, and vice versa (right of Figure 1). Empirical results of various imbalance scenarios (TINL, QINL, large-scale graph) and multiple graph neural networks (GNNs) demonstrate the effectiveness and generalizability of our method. Besides, we provide the sensitivity to topology imbalance as a new evaluation perspective for different GNN architectures.

2 Topology-Imbalance Node Representation Learning

2.1 Notations and Preliminary

In this work, we follow the well-established semi-supervised node classification setting [47, 18] to conduct analyses and experiments. Given an undirected and unweighted graph $G = (V, E, L)$, where $V$ is the node set represented by the feature matrix $X \in \mathbb{R}^{n \times d}$ ($n = |V|$ is the node size and $d$ is the node embedding dimension), $E$ is the edge set which is represented by an adjacency matrix $A \in \mathbb{R}^{n \times n}$, $L \subset V$ is the labeled node set and usually we have $|L| \ll |V|$, the node classification
Figure 2: Node influence and boundary shift caused by quantity- and topology-imbalance. (a): The prediction results of GCN and LP are highly consistent (t-SNE [39] visualization of the CORA dataset). (b): The node influence boundary (the yellow dotted line) is shifted towards the small class from the true class boundary (the black dotted line) under the quantity- and topology-imbalance scene. (c): The node influence boundary is shifted towards the large class under the quantity-balanced, topology-imbalanced scene. We regard the large class as positive class to indicate the results.

2.2 Understanding Topology Imbalance via Label Propagation

From Figure 1, we can intuitively perceive the imbalance brought by the positions of labeled nodes; in this part, we further explore the nature of topology imbalance with the well-known Label Propagation [50] algorithm (short as LP) and provide a uniform analysis framework for the comprehensive node imbalance issue. In LP, labels are propagated from the labeled nodes and aggregated along edges, which can also be viewed as a random walk process from labeled nodes. The convergence result $Y$ after repeated propagation is regarded as the nodes soft-labels:

$$Y = \alpha (I - (1 - \alpha)A')^{-1}Y^0,$$

where $I$ is the identity matrix, $\alpha \in (0, 1]$ is the random walk restart probability, $A' = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is the adjacency matrix normalized by the diagonal degree matrix $D$, $Y^0$ is the initial label distribution where labeled nodes are represented by the one-hot vectors. The prediction label for the $i$-th node is $q_i = \arg \max_j Y_{ij}$. LP is a simple yet successful model [37] and can be unified with GNN models owning the message-passing mechanism [41]. From Figure 2(a), we can empirically find that there is a significant correlation between the results of LP and GCN (T/F indicates prediction is True/False).

The LP prediction $q$ can be viewed as the distribution of the (labeled) node influence [41] (i.e., each node is mostly influenced by which class’s information); hence the boundaries of the node influence can act as an effective reflection for the GNN model decision boundaries considering the high consistency between LP and GNN. Moreover, node influence offers a unified view of TINL and QINL: ideally, the node influence boundaries should be consistent with the true class boundaries, but both the labeled nodes’ numbers (QINL) and positions (TINL) can cause a shift of the node influence boundaries from the true one, resulting in deviation of the model decision boundaries.

Node imbalance issue is composed of topology- and quantity-imbalance. Figure 2 illustrates two examples of node influence boundary shift. In Figure 2(b), when the uniform selection is adopted to generate training set, both the quantity and the topology are imbalanced for model training; then the large class with more total nodes (denotes by blue color) will own stronger influence than the small class with fewer total nodes (denotes by red color) due to the quantity advantage and the node influence boundary is shifted towards the small class. In Figure 2(c), when the quantity-balanced strategy is adopted for sampling training nodes, it will be easier for the small class to has more labeled nodes close to the class boundary and the boundary of the node influence is shifted into the large class. We can find that even when the training set is quantity-balanced, the topology-imbalance issue still exists and hinders the node classification learning. Hence, we can conclude that node imbalance
We visualize the node labels and the Totoro values (scaled to where $y$ is the ground-truth label of node $v$, $P_v$ indicates the personalized PageRank probability vector for the node $v$. A larger Totoro value $T_v$ indicates that node $v$ is topologically closer to class boundaries, and vice versa. The normalization item $1/|C_j|$ is added to make the influence from the different classes comparable when computing conflict.

