# RIM: Reliable Influence-based Active Learning on Graphs

Wentao Zhang, Yexin Wang,
Zhenbang You, Meng Cao,
Ping Huang, Jiulong Shan,
Zhi Yang, Bin Cui

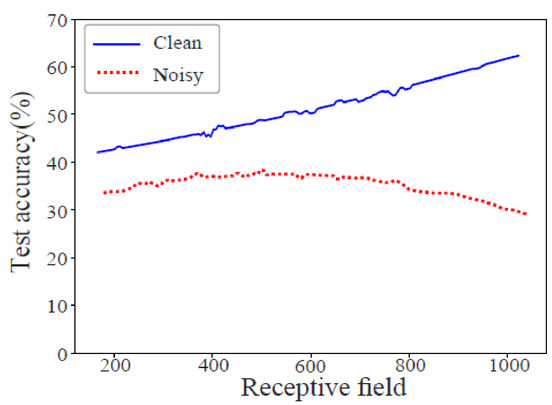
NEURAL INFORMATION PROCESSING SYSTEMS

## 1. Abstract

Message passing is the core of most graph models such as Graph Convolutional Network (GCN) and Label Propagation (LP), which usually require a large number of clean labeled data to smooth out the neighborhood over the graph. However, the labeling process can be tedious, costly, and error-prone in practice. In this paper, we propose to unify active learning (AL) and message passing towards minimizing labeling costs, e.g., making use of few and unreliable labels that can be obtained cheaply. We make two contributions towards that end. First, we open up a novel perspective by drawing a connection between AL enforcing message passing and social influence maximization, ensuring that the selected samples effectively improve the model performance. Second, we propose an extension to the influence model that incorporates an explicit quality factor to model label noise. In this way, we derive a fundamentally new AL selection criterion for GCN and LP–reliable influence maximization (RIM)–by considering both quantity and quality of influence simultaneously. Empirical studies on public datasets demonstrate that RIM significantly outperforms the state-of-the-art AL methods in terms of accuracy and efficiency.

## 2. Introduction

Graph Convolutional Neural Network (GCN) and Label Propagation (LP) [are classic message passing algorithm for graph data and typically requires a large amount of labeled data to achieve satisfactory performance. However, labeling data, be it by specialists or crowd-sourcing, often consumes too much time and money. The process is also tedious and error-prone. As the figure shows, if the label is noisy, the performance could even drop with the increase of node influence. So it is desirable to achieve good classification results with labeled data that is both few and unreliable, and we propose our method RIM to solve it.




The influence between feature propagation and test accuracy with clean/noisy label

## 3. Proposed Framework

This section presents RIM, the first graph-based AL framework that considers both the influence quality and influence quantity. At each batch of node selection, RIM first measures the proposed reliable influence quantity and selects a batch of nodes that can maximize the number of activated nodes, and then it updates the influence quality for the next iteration. The above process is repeated until the labeling budget B runs out. We will introduce each component of RIM in detail below.

### 3.1 Influence Propagation

We measure the feature/label influence of a node $v_i$ on $v_j$ by how much change in the input feature/label of $v_i$ affects the aggregated feature/label of $v_j$after $k$ iterations propagation.

### 3.2 Influence Quality Estimation

We further estimate the label reliability and associate it with influence quality. The intuition behind our method is to exploit the assumption that labels and features vary smoothly over the edges of the graph; in other words, nodes that are close in feature space and graph structure are likely to have the same label. So we can recursively infer the newly selected node's quality based on the features/labels of previously selected nodes.

### 3.3 Reliable Influence

Different from the original social influence method which only considers the influence magnitude we measure the influence quantity and get the reliable influence quantity by introducing the influence quality since it is common to have noisy oracles in the labelling process of Active Learning.

### 3.4 Node Selection and Model Training

In order to increase the feature/label influence (smoothness) effect on the graph, we should select nodes that can influence more unlabeled nodes. Due to the impact of the graph structure, the speed of expansion or, equivalently, growth of the influence can change dramatically given different sets of label nodes. This observation motivates us to address the graph data selection problem in the viewpoint of influence maximization.

