Revisiting Heterophily For Graph Neural Networks

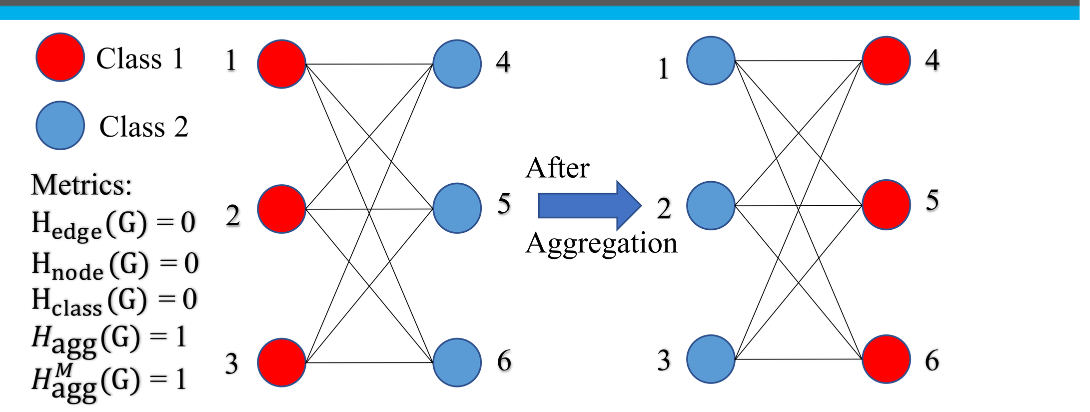
Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, Doina Precup

Mila · DeepMind · McGill

Abstract

GNNs have been commonly believed to outperform NNs in real-world tasks, but recent work has identified a non-trivial set of datasets where graph-aware models underperform graph-agnostic models. People commonly believe that the main cause of this empirical observation is heterophily (low homophily), i.e., there exist much more inter-class edges than inner-class edge that nodes from different classes will be falsely mixed and become indistinguishable. In this paper, we first revisit the widely used homophily metrics and point out their shortcomings. Then, we study heterophily from the perspective of post-aggregation node similarity and propose new homophily metrics, which are more informative. Based on this investigation, we prove that some harmful cases of heterophily can be effectively addressed by local diversification operation. Then, we propose the Adaptive Channel Mixing (ACM), a framework to adaptively exploit aggregation, diversification and identity channels node-wisely for diverse node heterophily situations. When evaluated on 10 benchmark node classification tasks, ACM-augmented baselines consistently achieve significant performance gain, exceeding state-of-the-art GNNs on most tasks without incurring significant computational burden.

Shortcomings of Homophily Metrics



The existing homophily metrics only describe graph-label consistency. The inconsistency relation is implied to have negative effect to the performance of GNNs. Nevertheless, it is not always the case, e.g., the bipartite graph shown above is highly heterophilic, but in fact, the nodes in classes 1 and 2 just exchange colors and are still distinguishable after mean aggregation.

Aggregation Homophily

We first define the post-aggregation node similarity matrix

$$S(\hat{A}, X) \equiv \hat{A}X(\hat{A}X)^T \in \mathbb{R}^{N\times N}$$

It strongly relates to the gradient of SGC and a large $[S(\hat{A}, X)]_{ij}$ means node i tends to be updated to the same class as node j. We then define the aggregation similarity score $S_{agg}\left(S(\hat{A}, X)\right)$ as:

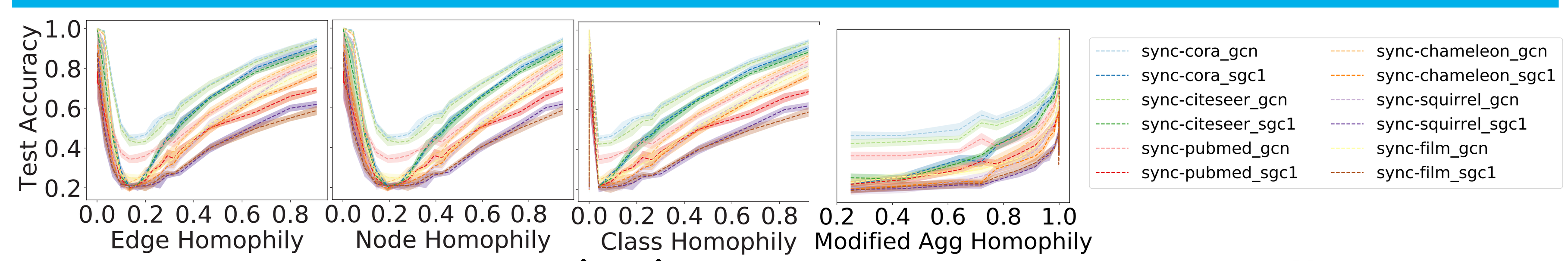
$$\frac{1}{|V|}\left|\left\{v \middle| \mathrm{Mean}_u\left(\{S(\hat{A},X)_{v,u} \middle| Z_{u,:} = Z_{v,:}\}\right) \geq \mathrm{Mean}_u\left(\{S(\hat{A},X)_{v,u} \middle| Z_{u,:} \neq Z_{v,:}\}\right)\right\}\right|$$

It measures the proportion of nodes who put the average weights on the nodes in the same class (including v) more than that in other classes. Aggregation homophily and its modified version are defined as:

$$H_{agg}(G) = S_{agg}\left(S(\hat{A}, Z)\right)$$

$$H^M_{agg}(G) = \left[2\, S_{agg}\left(S(\hat{A}, Z)\right) - 1\right]_+$$

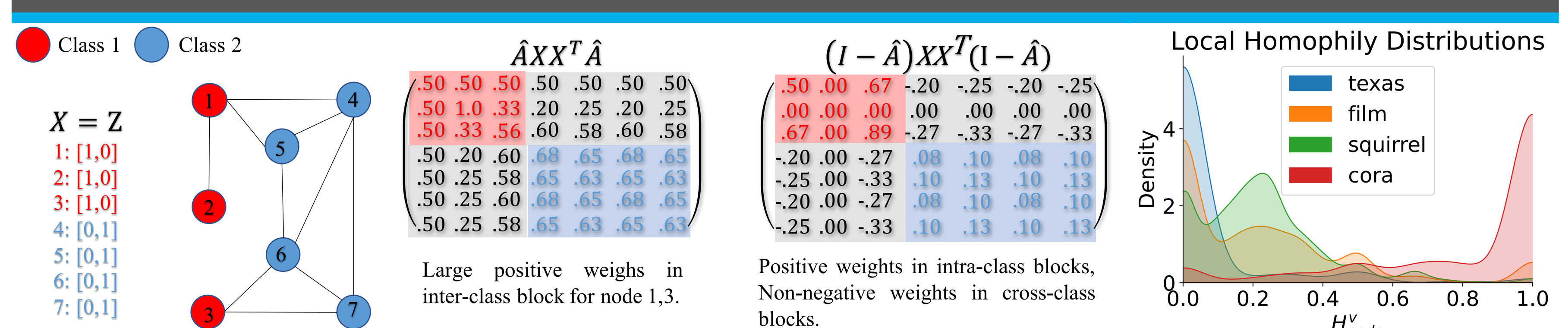
Comparison on Synthetic Graphs



In the previous example, when $\hat{A} = \hat{A}_{rw}$, it is easy to see that $H_{agg}(G) = H^M_{agg}(G) = 1$ and other metrics are 0. Thus, this new metric reflects the fact that nodes in classes 1 and 2 are still highly distinguishable after aggregation, while other metrics fail to capture such information and misleadingly give value 0.

To comprehensively compare the metrics, we generate synthetic graphs with different homophily levels and evaluate 1-hop SGC and GCN on them. The performance of GNNs is expected to be monotonically increasing if the homophily metric is informative. However, the above figures show that the performance curves under edge, node and class homophily are U-shaped, while for $H^M_{agg}(G)$, it reveals a nearly monotonic curve. This indicates that $H^M_{agg}(G)$ provides a better indication of the way in which the graph structure affects the performance of SGC-1 and GCN than existing metrics.