We visualize the node labels and the Totoro values (scaled to $[0, 1]$) of labeled nodes in Figure 3(a). We can find that the labeled nodes with smaller Totoro values are farther away from the class boundaries, demonstrating the effectiveness of Totoro in locating the positions of labeled nodes. Besides, we sum learning is caused by the joint effect of TINL and QINL. Separately considering TINL or QINL will lead to a one-sided solution to node imbalance learning.

2.3 Measuring Topology Imbalance by Influence Conflict

Although we have realized that the imbalance of node topology interferes with model learning, how to measure the labeled node’s relative topological position to its class (being far away from or close to the class center) remains the key challenge in handling the topology-imbalance issue due to the complex graph connections and the unknown class labels for most nodes in the graph. As the nodes are homogeneously connected when constructing the graph, even nodes close to the class boundaries own similar characteristics to their neighbors. Thus it is unreliable to leverage the difference between the characteristics of one labeled node and its surrounding subgraphs to locate its topological position. Instead, we propose to utilize the node topology information by considering the node influence conflict across the whole graph and devise the Conflict Detection-based Topology Relative Location metric (Totoro).

Similar to Eq (1), we calculate the Personalized PageRank matrix $P$ to measure node influence distribution from each labeled node:

$$P = \alpha(I - (1 - \alpha)A')^{-1}. \quad (2)$$

Node influence conflict denotes topological position. According to related studies [41, 19, 2], $P$ can be viewed as the distribution of influence exerted outward from each node. We assume that if a labeled node $v \in V$ encounters strong heterogeneous influence from the other classes’ labeled nodes in the subgraph around node $v$ where node $v$ itself owns great influence, we have the conclusion that node $v$ meets large influence conflict in message passing and it is close to topological class boundaries, and vice versa. Based on this hypothesis, we take the expectation of the influence conflict between the node $v$ and the labeled nodes from other classes when node $v$ randomly walks across the entire graph as a measurement of how topologically close node $v$ is to the center of the class it belongs to. The Totoro value of node $v$ is computed as:

$$T_v = \mathbb{E}_{x \sim P_x} \left[ \sum_{j \in [1, k], j \neq y_v} \frac{1}{|C_j|} \sum_{i \in C_j} P_{i,x} \right], \quad (3)$$

where $y_v$ is the ground-truth label of node $v$, $P_v$ indicates the personalized PageRank probability vector for the node $v$. A larger Totoro value $T_v$ indicates that node $v$ is topologically closer to class boundaries, and vice versa. The normalization item $1/|C_j|$ is added to make the influence from the different classes comparable when computing conflict.

We visualize the node labels and the Totoro values (scaled to $[0, 1]$) of labeled nodes in Figure 3(a). We can find that the labeled nodes with smaller Totoro values are farther away from the class boundaries, demonstrating the effectiveness of Totoro in locating the positions of labeled nodes. Besides, we sum
the conflict of all the labeled nodes $\sum_{v \in \mathcal{L}} T_v$ to measure the overall conflict of the dataset, which can be viewed as the metric for the overall topology imbalance given the graph $\mathcal{G}$ and the training set $\mathcal{L}$. Figure 3(b) shows that there is a significant negative correlation between the overall conflict and the model performance, which further demonstrates the effectiveness of Totoro in measuring the intensity of topology imbalance at the dataset level.

### 2.4 Alleviate Topology Imbalance by Instance-wise Node Re-weighting

In this section, we introduce ReNode, a model-agnostic training weight schedule mechanism to address TINL for general GNN encoder in a plug-and-play manner. Inspired by the analysis in Section 2.2, the ReNode method is devised to promote the training weights of the labeled nodes that are close to the topological class centers, so as to make these nodes play a more active role in model learning, and vice versa. Specifically, we devise a cosine annealing mechanism for the training node weights based on their Totoro values:

$$w_v = w_{\text{min}} + \frac{1}{2}(w_{\text{max}} - w_{\text{min}})(1 + \cos(\frac{\text{Rank}(T_v)}{|\mathcal{L}|}\pi)), \quad v \in \mathcal{L}$$  \hspace{1cm} (4)

where $w_v$ is the modified training weight for the labeled node $v$, $w_{\text{min}}$, $w_{\text{max}}$ are the hyper-parameters indicating the lower bound and upper bound of the weight correction factor, $\text{Rank}(T_v)$ is the ranking order of $T_v$ from the smallest to the largest. The training loss $L_T$ for the quantity-balanced, topology-imbalanced node classification task is computed by the following equations:

$$L_T = -\frac{1}{|\mathcal{L}|} \sum_{v \in \mathcal{L}} w_v \sum_{c=1}^{k} y^c_v \log g^c_v, \quad g = \text{softmax}(\mathcal{F}(X, A, \theta)),$$  \hspace{1cm} (5)

where $\mathcal{F}$ denotes any GNN encoder, $\theta$ is the parameter of $\mathcal{F}$, $g_i$ is the GNN output for node $i$, $y^c_i$ is the gold label for node $i$ in one-hot embedding. By encouraging the positive effects of the labeled nodes near the class topological centers, and reducing the negative effects of those near the topological class boundaries, our ReNode method is expected to minimize the deviation between the node influence boundaries and the true class boundaries, so as to correct the class imbalance caused by the positions of labeled nodes.

**ReNode to Jointly Handle TINL and QINL.** In this part, we introduce the application of the ReNode method in a more general graph imbalance scenario where both the topology- and quantity-imbalance issues exist. As analyzed in previous sections, the TINL and QINL are orthogonal problems. Therefore, we propose that our ReNode method based on (labeled) node topology can be seamlessly combined with the existing methods designed for the quantity-imbalance learning. Without loss of generality, we present how our ReNode method can be combined with the vanilla $\hat{P}$ and the corresponding training ReNode factor $\hat{w}$ are generated by the estimation method from Andersen et al. \cite{PPRGo} and then $\hat{P}$ is

---

\*\*We also try other schedules and the cosine method works best. More details can be found in Appendix B.\*\*
directly employed as the aggregation weights from all the other nodes regardless of their topology distance from the current node:

\[ g' = \text{softmax}(\tilde{\mathbf{P}}\mathcal{F}'(\mathbf{X}, \theta')), \tag{7} \]

where \( \mathcal{F}' \) can be a linear layer or a multi-layer perceptron with parameter \( \theta' \). The final training loss for large-scale graph \( L_L \) follows Eq (5) and (6), and replaces \( w \) and \( g \) with \( \tilde{w} \) and \( g' \).

### 3 Experiments

In this section, we will first introduce the experimental datasets for both transductive and inductive semi-supervised node classification. Then we introduce the experiments to verify the effectiveness of the proposed ReNode method in three different imbalance situations: (1) TINL only, (2) TINL and QINL, (3) Large-scale Graph.

#### 3.1 Datasets

We adopt two sets of graph datasets to conduct experiments. For the transductive setting [13], we take the widely-used Plantoid paper citation graphs [33] (CORA, CiteSeer, Pubmed) and the Amazon co-purchase graphs [24] (Photo, Computers) to verify the effectiveness of our method. For the inductive setting, we conduct experiments on the popular Reddit dataset [13] and the enormous MAG-Scholar dataset (coarse-grain version) [2] which owns millions of nodes and features. For each of these datasets, we repeat experiments on 5 different datasets splittings [34] and we run 3 times for each splitting to reduce the random variance. More details about the datasets and experiment settings are presented in Appendix A.

#### 3.2 ReNode for the Pure Topology-imbalance Issue

**Settings** When considering topology-imbalance only, the labeling set takes a balanced setting and the annotation size for each class is all equal to \( |\mathcal{L}|/k \). Following the most widely-used semi-supervised setting in node classification studies [47][18], we randomly select 20 nodes in each class for training and 30 nodes per class for validation: all the remaining nodes form the test set. We display the experiment results for the 5 transductive datasets on 6 widely-used GNN models: GCN [18], GAT [40], PPNP [19], GraphSAGE [13] (short as SAGE), ChebGCN [9] (short as CHEB) and SGC [43]. We strictly align the hyperparameters in each group of experiments to show the pure improvement brought by our ReNode method (similarly hereinafter). The training loss \( L_T \) from section 2.4 is adopted.

Table 1: Result of different dataset conflict levels (High/Middle/Low). Our ReNode method improve the GNN (GCN) performance most when the conflict level of graph is high.

Table 2: Result of different dataset conflict levels (High/Middle/Low). Our ReNode method improve the GNN (GCN) performance most when the conflict level of graph is high.
Table 3: ReNode method for the compound scene of TINL and QINL. The imbalance ratio $\rho$ is set to different levels ([5, 10]) to test the effect of our method under different imbalance intensities.

