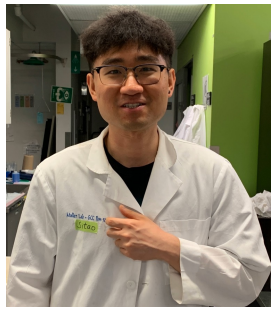
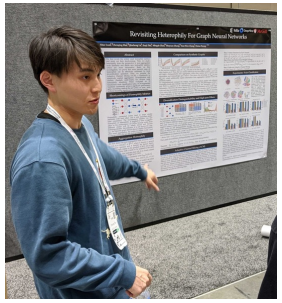
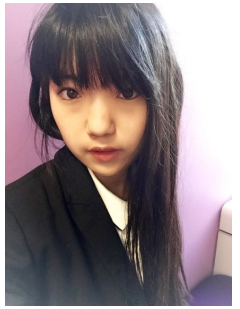
When Do Graph Neural Networks Help with Node Classification?

- Investigating the Impact of Homophily Principle on Node Distinguishability

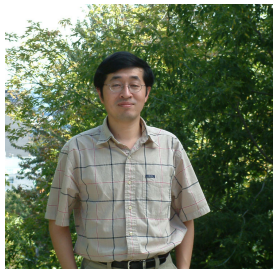

Sitao Luan


Chenqing Hua


Minkai Xu


Qincheng Lu


Jiaqi Zhu


Xiao-Wen Chang


Jie Fu


Jure Leskovec

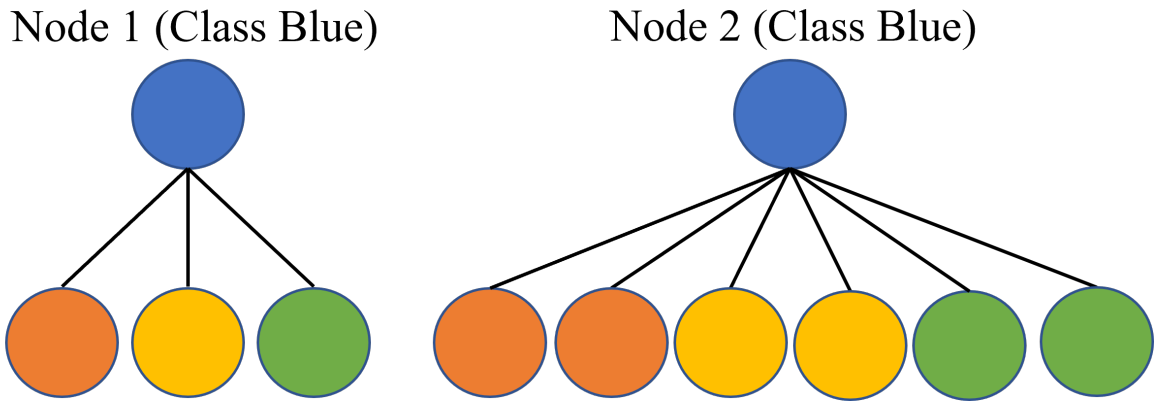

Doina Precup

Introduction

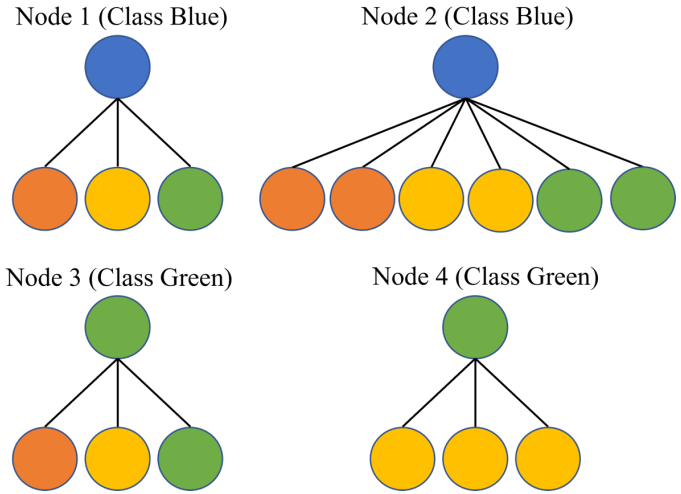
- Homophily principle
  - Nodes with the same labels are more likely to be connected.
  - Believed to be the main reason for the performance superiority of GNNs over NNs on node classification tasks.
- Heterophily
  - Lack of homophily
  - Considered as the main cause of the inferiority of GNNs on heterophilic graphs
  - Nodes from different classes are connected and mixed → Indistinguishable node embeddings → Pose a challenge to the classifier