Diversification Distinguishability and High-pass Filters



Diversification operation (high-pass(HP) filter), $I - \hat{A}$ is empirically found to be useful for heterophily problem. In the above example, $S(\hat{A}, X)$ assigns large positive weights to nodes from different classes, while $S(I - \hat{A}, X)$ assigns negative values to nodes from different classes. This indicates that in some heterophily cases, nodes are still distinguishable through their local surrounding dissimilarities. Thus, we define Diversification Distinguishability (DD) based on $S(I - \hat{A}, X)$.

Definition. Given $S(I - \hat{A}, X)$, a node v is diversification distinguishable if the following two conditions are satisfied at the same time.

1. $Mean_u\left(\{S(I-\hat{A},X)_{v,u} \middle| u \in V, Z_{u,:} = Z_{v,:}\}\right) \geq 0$; 2. $Mean_u\left(\{S(I-\hat{A},X)_{v,u} \middle| u \in V, Z_{u,:} \neq Z_{v,:}\}\right) < 0$;

Then, graph diversification distinguishability value is defined as

$$DD_{\hat{A},X}(G) = \frac{1}{|V|}\left|\{v | v \in V, v \text{ is diversification distinguishable}\}\right|$$

Theorem: For $C = 2$, suppose $X = Z$, $\hat{A} = \hat{A}_{rw}$, then all nodes are diversification distinguishable and $DD_{\hat{A},Z}(G) = 1$.

Adaptive Channel Mixing (ACM)

We propose to include low-pass, high-pass and identity channels together in each GNN layers instead of using the uni-channel architecture. Besides, different nodes may have different local heterophily situations and may have different needs for the information from different channels. So we propose a node-wise channel mixing mechanism to adaptively exploit the channel information.

We will use GCN as an example to introduce the ACM framework:

Step 1. Feature Extraction for Each Channel:

Option 1: $H^l_L = \mathrm{ReLU}(H_{LP}H^{l-1}W^{l-1}_L), H^l_H = \mathrm{ReLU}(H_{HP}H^{l-1}W^{l-1}_H), H^l_I = \mathrm{ReLU}(IH^{l-1}W^{l-1}_I)$;

Option 2: $H^l_L = H_{LP}\mathrm{ReLU}(H^{l-1}W^{l-1}_L), H^l_H = H_{HP}\mathrm{ReLU}(H^{l-1}W^{l-1}_H), H^l_I = I\mathrm{ReLU}(H^{l-1}W^{l-1}_I)$;

$H^0 =\in \mathbb{R}^{N\times F_0}, W^{l-1}_L, W^{l-1}_H, W^{l-1}_I \in \mathbb{R}^{F_{l-1}\times F_l}, l = 1, \ldots, L$;

Step 2. Row-wise Feature-based Weight Learning:

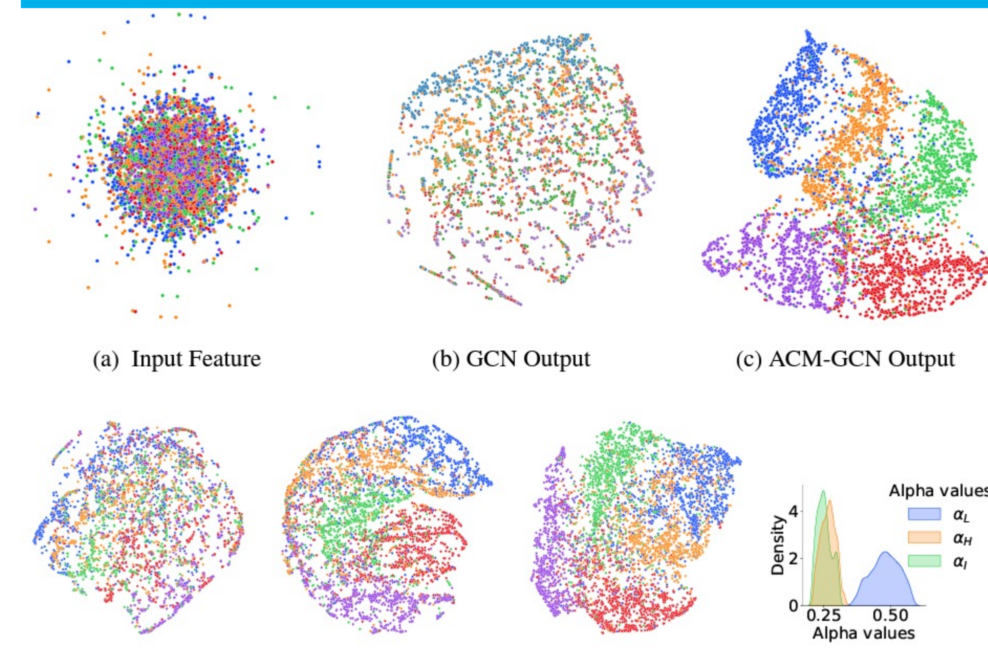
$\tilde{\alpha}^l_L = \mathrm{Sigmoid}(H^l_L\tilde{W}^l_L),\ \tilde{\alpha}^l_H = \mathrm{Sigmoid}(H^l_H\tilde{W}^l_H),\ \tilde{\alpha}^l_I = \mathrm{Sigmoid}(H^l_I\tilde{W}^l_I),\ \tilde{W}^l_L, \tilde{W}^l_H, \tilde{W}^l_I \in \mathbb{R}^{F_l\times 1}$;

$[\alpha^l_L, \alpha^l_H, \alpha^l_I] = \mathrm{Softmax}\left(\left(\frac{[\tilde{\alpha}^l_L, \tilde{\alpha}^l_H, \tilde{\alpha}^l_I]}{T}\right)W^l_{\mathrm{Mix}}\right) \in \mathbb{R}^{N\times 3}, T \in \mathbb{R}$ is temperature, $W^l_{\mathrm{Mix}} \in \mathbb{R}^{3\times 3}$;

Step 3. Node-wise Adaptive Channel Mixing:

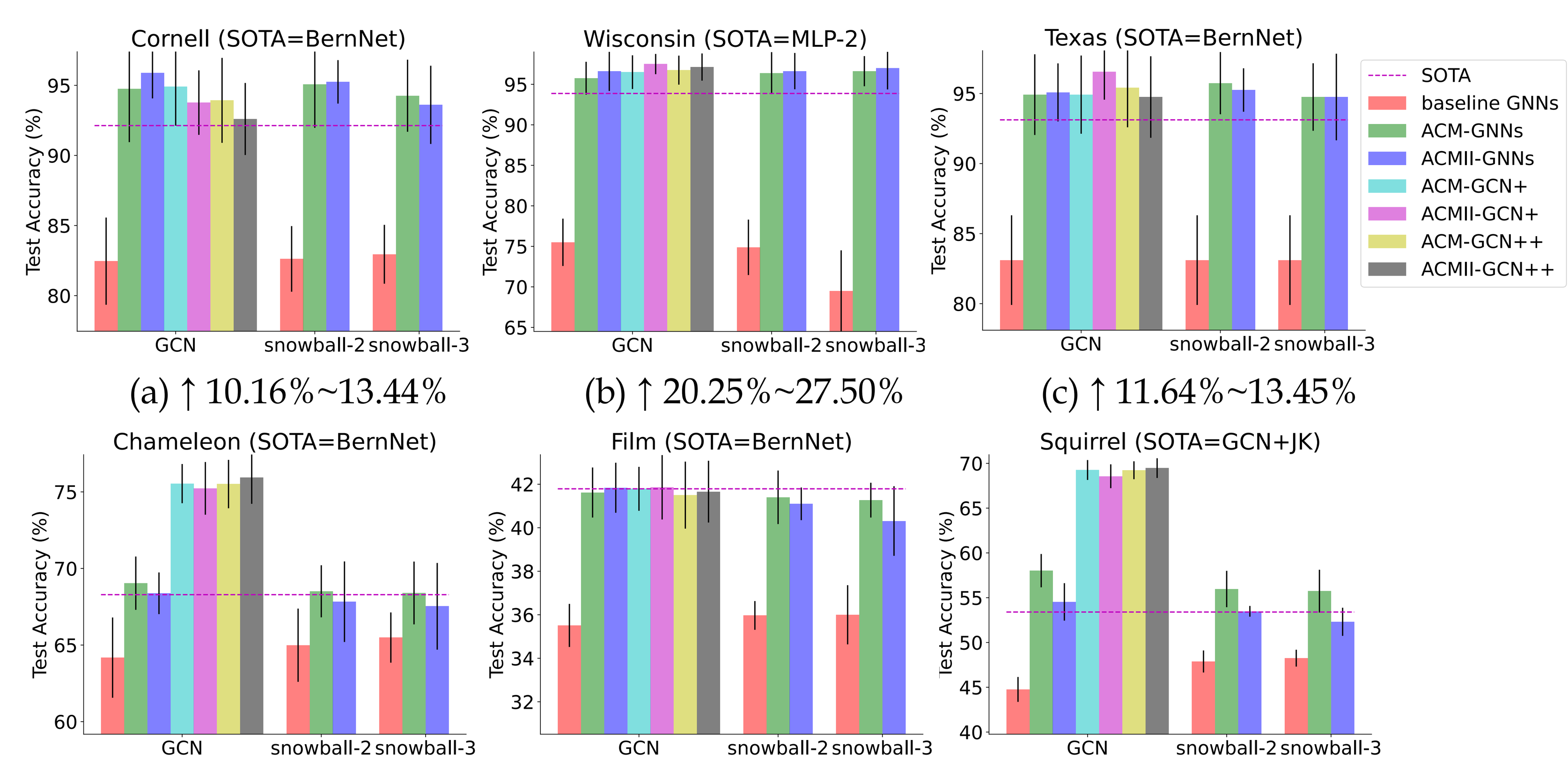
$H^l = \mathrm{ReLU}\left(\mathrm{diag}(\alpha^l_L)H^l_L + \mathrm{diag}(\alpha^l_H)H^l_H + \mathrm{diag}(\alpha^l_I)H^l_I\right)$.

Experiments: Node Classification



(a) Input Feature (b) GCN Output (c) ACM-GCN Output (d) Low-pass Channel (e) High-pass Channel (f) Identity Channel (g) Learned Weights α

The t-SNE visualization on the left demonstrates that the high-pass channel (e) and identity channel (f) can extract meaningful patterns, which the low-pass channel (d) is not able to capture. The output of ACM-GCN (c) shows clearer boundaries among classes than GCN (b) and any channel. See more t-SNE visualizations and comparisons in our paper.



(a) ↑ 10.16%~13.44% (b) ↑ 20.25%~27.50% (c) ↑ 11.64%~13.45%

(d) ↑ 2.04%~11.90% (e) ↑ 4.31%~6.35% (f) ↑ 4.06%~25.22%

To visualize the performance, we plot the bar charts of the test accuracy of SOTA models, three selected baselines (GCN, snowball-2, snowball-3), their ACM(II) augmented models, ACM(II)-GCN+ and ACM(II)-GCN++ on the 6 most commonly used benchmark heterophily datasets. From the above figures we can see that (1) after being combined with the ACM or ACMII framework, the performance of the three baseline models is ***significantly boosted, by 2.04%~27.50%*** on all the 6 tasks. The ACM and ACMII in fact achieve SOTA performance. (2) On *Cornell, Wisconsin, Texas, Chameleon* and *Squirrel*, the augmented baseline models ***significantly outperform the current SOTA models***. Overall, these results suggest that the proposed approach can help GNNs to generalize better on node classification tasks on heterophilic graphs.

See our paper for ablation study, more performance comparisons, details of implementation and discussion of limitations.