<table>
<thead>
<tr>
<th>Imbalance Ratio</th>
<th>CORA</th>
<th>CiteSeer</th>
<th>PubMed</th>
<th>Photo</th>
<th>Computers</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>60.9±1.2</td>
<td>41.0±1.0</td>
<td>53.6±2.5</td>
<td>47.6±2.8</td>
<td>61.0±1.4</td>
</tr>
<tr>
<td>DR-GCN</td>
<td>67.7±1.5</td>
<td>51.3±1.4</td>
<td>54.7±2.5</td>
<td>52.5±2.6</td>
<td>79.4±2.7</td>
</tr>
<tr>
<td>RA-GCN</td>
<td>69.0±1.5</td>
<td>51.7±1.7</td>
<td>55.6±3.1</td>
<td>52.7±2.1</td>
<td>80.6±1.8</td>
</tr>
<tr>
<td>G-SMOTE</td>
<td>68.1±1.0</td>
<td>49.6±1.2</td>
<td>54.0±4.6</td>
<td>51.8±1.3</td>
<td>79.7±2.2</td>
</tr>
<tr>
<td>RW (w/o RN)</td>
<td>70.0±1.5</td>
<td>50.1±1.5</td>
<td>55.2±3.1</td>
<td>54.0±3.1</td>
<td>81.2±1.5</td>
</tr>
<tr>
<td>RW (w/ RN)</td>
<td>66.4±4.4</td>
<td>51.9±1.3</td>
<td>54.3±3.1</td>
<td>54.0±3.0</td>
<td>80.5±2.7</td>
</tr>
<tr>
<td>FOCAL (w/o RN)</td>
<td>68.7±3.7</td>
<td>52.6±1.9</td>
<td>54.6±2.9</td>
<td>54.7±2.7</td>
<td>80.9±3.0</td>
</tr>
<tr>
<td>FOCAL (w/ RN)</td>
<td>71.1±4.4</td>
<td>51.9±1.2</td>
<td>54.1±2.0</td>
<td>53.5±2.6</td>
<td>80.6±2.8</td>
</tr>
</tbody>
</table>

Results From Table[1] we can find that our ReNode method can effectively improve the overall performance (Weighted-F1) and the class-balance performance (Macro-F1) for all the 6 experiment GNNs in most cases, which proves the effectiveness and generalizability of our method. Our method considers the graph-specific topology imbalance issue which has been usually neglected in existing methods and conducts a fine-grained and self-adaptive adjustment to the training node weights based on their topological positions. We notice that the improvement for the CiteSeer dataset is less than the other datasets. We analyze the reason lies in that the connectivity of CiteSeer is poor, which makes the conflict detection–based method fail to reflect the node topological position well. To verify the motivation of relieving topology-imbalance, we set training sets with different levels of topology-imbalance to test our method[1]. Table[2] displays that our ReNode method improves the performance of GNN (GCN) most when the dataset is highly topologically imbalanced, which demonstrates that our method can effectively alleviate topology-imbalance and improve GNN performance.

3.3 ReNode for the Compound Scene of TINL and QINL

Settings When jointly considering both topology- and quantity-imbalance issues, following existing studies [5, 4], we take the step imbalance setting, in which all the minority classes have the same labeling size $n_i$ and all the majority classes have the same labeling size $n_a = \rho \cdot n_i$. The imbalance ratio $\rho$ denotes the intensity of quantity imbalance which is equal to the ratio of the node size of the most frequent to least frequent class. In this work, the imbalance ratio $\rho$ is set to [5, 10] for each dataset. The fraction of the majority classes is $\mu$, and for all experiments, we set $\mu = 0.5$ and round down the result $\mu \cdot k$. The training loss $L_Q$ from section [2.4] is adopted. We implement two groups of baselines for comparison: (1) Popular quantity-imbalance methods for general scenarios: Re-weight (RW), Focal Loss (Focal) and Class Balanced Loss (CB); (2) Graph-specific quantity-imbalance methods: DR-GCN [35], RA-GCN [11] and GraphSMOTE [49]. To jointly handle the topology- and quantity-imbalance issues and demonstrate the orthogonality of them, we combine our ReNode method with these three general quantity-imbalance methods (RW, Focal, CB). The backbone model is GCN [18], and the labeling ratio $\delta$ is set to 5%.

