Robust Offline Active Learning on Graphs

Yuanchen Wu, Yubai Yuan

Department of Statistics, The Pennsylvania State University

Summary

- We propose an active learning on graphs framework for node-level prediction
- Introduce **informativeness** and **representativeness** criteria for node querying
- **robust** to both node feature and labeling noise
- theoretical guarantee on prediction performance under mild assumption

Background

Active Learning: prioritize informative nodes for labeling

Graph signal space: node feature + network information

$$\mathbf{H}_\omega(\mathbf{X},\mathbf{A}) = \mathrm{Proj}_{\mathbf{L}_\omega}\mathrm{Span}(\mathbf{X}) := \mathrm{Span}\{\mathrm{Proj}_{\mathbf{L}_\omega}X_1,\cdots,\mathrm{Proj}_{\mathbf{L}_\omega}X_p\},$$
$$\text{where } \mathrm{Proj}_{\mathbf{L}_\omega}X_i = \sum_{j:\lambda_j\le\omega}\langle X_i,U_j\rangle U_j.$$

Informativeness

- **Information gain of labeling |S|:** the maximum recoverable dimension of the subspace by labeling |S|
- Select nodes to maximize information gain using **graph signal recovery** theory

Representativeness

- **Representative sampling** to control generalization error due to labeling noise
- Tool: graph sparsification

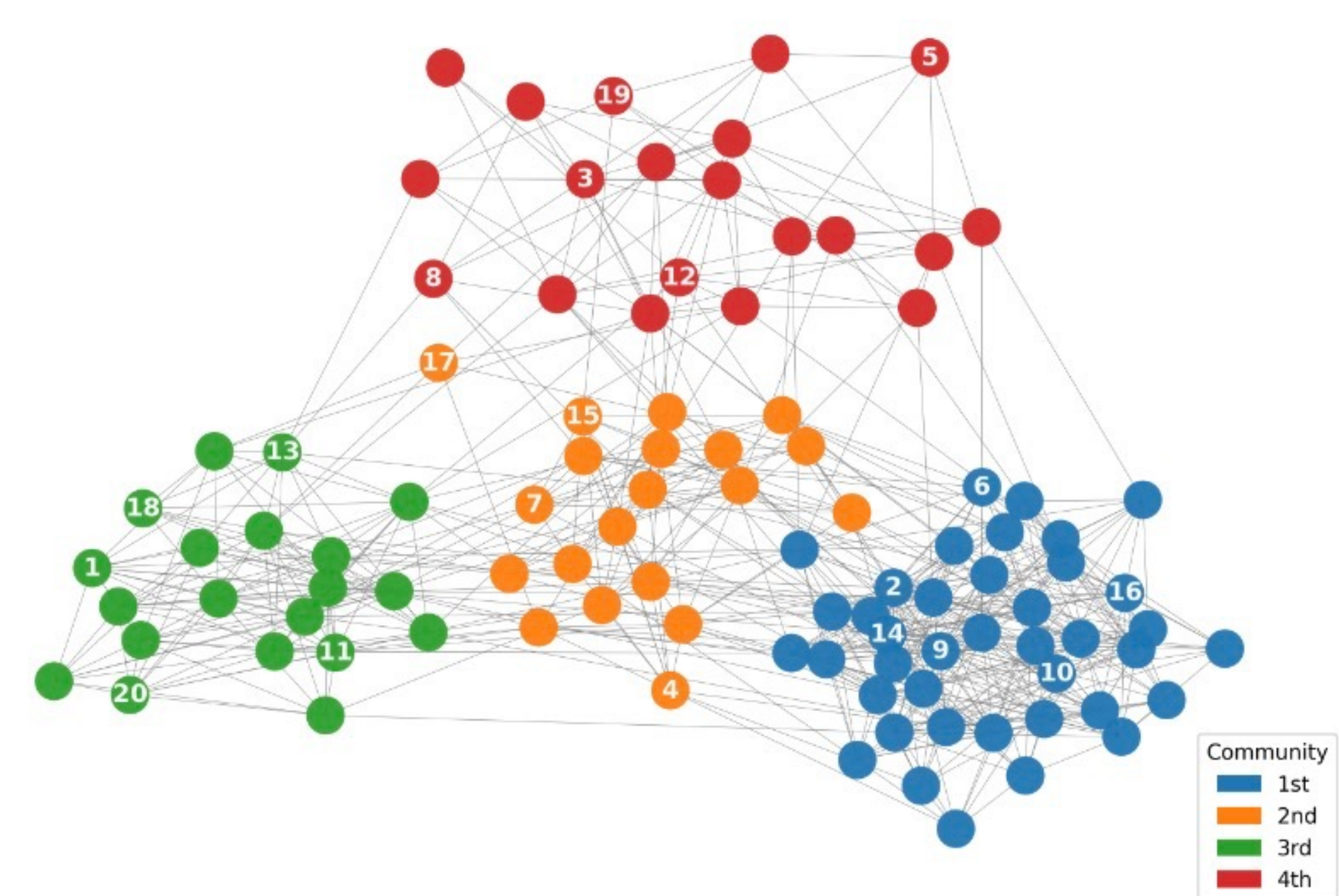
Ultimate bias-variance tradeoff:
Informativeness vs robustness

