When Do Graph Neural Networks Help with Node Classification? Investigating the Impact of Homophily Principle on Node Distinguishability

Abstract

Homophily principle is considered to be the reason behind good performance of GNNs on graph related machine learning tasks. In this project we have tried to evaluate this claim by testing the performance of a CSBM-H model on varying levels of homophily values and evaluating the Probabilistic Bayes Error to check if the claims hold.

1 Introduction

It is observed that GNNs generally perform well on graph related machine learning tasks as compared to graph-agnostic models like MLPs and the reason was believed to be due to the homophily principle. But the relationship between homophily in the graph and performance of GNN is not that simple.

We have demonstrated this by performing various experiments using a modified CSBM model for homophily parameter CSBM-H. This model takes an explicit parameter for value of homophily and also type of filter i.e full-pass, low-pass and high-pass so that we can test the performance of model on varying levels of homophily and on various filters. We plot the value of Probabilistic Bayes Error on varying values of homophily for low, high and full pass graph filters.

1.1 Metrics for Homophily

The homophily metrics studied have been described below. These metrics have been used to re-evaluate the Citeseer graph, after varying its structure by adjusting the homophily level.

1.1.1 Edge Homophily

Edge homophily in network theory is the tendency for connections (edges) to occur more frequently between nodes with similar attributes. This concept quantifies the likelihood of links forming based on shared characteristics of nodes, indicating a preference in network structure for assortative connections based on these attributes. It is given by:

$$H_{edge}(\mathcal{G}) = \frac{|\{e_{uv}|e_{uv} \in \mathcal{E}, Z_{u,:} = Z_{v,:}\}|}{|\mathcal{E}|}$$

1.1.2 Node Homophily

Node homophily, in network theory, refers to the propensity of nodes within a network to establish connections predominantly with other nodes possessing analogous attributes. This principle posits that the likelihood of a link forming between two nodes is significantly higher when their respective attributes exhibit similarity, thereby impacting the network's structural formation and dynamics. It is given by:

$$H_{node}(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \frac{\{u|u \in \mathcal{N}_v, Z_{u,:} = Z_{v,:}\}}{d_v}$$

1.1.3 Class Homophily

Class homophily in network analysis denotes the tendency for nodes (individuals) within a network to form connections predominantly with others in the same sociodemographic category, such as race, ethnicity, or socioeconomic status. This concept highlights that links in social networks are often more prevalent among individuals sharing similar class-based characteristics. It is given by:

$$H_{class}(G) = \frac{1}{C-1} \sum_{k=1}^{C} [h_k - \frac{|\{v|Z_{v,k=1}\}|}{N}]$$

1.1.4 Generalised Edge Homophily

Generalized edge homophily in network theory refers to the propensity for edges in a network to connect nodes that are similar not just in their intrinsic attributes but also in their relational or structural properties. This concept extends beyond individual characteristics to encompass similarities in nodes' positions or roles within the network's overall structure. It is given by:

$$H_{GE}(\mathcal{G}) = \frac{\sum_{(i,j)\in\mathcal{E}} cos(x_i, x_j)}{|\mathcal{E}|}$$

1.1.5 Adjusted Homophily

Adjusted homophily in network analysis quantifies the extent of homophily after accounting for structural constraints of the network. It differentiates between homophily arising from individual preference and that due to network topology, offering a more accurate measure of the intrinsic tendency for similar nodes to form connections beyond network-imposed limitations.

$$H_{adj} = \frac{H_{edge} - \sum_{c=1}^{C} \bar{p}_c^2}{1 - \sum_{k=1} C\bar{p}_c^2}$$

1.2 Measuring Node Distinguishability of CSBM-H

1.2.1 Probabilistic Bayes Error

Bayes error rate is the probability of a node being misclassified by the bayes classifier. We use this metric to measure the node distinguishability of the CSBM-H generated graph. In CSBM-H we have nodes belonging to two classes, so the PBE for CSBM-H is the expected value of node being misclassified given the node belongs to either of the two classes. $\mathcal{P}(CL_{Bayes}(x) = 0|x \in \mathcal{C}_0) = 1 - CDF(-\xi)_{\mathcal{X}(w_0,F_h,\lambda_0)}$ $\mathcal{P}(CL_{Bayes}(x) = 1|x \in \mathcal{C}_1) = CDF(-\xi)_{\mathcal{X}(w_1,F_h,\lambda_1)}$

Assuming we have $\mathcal{P}(x \in \mathcal{C}_0) = \mathcal{P}(x \in \mathcal{C}_1) = 1/2$

PBE is given by

$\frac{1}{2}((1 - \mathcal{P}(CL_{Bayes}(x) = 0|x \in \mathcal{C}_0)) + (1 - \mathcal{P}(CL_{Bayes}(x) = 1|x \in \mathcal{C}_1)))$

Higher PBE implies lesser node distinguishability.