Results From Table[3] (Macro-F1) is reported here for a fair comparison with these methods designed for class-balance performance, we can find that our ReNode method significantly outperforms both the general and the graph-specific quantity-imbalance methods in most situations by simultaneously alleviating the topology- and quantity-imbalance issues. Even when the training set is severely quantity-imbalanced ($\rho=10$), our method still effectively alleviates the imbalance issue and promotes model performance well. The performance of the quantity-imbalance methods from the general field (RW, Focal, CB) is on par with or less effective than the graph-specific quantity-imbalance methods (DR-GCN, RA-GCN, G-SMOTE), while the combination of our ReNode method and these general quantity-imbalance methods can surpass the graph-specific quantity-imbalance methods, which demonstrates that the node imbalance learning can be further solved by jointly handling the topology- and quantity-imbalance issues instead of considering the quantity-imbalance issue only.

The settings of the dataset topology-imbalance levels is shown in Appendix C. The three graph-specific methods have special operations for the training loss, which can be hardly combined with our method.
Figure 4: Experimental results (Weighted-F1,%) on the large-scale Reddit and MAG-Scholar graphs. Our ReNode method can effectively improve the model performance under different labeling sizes.

Figure 5: Evaluating GNNs from the aspect of topology-imbalance sensitivity (Metric: Weighted-F1 (%)). We can summarize the ranking of topology-imbalance sensitivity: GCN > PPNP > GAT.

3.4 ReNode for Large-scale Graphs

Settings We conduct experiments on the two large-scale datasets: Reddit and MAG-Scholar, to verify the effectiveness of our ReNode method in the inductive setting. We conduct experiments with different labeling sizes (20/50/100 training nodes per class) and imbalance settings (TINL-only, TINL and QINL). The backbone GNN model is PPRGo. For QINL, we take the uniform selection to sample training nodes to be consistent with PPRGo. The training loss $L_L$ from section 2.4 is adopted. Both baseline and our methods are not combined with any quantity-imbalance method.

Results In Figure 4, we present the experiment results with different labeling sizes and imbalance settings, we can find that our method can effectively promote the performance on the large-scale graphs comparing to the popular PPRGo model across different settings, which demonstrates the applicability of our method for extremely-large graphs. We also notice that our method can bring greater improvement when the labeling size is large. We explain the reason lies in that when the labeling size is large, the positions located by the conflicts among nodes will be more accurate, thus bringing more reasonable weight adjustments. On the other hand, when the labeling ratio is extremely small (especially for the enormous MAG-Scholar graph) and the influence conflict between the labeled nodes is negligible, our method exhibits the cold start problem.

4 Discussions

4.1 Evaluating GNNs from the Aspect of Topology-Imbalance Sensitivity

In Figure 5, we evaluate the GNN’s capability for handling topology-imbalance and find that different GNNs present significant difference in the topology-imbalance sensitivity across multiple datasets. The GCN model is susceptible to the topology-imbalance level of the graph and its performance decays greatly when the topology-imbalance increases. On the opposite, the GAT model is less sensitive to the topology-imbalance level and can achieve the best results when the topology-imbalance level is high. The PPNP model can achieve ideal performance when the topology-imbalance level is low, and its performance does not drop as sharply as GCN when the topology-imbalance level is high. We analyze the reason lies in that: (1) the aggregation operation of GCN is equivalent to directly

---

3We do not implement other inductive backbone GNN models like Cluster-GCN or APPNP due to the low efficiency of them when handling the enormous MAG-Scholar dataset.
averaging neighbor features \cite{45} that lacks the noise filtering mechanism, so it is more sensitive to the topology-imbalance level of the graph; (2) the GAT model can dynamically adjust the aggregation weight from different neighbors, which increases its robustness to the high topology-imbalance situation but hinders the model performance when the graph topology-imbalance level is low and there is less need to filter neighbor information; (3) the infinite convolution mechanism of the PPNP model makes it possible to aggregate the information from distant nodes to enhance its robustness to the graph topology imbalance.

Shchur et al. \cite{34} notice that the performance ranking of GNNs varies with the training set selection. Hence, existing node classification studies \cite{32, 14} usually repeat experiments multiple times with different training sets to reduce this randomness. The results from Figure 5 inspire us that the topology imbalance can partly explain the randomness of GNN performance caused by the training set selection and we can adopt the topology-imbalance sensitivity as a new aspect in evaluating the performance of different GNN architectures.

4.2 Limitations of Method

Although our ReNode method has proven effective in multiple scenarios, we also notice some limitations of it because of the complexity of node imbalance learning. First, the ReNode method is devised for homogeneously-connected graphs (linked nodes are expected to be similar, such as the various datasets in experiments), and it needs a further update for heterogeneously-connected graphs (such as protein networks). Besides, the ReNode method improves less when the graph connectivity is poor (Section 3.2) or the labeling ratio is extremely low (Section 3.3) because in these cases, the conflict level among nodes is low thus the nodes topological positions are insufficiently reflected.

