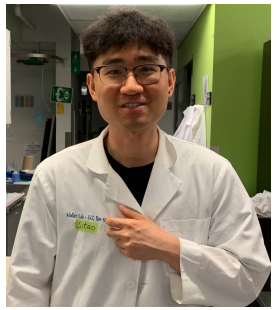
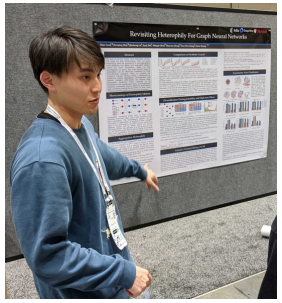
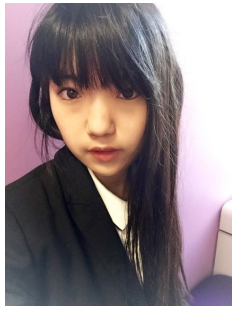
When Do Graph Neural Networks Help with Node Classification?

- Investigating the Impact of Homophily Principle on Node Distinguishability

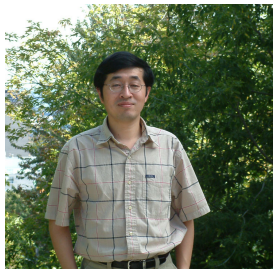

Sitao Luan


Chenqing Hua


Minkai Xu


Qincheng Lu


Jiaqi Zhu


Xiao-Wen Chang


Jie Fu


Jure Leskovec


Doina Precup

Performance Degradation

Datasets\Models	MLP Acc	GCN Acc	GAT Acc	GraphSAGE Acc	Baseline Average	Edge Homophily	Node Homophily
Cornell	85.14	60.81	59.19	82.97	67.66	0.3	0.11
Wisconsin	87.25	63.73	60.78	87.84	70.78	0.21	0.16
Texas	84.59	61.62	59.73	82.43	67.93	0.11	0.06
Film	36.08	30.98	29.71	35.28	31.99	0.22	0.24
Chameleon	46.21	61.34	61.95	47.32	56.87	0.23	0.25
Squirrel	29.39	41.86	43.88	30.16	38.63	0.22	0.22
Cora	74.81	87.32	88.07	85.98	87.12	0.81	0.83
Citeseer	73.45	76.70	76.42	77.07	76.73	0.74	0.71
Pubmed	87.86	88.24	87.81	88.59	88.21	0.8	0.79

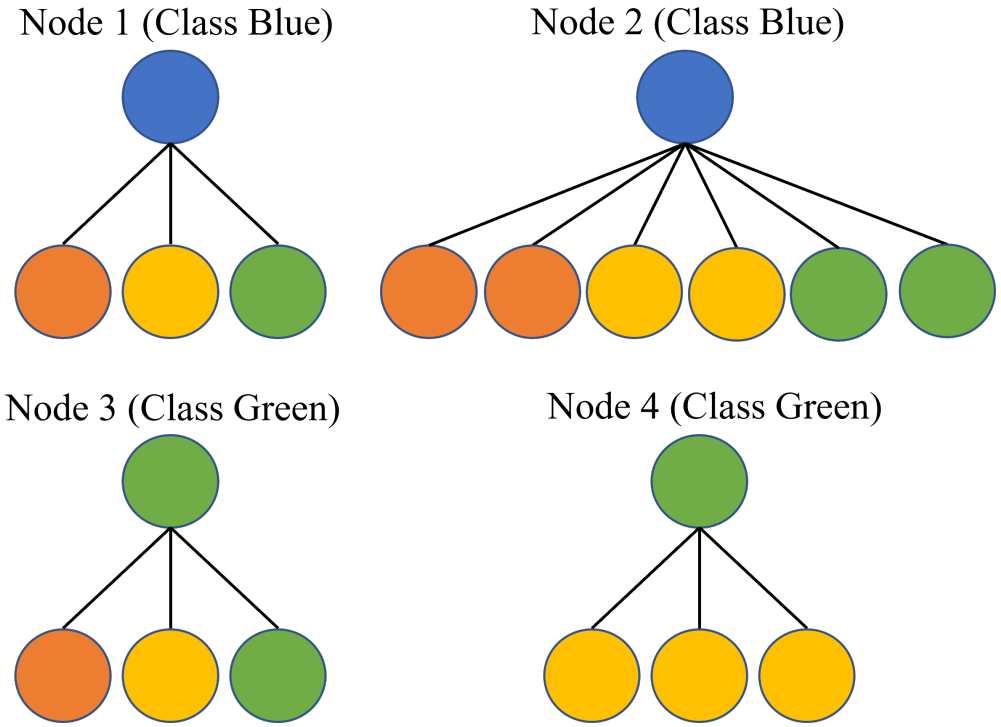
Shallow baseline GNNs underperform MLP-2 on some datasets

Homophily wins, Heterophily loses?