Introduction

- Deficiency in current literature: Homophily wins, Heterophily loses?
  - As long as nodes within the same class share similar neighborhood patterns, their embeddings will be similar after aggregation. E.g., Node 1,2.
  - Existing homophily metrics cannot accurately indicate the superiority of GNNs.
  - Cannot show when full-pass (FP), low-pass (LP) and high-pass (HP) filters are able to address heterophily.



Motivation



- Deficient understanding on node distinguishability (ND):
  - Only consider intra-class ND but ignore inter-class ND.
  - E.g., node 3 share the same heterophilic neighborhood pattern as nodes 1,2, but it is from class green.
- Our Claim
  - An ideal case is to have smaller intra-class ND than inter-class ND. E.g., nodes 1,2,4.
  - We need to quantify ND and study its relationship with homophily

Quantify ND on A Toy Model

- CSBM-H
  - The features $\boldsymbol{x}$ of two disjoint sets of nodes are generated from two normal distributions $N(\mu_0, \sigma_0^2)$ and $N(\mu_1, \sigma_1^2)$.
  - The degree of nodes in $\mathcal{C}_0$ and $\mathcal{C}_1$ are d_0, d_1 respectively.
  - For $i \in \mathcal{C}_0$, its neighbors are generated by independently sampling from hd_0 intra-class nodes and $(1-h)d_0$ inter-class nodes. The neighbors of $j \in \mathcal{C}_1$ are generated in the same way.
  - The LP and HP filter features are $\boldsymbol{h}$ and $\boldsymbol{h}^{\mathrm{HP}}$.

Optimal Bayes Classifier

[Theorem 1] Suppose $\sigma_0^2 \neq \sigma_1^2$ and $\sigma_0^2, \sigma_1^2 > 0$, the prior distribution for $\boldsymbol{x}$ is $\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_0) = \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_1) = 1/2$, then the optimal Bayes Classifier ($\mathrm{CL}_{\mathrm{Bayes}}$) for CSBM-H($\mu_0, \mu_1, \sigma_0^2 I, \sigma_1^2 I, d_0, d_1, h$) is

$$\mathrm{CL}_{\mathrm{Bayes}}(\boldsymbol{x}) = \begin{cases} 1, \eta(\boldsymbol{x}) \geq 0.5 \\ 0, \eta(\boldsymbol{x}) < 0.5 \end{cases} \text{ and } \eta(\boldsymbol{x}) = \mathbb{P}(z = 1|\boldsymbol{x}) = \frac{1}{1+exp(Q(\boldsymbol{x}))}$$

where $Q(\boldsymbol{x}) = a\boldsymbol{x}^\top \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c, a = \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right), \boldsymbol{b} = \frac{\boldsymbol{\mu_0}}{\sigma_0^2} - \frac{\boldsymbol{\mu_1}}{\sigma_1^2}, c = \frac{\boldsymbol{\mu_1}^\top \boldsymbol{\mu_1}}{2\sigma_1^2} - \frac{\boldsymbol{\mu_0}^\top \boldsymbol{\mu_0}}{2\sigma_0^2} + ln(\frac{\sigma_1^{F_h}}{\sigma_0^{F_h}})$

Probabilistic Bayes Error (PBE)

- Bayes error (BE) for CSBM-H is

$$\mathrm{BE} = \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_0)\,(1 - \mathbb{P}\left(\mathrm{CL}_{\mathrm{Bayes}}(\boldsymbol{x}) = 0 \middle| \boldsymbol{x} \in \mathcal{C}_0\right)) + \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_1)\,(1 - \mathbb{P}\left(\mathrm{CL}_{\mathrm{Bayes}}(\boldsymbol{x}) = 1 \middle| \boldsymbol{x} \in \mathcal{C}_1\right))$$

- BE can be estimated by Probabilistic Bayes Error (PBE)

$$\frac{CDF_{\widetilde{\chi}^2(w_0,F_h,\lambda_0)}(-\xi) + (1-CDF_{\widetilde{\chi}^2(w_1,F_h,\lambda_1)}(-\xi))}{2}$$

where CDF is the Cumulative Distribution Function of generalized chi-square distribution.