**Algorithm 1:** Batch Node Selection.

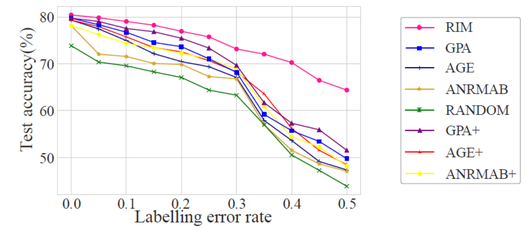
**Input:** Initial labeled set $\mathcal{V}_0$, query batch size $b$, and labelling accuracy $\alpha$.
**Output:** Labeled set $\mathcal{V}_l$

$\mathcal{V}_l = \mathcal{V}_0$;
**for** $t = 1, 2, \ldots, b$ **do**
  Select the most valuable node $v^* = \arg\max_{v \in \mathcal{V}_{train} \setminus \mathcal{V}_l} \quad F(\mathcal{V}_l \cup \{v\})$;
  Set the influence quality of $v^*$ to the labelling accuracy $\alpha$;
  Update the labeled set $\mathcal{V}_l = \mathcal{V}_l \cup \{v^*\}$;
Update the influence quality of nodes in $\mathcal{V}_l \setminus \mathcal{V}_0$ according to E.q. 8;
**return** $\mathcal{V}_l$

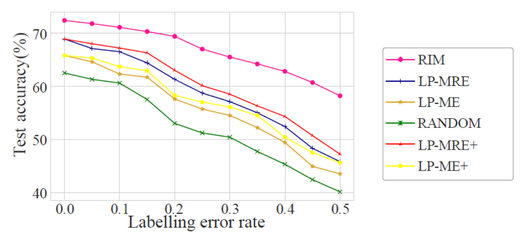
## 4. Experiments

We now verify the effectiveness of RIM on four real-world graphs. We aim to answer four questions.Q1: Compared with other state-of-the-art baselines, can RIM achieve better predictive accuracy? Q2: How does the influence quality and quantity influence RIM? Q3: Is RIM faster than the compared baselines in the end-to-end AL process? Q4: If RIM is more effective than the baselines, what should be the reason?

### 4.1 Performance in GCN and LP.

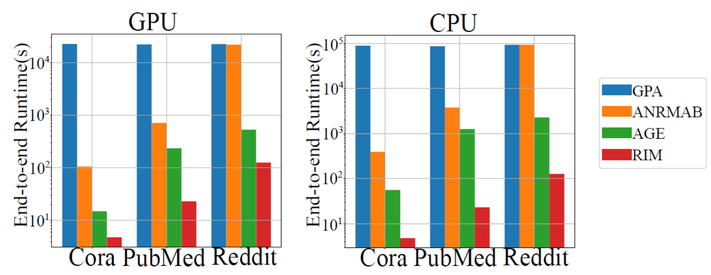



The test accuracy with different labelling error rate of labeled nodes for GCN




The test accuracy with different labelling error rate of labeled nodes for LP

### 4.2 Efficiency comparison




The end-to-end runtime (at logscale) on different datasets.

References:

C. C. Aggarwal, X. Kong, Q. Gu, J. Han, and S. Y. Philip. Active learning: A survey. In Data Classification: Algorithms and Applications, pages 571–605. CRC Press, 2014

M.-R. Bouguelia, S. Nowaczyk, K. Santosh, and A. Verikas. Agreeing to disagree: active learning with noisy labels without crowdsourcing. International Journal of Machine Learning and Cybernetics, 9(8):1307–1319, 2018.

H. Cai, V. W. Zheng, and K. C.-C. Chang. Active learning for graph embedding. arXiv preprint arXiv:1705.05085, 2017

H. Dong, J. Chen, F. Feng, X. He, S. Bi, Z. Ding, and P. Cui. On the equivalence of decoupled graph convolution network and label propagation. arXiv preprint arXiv:2010.12408, 2021.

J. Du and C. X. Ling. Active learning with human-like noisy oracle. In 2010 IEEE International Conference on Data Mining, pages 797–802. IEEE, 2010