- Homophily: nodes from the same class are more likely to be connected. Foundation of the success of graph neural networks.
- Heterophily: nodes from different classes become less distinguishable. Main cause of performance degradation.
- Question: homophily wins, heterophily loses? No

Node Distinguishability

- Current understanding: As long as nodes within the same class share similar neighborhood patterns, their embeddings will be similar after aggregation (*Ma et al., ICLR 2022*). E.g., nodes {1,2}.
- Deficiency: Only consider intra-class ND but ignore inter-class ND, e.g., node 3 and nodes {1,2}.
- Our Claim: An ideal case is to have smaller intra-class ND than inter-class ND. E.g., nodes {1,2,4}.



Study Homophily on ND

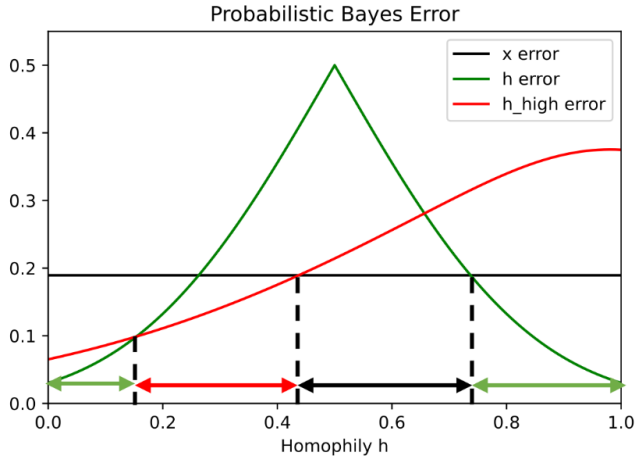
- Quantify ND on CSBM-H
  - Features of two sets of nodes are generated from two normal distributions.
  - The neighbors are generated by independently sampling from intra-class and inter-class nodes based on a given homophily value.
- Compute the Probabilistic Bayes Error (PBE) for the Optimal Bayes Classifier of CSBM-H. **PBE can be used to measure ND**.
- PBE can be numerically calculated, but it does not have an analytic expression, which makes it less explainable and intuitive.

Negative Generalized Jefferys Divergence

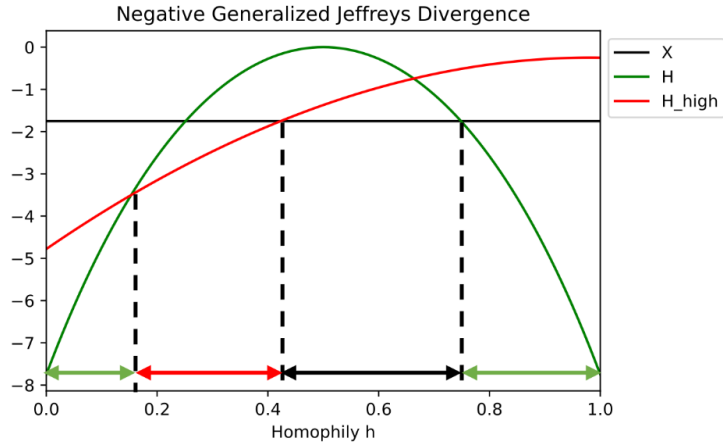
- D_{NGJ}: a KL-divergence based metric.
- D_{NGJ} relies on a normalized distance term and it depends on
  - A distance term which describes inter-class node distinguishability.
  - A normalization term which describes intra-class node distinguishability.

Analysis

- Plot the relation between homophily and ND
- **Mid-homophily pitfall**: a medium level of homophily has a more detrimental effect on ND than extremely low levels of homophily.



(a) PBE



(b) D_{NGJ}

- Ablation study on node degree, class variances and so on.