- PBE can be used to measure ND. It can be numerically calculated and visualized to show the relationship between h and ND precisely. However, we do not have an analytic expression for PBE, which makes it less explainable and intuitive.

Negative Generalized Jeffery's Divergence

[Generalized Jeffreys Divergence] For a random variable $\boldsymbol{x}$ which has either the distribution $\mathrm{P}(\boldsymbol{x})$ or the distribution $\mathrm{Q}(\boldsymbol{x})$, the generalized Jeffreys divergence is defined as

$$D_{GJ}(P,Q) = \mathbb{P}(\boldsymbol{x} \sim P) E_{\boldsymbol{x}\sim P}\left[\ln \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}\right] + \mathbb{P}(\boldsymbol{x} \sim Q) E_{\boldsymbol{x}\sim Q}[\ln \frac{Q(\boldsymbol{x})}{P(\boldsymbol{x})}]$$

With $\mathbb{P}(\boldsymbol{x} \sim P) = \mathbb{P}(\boldsymbol{x} \sim Q) = 1/2$, the negative generalized Jeffery's divergence for CSBM-H is

$$D_{NGJ}(\mathrm{CSBM{-}H}) = -d_X^2\left(\frac{1}{4\sigma_1^2} + \frac{1}{4\sigma_0^2}\right) - \frac{F_h}{4}\left(\rho^2 + \frac{1}{\rho^2} - 2\right)$$

where $d_X^2 = (\mu_0 - \mu_1)^\top(\mu_0 - \mu_1), \rho = \frac{\sigma_0}{\sigma_1}$

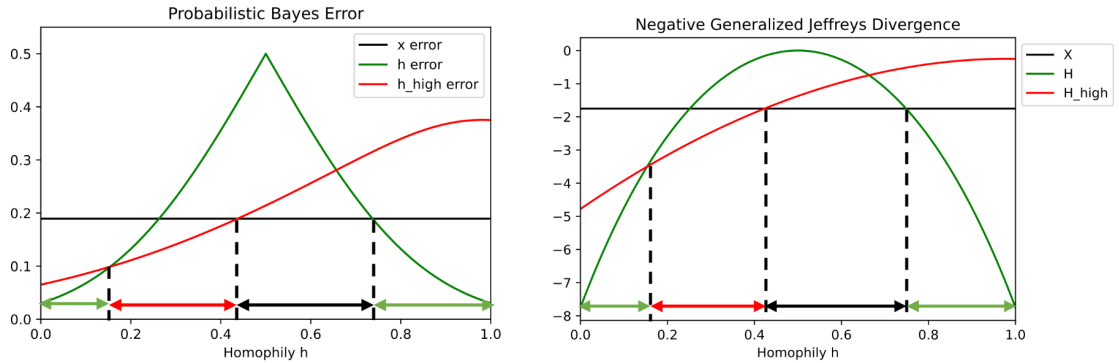
D_{NGJ}

$$D_{NGJ}(\text{CSBM}-\text{H}) = \underbrace{-d_X^2\left(\frac{1}{4\sigma_1^2}+\frac{1}{4\sigma_0^2}\right)}_{\textbf{Negative Normalized Distance}} \quad \underbrace{-\frac{F_h}{4}\left(\rho^2+\frac{1}{\rho^2}-2\right)}_{\textbf{Negative Variance Ratio}}$$

- D_{NGJ} indicates that ND relies on two terms: Expected Negative Normalized Distance (ENND) and the Negative Variance Ratio (NVR)
  - ENND depends on how large is the inter-class ND d_X^2 compared with the normalization term $\frac{1}{4\sigma_1^2}+\frac{1}{4\sigma_0^2}$, which is determined by intra-class ND;
  - NVR depends on how different the two intra-class NDs are, i.e., when they are significantly different (ρ is close to 0), NVR is small which means the nodes are more distinguishable and vice versa.

Analysis

- **Mid-homophily pitfall**: a medium level of homophily has a more detrimental effect on ND than extremely low levels of homophily.
- $\boldsymbol{x}, \boldsymbol{h}$ and $\boldsymbol{h}^{\mathrm{HP}}$ will get the lowest PBE and D_{NGJ} in different homophily intervals, which we refer to as the "FP regime (black)", "LP regime (green)", and "HP regime (red)". LP regime: low and very high homophily intervals (two ends); HP regime: low to medium homophily interval; FP regime: medium to high homophily area.