5 Related Work

Imbalanced classification problems are widespread in real scenarios and have attracted extensive attention from both academia and industry. Most existing studies on this topic focus on the class-imbalanced quantity distribution \cite{15}, where the model’s inference ability for the majority classes will be significantly better than that of minority classes \cite{12}. The existing methods for solving the quantity-imbalance issue can be roughly divided into methods for the data selection phase and the model training phase. Active learning \cite{31, 10, 42} and Re-sampling \cite{6, 16, 25} are two classical examples designed to construct a quantity-balanced training set. On the other hand, Re-weighting is a simple but effective solution for the model training phase, which adjusts the weights of training samples in different classes based on the labeling sizes \cite{17, 30, 8, 5}. However, directly applying these methods into the graph scene lacks the consideration for the graph-specific topology-imbalance issue. Unlike the re-weight methods which conduct class-lever re-weighting, our ReNode method is a more fine-grained one and assign weights to each node individually.

There have been quantity-imbalance studies (Tomek links \cite{38}, NearMiss \cite{23}, One-Sided Selection \cite{20}) trying to exclude the negative influence of labeling samples close to class boundaries by measuring the similarity of sample features. However, in the graph scene, the prior knowledge contained in node connections is more reliable than directly calculating the feature similarity. Besides, the number of labeled nodes is quite small in the semi-supervised setting. Thus it is not robust to locate their positions by computing similarity among a small number of nodes and we propose to leverage the influence conflict across the whole graph to locate node position to boundaries.

Graph data structure owns a wide range of applications, such as social media \cite{13}, stock exchange \cite{21}, shopping \cite{34}, medicine \cite{44}, transportation \cite{28} and so on. Similar to other data structures, graph node representation learning also suffer from the quantity-imbalance issue \cite{35}. Apart from the universal quantity-balance approaches introduced in Section 5 which can be transferred to the graph scene, there are some graph-specific quantity-imbalance methods recently proposed. DR-GCN \cite{35} propose two types of regularization to tackle quantity imbalance: class-conditioned adversarial training and unlabeled nodes latent distribution constraint. RA-GCN \cite{11} propose to automatically learn to weight the training samples in different classes in an adversarial training manner. AdaGCN Shi et al. \cite{36} propose to leverage the boosting algorithm to handle the quantity-imbalance issue for the node classification task. GraphSMOTE \cite{49} combines the synthetic node generation and the edge generation to up-sample nodes for the minority classes. However, these studies only pay attention to the quantity imbalance and overlook the topology imbalance.

9
Different from these studies [48, 29, 26] that try to locate the absolute positions for all the nodes by measuring their distance from the selected anchor nodes, our Totoro metric is devised to locate the relative positions to the class boundary for the labeling nodes by considering the influence conflict and can get rid of the dependence on the anchor nodes. Besides, our relative positions can more accurately reflect node class information because we distinguish the information from different classes while existing studies [48, 29] treat all the anchor nodes the same and ignore the class difference.

6 Conclusion and Future Work

In this work, we recognize the topology-imbalance node representation learning (TINL) as a graph-specific imbalance learning problem that has not been studied so far. We find that the topology-imbalance issue widely exists in graphs and severely hinders the learning of node classification. We unify TINL with the quantity-imbalance node representation learning (QINL) by considering the shift of the node influence boundaries from true class boundaries. To measure the degree of topology imbalance, we devise a conflict detection–based metric Totoro to locate node position, and further propose the ReNode method to adaptively adjust the training weights of labeled nodes based on their topological positions. Extensive empirical results have verified the effectiveness of our method in various settings: TINL-only, both TINL and QINL, and large-scale graph. Besides, we also propose the topology-imbalance sensitivity as a new metric to evaluate GNNs.

Considering the importance of the topology-imbalance issue and the limitations of our approach, advanced methods with stronger theoretical or experimental support are expected in future work. Moreover, since topology imbalance is widespread in graph-related tasks other than node classification, how to measure and solve the topology-imbalance issues in broader graph scopes remains a meaningful challenge for future study.

7 Acknowledgement

We appreciate all the thoughtful and insightful suggestions from reviews. This work was supported in part by a Tencent Research Grant and National Natural Science Foundation of China (No. 61673028). Xu Sun is the corresponding author of this paper.

References