More General Settings

- We investigate ND in a broader setting by studying "how significant the intra-class embedding distance is smaller than the inter-class embedding distance".
  - The upper bound of node distinguishability depends on both intra- and inter-class node distinguishability.
- New discovery: the node distinguishability for HP filtered features depends on **relative center distance.** This explains how high-pass filter can address some heterophily cases.

Empirical Study

- To test whether "intra-class embedding distance is smaller than the inter-class embedding distance" strongly relates to the superiority of graph-aware models (GCN, SGC-1) to their coupled graph-agnostic models (MLP-2, MLP-1) ,
  - For different graph-aware and graph-agnostic models, we compute the proportion of nodes whose intra-class node distance is significantly smaller than inter-class node distance.
  - Conduct hypothesis testing on the proportion values.
  - In most cases, when the null hypothesis significantly not holds, the graph-aware models will underperform the graph-agnostic models and vice versa. This verifies the strong correlation.

Beyond Homophily Metrics

- P-value can provide an accurate statistical threshold, but it's computationally expensive in practice.
  - Use Gaussian Naïve Bayes (GNB) and Kernel Regression (KR) with Neural Network Gaussian Process, which do not require iterative training
  - Hypothesis testing on prediction accuracies of sampled nodes.

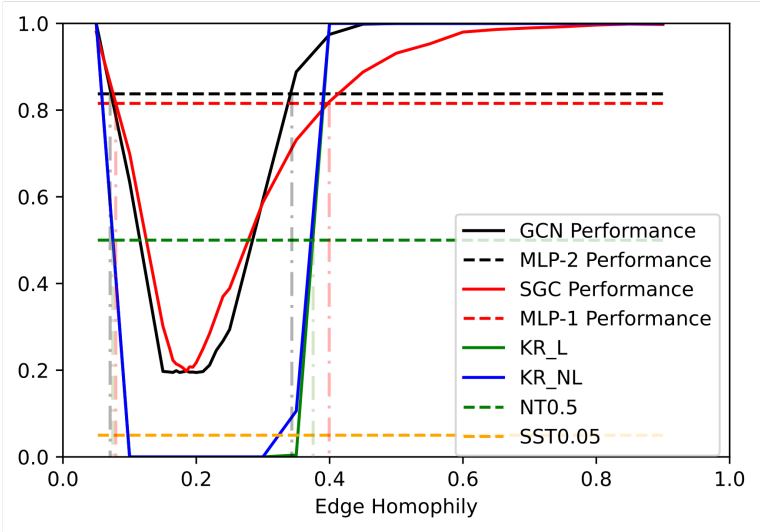
Results

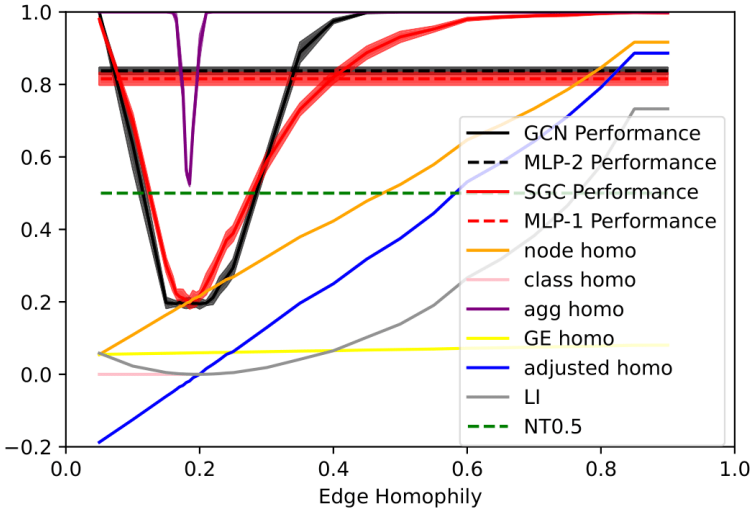
KR and GNB based metrics significantly outperform the existing homophily metrics, reducing the errors from at least 5 down to just 1 out of 18 cases.

		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

- Generate synthetic graphs with different homophily levels. Test if the metrics are consistent with the behavior of GNNs.
  - 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.
  - Other metrics fails. → The proposed metrics are much more effective than existing metrics on revealing the advantage and disadvantage of GNNs.





Q&A

Paper

Code