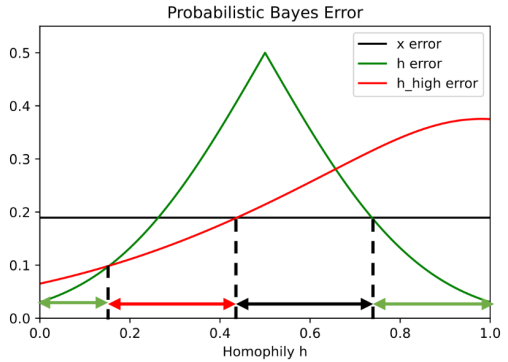


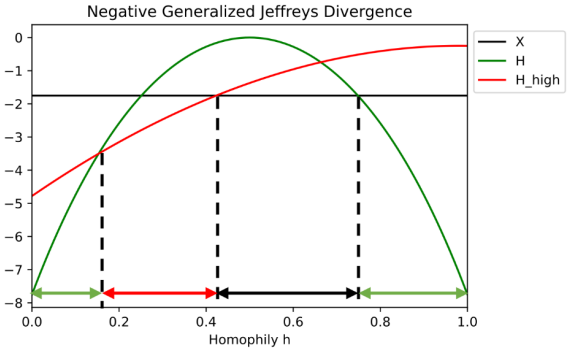
(a) PBE (b) D_{NGJ}

Fig 1: $\mu_0 = [-1,0], \mu_1 = [1,0], \sigma_0^2 = 1, \sigma_1^2 = 2, d_0 = 5, d_1 = 5$

Analysis (Ablation)

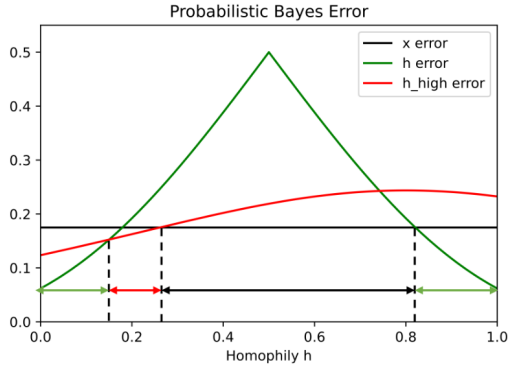
- In Figure 2, as σ_1^2 increases, the PBE and D_{NGJ} of the three curves all go up, which means the node embeddings become less distinguishable under HP, LP and FP filters.
- The significant shrinkage of the HP regimes and the expansion of the FP regime indicates that the original features are more robust to imbalanced variances, especially in the low homophily area.
- See more ablations in our paper.

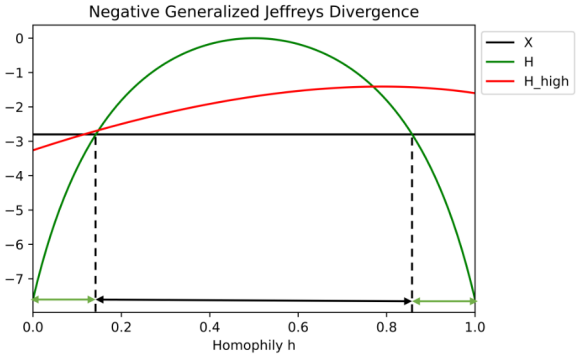




(a) PBE (b) D_{NGJ}

Fig 1: $\mu_0 = [-1,0], \mu_1 = [1,0], \sigma_0^2 = 1, \sigma_1^2 = 2, d_0 = 5, d_1 = 5$