$$\mathbf{E}_Y\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2 \le \mathcal{O}\Big(\frac{r_d t}{\mathcal{B}} + 2(\frac{r_d t}{\mathcal{B}})^{3/2} + (\frac{r_d t}{\mathcal{B}})^2\Big) \times (n\sigma^2 + \sum_{i>d,i\in supp(\mathbf{f})}\alpha_i^2) + \sum_{i>d,i\in supp(\mathbf{f})}\alpha_i^2$$

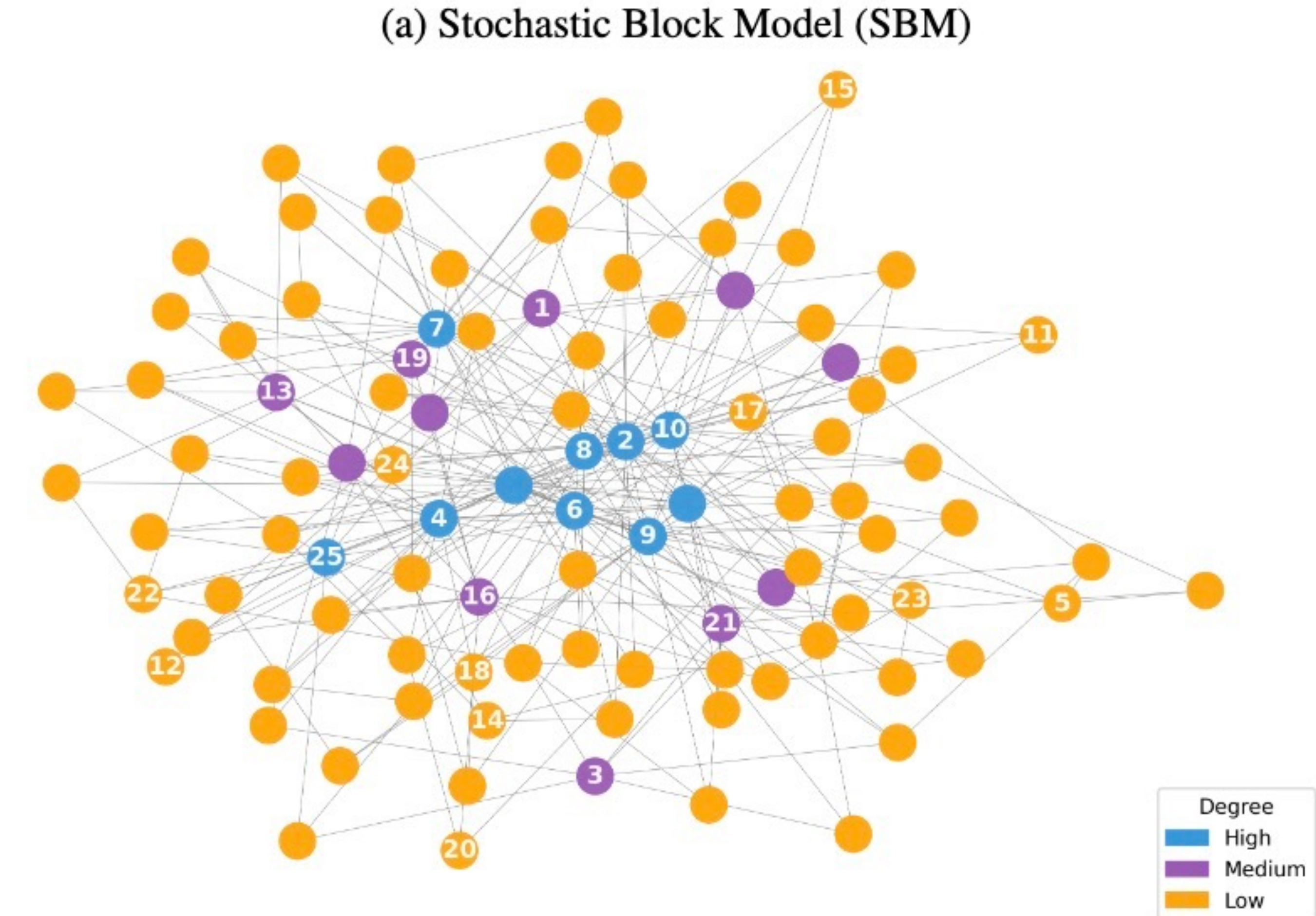
Algorithm