(a) PBE (b) D_{NGJ}

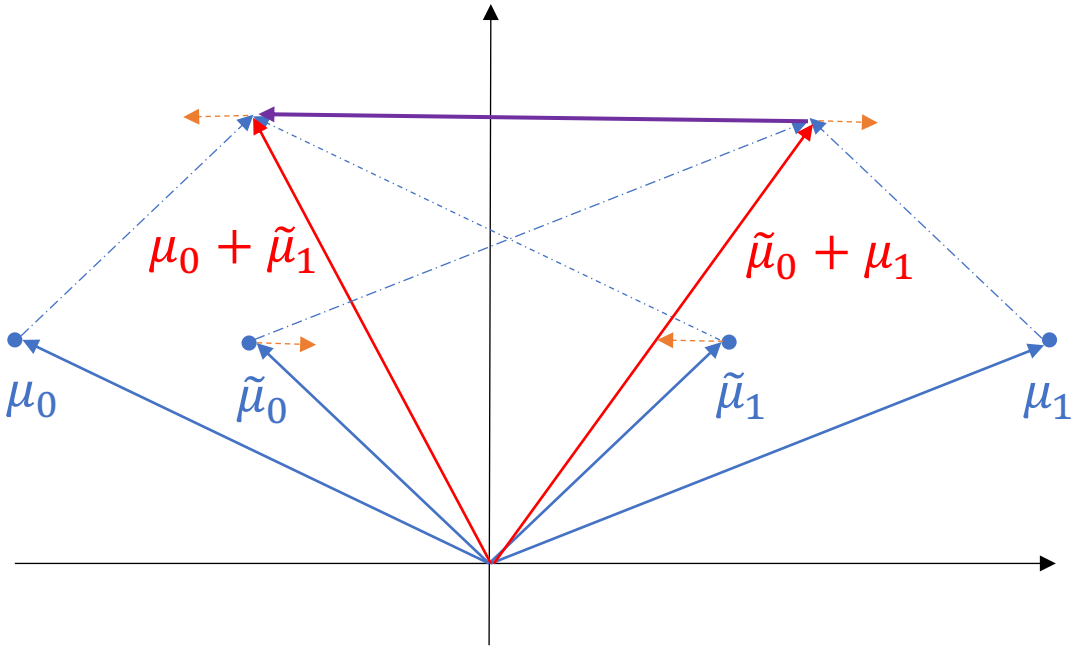
Fig 2: $\mu_0 = [-1,0], \mu_1 = [1,0], \sigma_0^2 = 1, \sigma_1^2 = 5, d_0 = 5, d_1 = 5$

More General Settings

- We investigate ND by studying how significant the intra-class embedding distance is smaller than the inter-class embedding distance.
  - The upper bound mainly depends on a distance term (inter-class ND) and normalized variance term (intra-class ND);
  - The normalized variance term of HP filter is less sensitive to the changes of node degree than that of LP filter, because there is an additional 1 in the constant term. (See details in our paper)

How Does HP Filters Work?

- We show that the distance term of HP filter depends on the **relative center distance**, which is a novel discovery.
- When homophily decreases, the aggregated centers will move away from the original centers, and the relative center distance (purple) will get larger which means the embedding distance of nodes from different classes will have larger probability to be big.



Empirical Study

- To test whether "intra-class embedding distance is smaller than the inter-class embedding distance" strongly relates to the superiority of G-aware models to their coupled G-agnostic models, we conduct the following hypothesis testing
  - Train two G-aware models GCN, SGC-1 and their coupled G-agnostic models MLP-2 and MLP-1 with fine-tuned hyperparameters;
  - For each trained model, we calculate the pairwise Euclidean distance of the node embeddings in output layers.
  - Next, we compute the proportion of nodes whose intra-class node distance is significantly smaller than inter-class node distance and obtain ***Prop*** values. Test

$$H_0: \text{Prop(G-aware model)} \geq \text{Prop(G-agnostic model)};$$
$$H_1: \text{Prop(G-aware model)} < \text{Prop(G-agnostic model)}$$

Results and Comparisons

In most cases (except for GCN vs. MLP-2 on PubMed), when H_1 significantly holds, G-aware models will underperform the coupled G-agnostic models and vice versa. This supports our claim that the performance of G-aware models is closely related to "intra-class vs. inter-class node embedding distances", no matter the homophily levels.

		Cornell	Wisconsin	Texas	Film	Chameleon	Squirrel	Cora	CiteSeer	PubMed
Baseline Homophily Metrics	H_{edge}	0.5669	0.4480	0.4106	0.3750	0.2795	0.2416	0.8100	0.7362	0.8024
	H_{node}	0.3855	0.1498	0.0968	0.2210	0.2470	0.2156	0.8252	0.7175	0.7924
	H_{class}	0.0468	0.0941	0.0013	0.0110	0.0620	0.0254	0.7657	0.6270	0.6641
	H_{agg}	0.8032	0.7768	0.694	0.6822	0.61	0.3566	0.9904	0.9826	0.9432
	H_{GE}	0.31	0.34	0.35	0.16	0.0152	0.0157	0.17	0.19	0.27
	H_{adj}	0.1889	0.0826	0.0258	0.1272	0.0663	0.0196	0.8178	0.7588	0.7431
	LI	0.0169	0.1311	0.1923	0.0002	0.048	0.0015	0.5904	0.4508	0.4093
Classifier-based Performance Metrics	KR_{NNGP}	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
	GNB	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
SGC v.s. MLP-1	p-value	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00
	ACC SGC	70.98 ± 8.39	70.38 ± 2.85	83.28 ± 5.43	25.26 ± 1.18	64.86 ± 1.81	47.62 ± 1.27	85.12 ± 1.64	79.66 ± 0.75	85.5 ± 0.76
	ACC MLP-1	93.77 ± 3.34	93.87 ± 3.33	93.77 ± 3.34	34.53 ± 1.48	45.01 ± 1.58	29.17 ± 1.46	74.3 ± 1.27	75.51 ± 1.35	86.23 ± 0.54
	Diff Acc	-22.79	-23.49	-10.49	-9.27	19.85	18.45	10.82	4.15	-0.73
GCN v.s. MLP-2	p-value	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00
	ACC GCN	82.46 ± 3.11	75.5 ± 2.92	83.11 ± 3.2	35.51 ± 0.99	64.18 ± 2.62	44.76 ± 1.39	87.78 ± 0.96	81.39 ± 1.23	88.9 ± 0.32
	ACC MLP-2	91.30 ± 0.70	93.87 ± 3.33	92.26 ± 0.71	38.58 ± 0.25	46.72 ± 0.46	31.28 ± 0.27	76.44 ± 0.30	76.25 ± 0.28	86.43 ± 0.13
	Diff Acc	-8.84	-18.37	-9.15	-3.07	17.46	13.48	11.34	5.14	2.47

Performance Metrics Beyond Homophily

- The p-value can be a better performance metric for GNNs beyond homophily. Moreover, the p-value can provide a statistical threshold, and this property is not present in existing homophily metrics.
- However, it is required to train and fine-tune the models to obtain the p-values, which make it less practical because of computational costs.
- We choose Gaussian Naïve Bayes (GNB) and Kernel Regression (KR) with Neural Network Gaussian Process (NNGP), which do not require iterative training to capture the **feature-based linear or non-linear** information.

Method

$$H_0: ACC(Classifier(H)) \geq ACC(Classifier(X));$$
$$H_1: ACC(Classifier(H)) < ACC(Classifier(X))$$

- We first randomly sample 500 labeled nodes and splits them into 60%/40% as "training" and "test" data. The original features X and aggregated features H of the sampled training and test nodes can be calculated and are then fed into a given classifier.
- The prediction accuracy of the test nodes will be computed directly with feedforward method. We repeat this process for 100 times to get 100 samples of prediction accuracy for X and H.
- The p-value can provide a statistical threshold value, such as 0.05, to indicate whether H is significantly better than X for node classification.