- Initialize $S=\emptyset$, $\mathcal{B}$=query budget
- While $|S|\le\mathcal{B}$:
  1. Sample m nodes as candidate sets based on **representative** sampling
  2. From 1, choose the node i that maximizes **information gain**
  3. Update $S = S\cup i$
- Label all nodes in S

Visualization



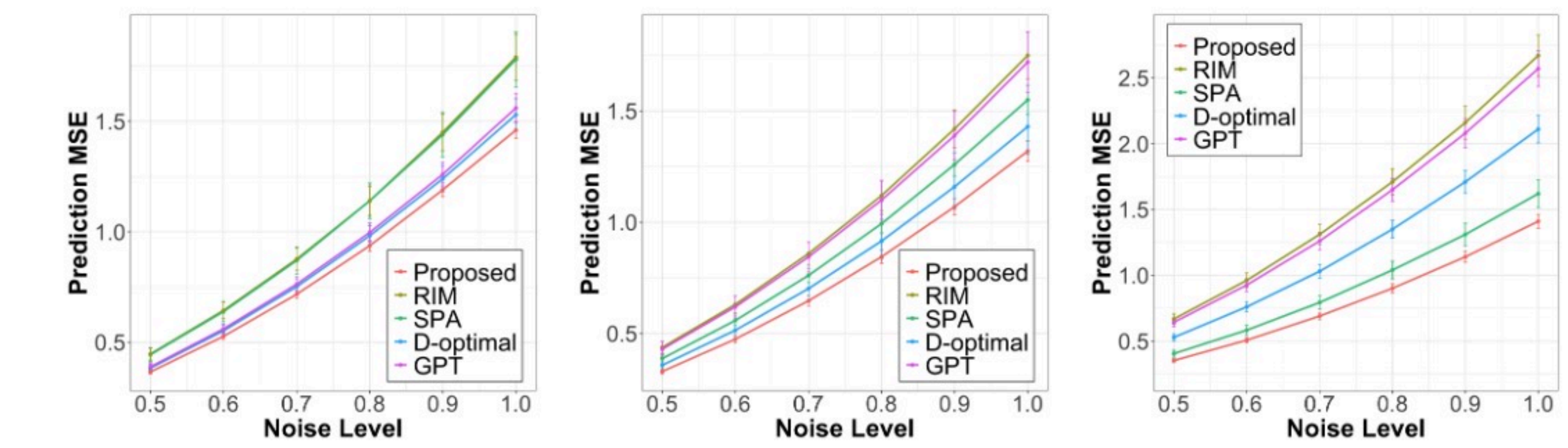
(a) Stochastic Block Model (SBM)



(b) Barabási–Albert model (BA)