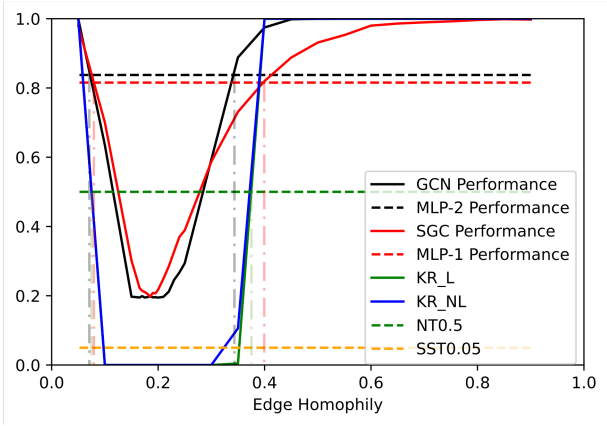
Results

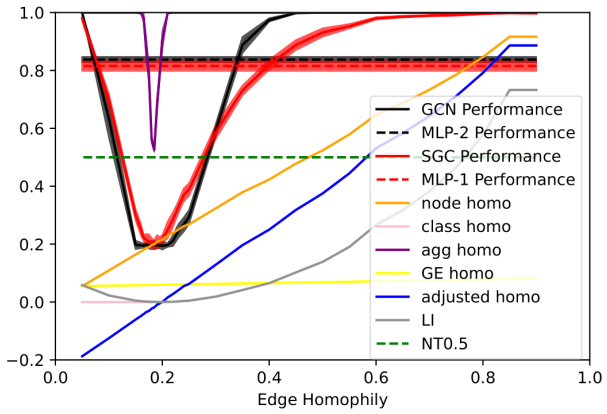
- KR and GNB based metrics significantly outperform the existing homophily metrics, reducing the errors from at least 5 down to just 1 out of 18 cases. Besides, we only need a small set of the labels to calculate the p-value, which makes it better for sparse label scenario.

		Cornell	Wisconsin	Texas	Film	Chameleon	Squirrel	Cora	CiteSeer	PubMed
Baseline Homophily Metrics	H_{edge}	0.5669	0.4480	0.4106	0.3750	0.2795	0.2416	0.8100	0.7362	0.8024
	H_{node}	0.3855	0.1498	0.0968	0.2210	0.2470	0.2156	0.8252	0.7175	0.7924
	H_{class}	0.0468	0.0941	0.0013	0.0110	0.0620	0.0254	0.7657	0.6270	0.6641
	H_{agg}	0.8032	0.7768	0.694	0.6822	0.61	0.3566	0.9904	0.9826	0.9432
	H_{GE}	0.31	0.34	0.35	0.16	0.0152	0.0157	0.17	0.19	0.27
	H_{adj}	0.1889	0.0826	0.0258	0.1272	0.0663	0.0196	0.8178	0.7588	0.7431
	LI	0.0169	0.1311	0.1923	0.0002	0.048	0.0015	0.5904	0.4508	0.4093
Classifier-based Performance Metrics	KR_{NNGP}	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
	GNB	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
SGC v.s. MLP-1	p-value	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00
	ACC SGC	70.98 ± 8.39	70.38 ± 2.85	83.28 ± 5.43	25.26 ± 1.18	64.86 ± 1.81	47.62 ± 1.27	85.12 ± 1.64	79.66 ± 0.75	85.5 ± 0.76
	ACC MLP-1	93.77 ± 3.34	93.87 ± 3.33	93.77 ± 3.34	34.53 ± 1.48	45.01 ± 1.58	29.17 ± 1.46	74.3 ± 1.27	75.51 ± 1.35	86.23 ± 0.54
	Diff Acc	-22.79	-23.49	-10.49	-9.27	19.85	18.45	10.82	4.15	-0.73
GCN v.s. MLP-2	p-value	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00
	ACC GCN	82.46 ± 3.11	75.5 ± 2.92	83.11 ± 3.2	35.51 ± 0.99	64.18 ± 2.62	44.76 ± 1.39	87.78 ± 0.96	81.39 ± 1.23	88.9 ± 0.32
	ACC MLP-2	91.30 ± 0.70	93.87 ± 3.33	92.26 ± 0.71	38.58 ± 0.25	46.72 ± 0.46	31.28 ± 0.27	76.44 ± 0.30	76.25 ± 0.28	86.43 ± 0.13
	Diff Acc	-8.84	-18.37	-9.15	-3.07	17.46	13.48	11.34	5.14	2.47

Results on Synthetic Datasets

- The intersections of performance curves of GCN vs. MLP-2 and SGC vs. MLP-1 perfectly match with the intersection of KR curves with the threshold 0.05. However, the existing homophily metrics fails.
- The new classifier-based performance metrics are much more effective than the existing homophily metrics on revealing the advantage and disadvantage of GNNs.





Performance Metrics	Linear or Non-linear	Feature Dependency	Sparse Labels	Statistical Threshold
H_{node}	linear	✗	✗	✗
H_{edge}	linear	✗	✗	✗
H_{class}	linear	✗	✗	✗
H_{agg}	linear	✗	✓	✗
H_{GE}	linear	✓	✓	✗
H_{adj}	linear	✗	✗	✗
LI	linear	✗	✗	✗
Classifier	both	✓	✓	✓

Q&A

Paper