Experiments

Synthetic network

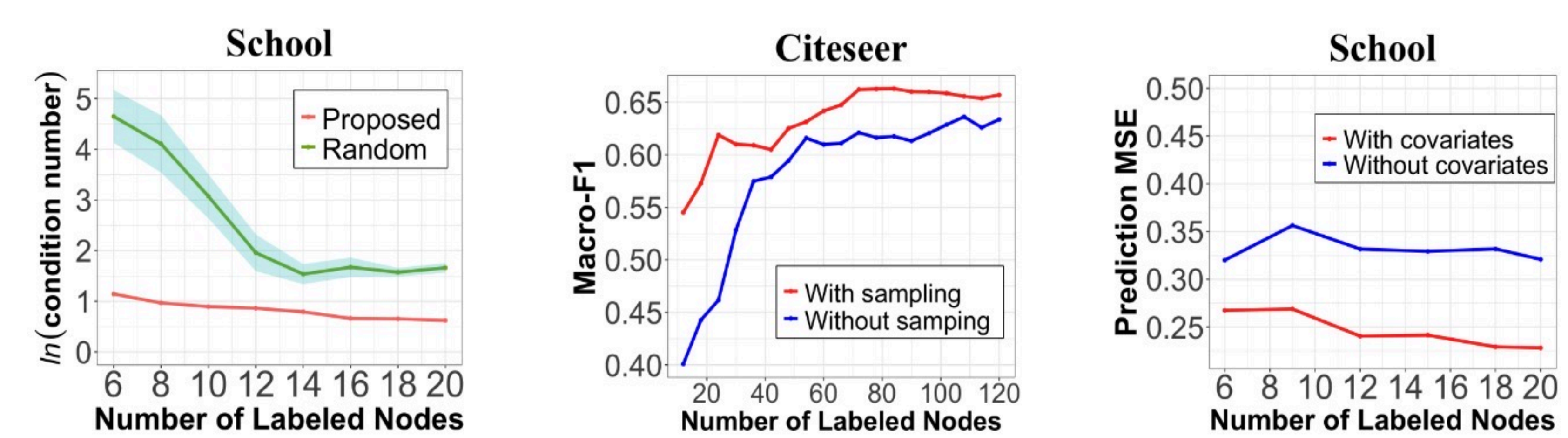


(a) Small World (b) Community (c) Scale-free

Real network

# Labeled Nodes	Cora ($h=0.81$) 35	Cora 70	Cora 140	Chameleon ($h=0.23$) 50	Chameleon 75	Chameleon 100	Texas ($h=0.11$) 15	Texas 30	Texas 45
Random	68.2 ± 1.3	74.5 ± 1.0	78.9 ± 0.9	22.4 ± 2.6	22.1 ± 2.5	21.8 ± 2.1	67.0 ± 3.3	69.9 ± 3.3	73.8 ± 3.2
AGE	72.1 ± 1.1	78.0 ± 0.9	82.5 ± 0.5	30.0 ± 4.5	28.2 ± 4.9	28.6 ± 5.0	67.9 ± 2.6	68.8 ± 3.3	72.1 ± 3.6
GPT	77.4 ± 1.6	81.6 ± 1.2	**86.5** ± 1.2	14.1 ± 2.5	15.8 ± 2.2	16.4 ± 2.4	72.6 ± 2.0	72.5 ± 3.6	74.6 ± 1.8
RIM	77.5 ± 0.8	81.6 ± 1.1	84.1 ± 0.8	**35.5** ± 3.7	**42.8** ± 3.0	34.4 ± 3.5	68.5 ± 3.7	78.4 ± 3.0	74.6 ± 3.7
IGP	77.4 ± 1.7	81.7 ± 1.6	86.3 ± 0.7	32.5 ± 3.6	33.7 ± 3.1	33.4 ± 3.5	70.8 ± 3.7	69.9 ± 3.3	76.1 ± 3.6
SPA	76.5 ± 1.9	80.3 ± 1.6	85.2 ± 0.6	30.2 ± 3.2	28.5 ± 2.9	31.0 ± 4.4	72.0 ± 3.2	72.5 ± 3.1	74.6 ± 2.1
Proposed	**78.4** ± 1.7	**81.8** ± 1.8	**86.5** ± 1.1	35.1 ± 2.8	35.7 ± 3.0	**37.2** ± 3.0	**75.0** ± 1.9	**79.5** ± 0.8	**80.4** ± 2.7

Ablation study



Limitation & Future work

- **Limitation:** over-reliance on a priori knowledge of label construction
- **Future direction**: adaptive methods for designing the signal subspace to effectively handle both homophily and heterophily
- Extension to an **online** active learning setting that iteratively incorporates node response information to further enhance query efficiency