1.2.2 Generalized Jeffreys Divergence

The KL-divergence is a statistical measure of how a probability distribution P is different from another distribution Q. Using this the paper has defined another measure for node distinguishability, generalized jeffreys divergence. The negative generalized Jeffreys divergence D_{NGJ} for CSBM-H is computed as follows:

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

where $d_X^2 = (\mu_0 - \mu_1)^T(\mu_0 - \mu_1)$ is the squared euclidean distance between centers. $\rho = \frac{\sigma_0}{\sigma_1}$. For h and h^{HP} we have $d_H^2 = (2h - 1)^2 d_X^2$ and $d_{HP}^2 = 4(1 - h)^2 d_X^2$. We can see that D_{NGJ} depends on two terms. The first term is known as Expected Negative Normalized Distance (ENND). ENND depends on how large is the inter-class ND d_X^2 compared to normalization term which is determined by intra-class ND i.e class variances σ_0 and σ_1. The second term is known as Negative Variance Ratio (NVR). NVR depends on how different the two intra-class NDs (σ_0 and σ_1) are. If D_{NGJ} is smaller it implies nodes are more distinguishable and vice-versa.

2 Dataset

1. **Citeseer Dataset :** citeseer dataset is a citation network which consists of 3327 publications classified into 6 classes and consists of 4732 edges. Each citation consists of 3703 features.

3 Experiments

3.1 Homophily modification

We tried to modify the edge homophily of citeseer dataset by removing and adding intraclass edges to the given graph.

3.2 CSBM-H

CSBM-H is a variation of CSBM which takes additional homophily parameter and generate the data. CSBM-H generates graphs consisting of two disjoint sets of nodes corresponding to two separate classes, features for these nodes are sampled from two separate normal distributions for which means and variances are passed as parameters. The degree of nodes in each class is also passed as a parameter. For each node, $h.d$ are intra-class neighbors and $(1 - h).d$ are inter-class neighbours, where h is homophily parameter and d is the degree of the node of the particular class.

We have performed all the experiments in the google colab environment. In each experiment we test the accuracy of the CSBM-H model using the Probabilistic Bayes Error rate and generalized Jeffrey's Divergence on varying values of homophily and different filters. In each experiment, we also vary the standard deviation values of features generated and degree of classes of the CSBM-H model and test it's effect on the error.

3.3 Results

3.3.1 Citeseer Homophily

For citeseer dataset, we obtained the homophily value for above mentioned metrics as follows:

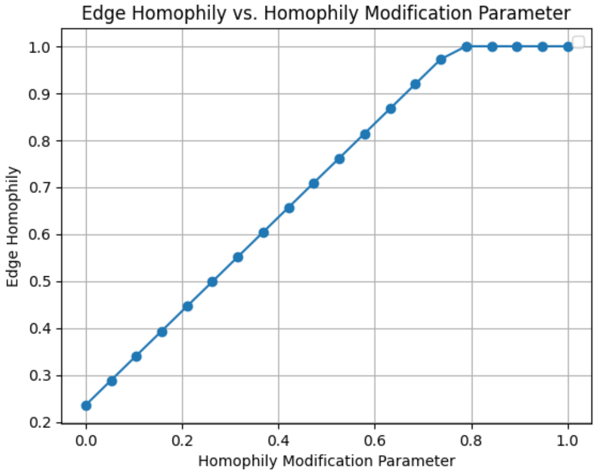

Figure 1: Edge homophily vs modification parameter

Metric	Value
Node Homophily	0.706249161762248
Edge Homophily	0.7355008787346221
Class Homophily	0.6283437645725868
Generalised Edge Homophily	0.19414880454437136
Adjusted Homophily	0.7355008787346221

Table 1: Homophily Metrics for the Citeseer Dataset

3.3.2 Homophily modification

We first tried to reduce homophily by removing 10% intra-class edges and adding 10% inter-class edges from the graph and homophily value reduced to 0.6355448154657294.

We tried to increase homophily by adding 10% intra-class edges and removing 10% inter-class edges from the graph and homphily value increased to 0.835456942003515.

Figure 1 plots the edge homophily when modification parameter is changed.

3.3.3 Plots for normal distribution

Figure 2 contains a plot for PDF of 3 normal distributions with mean and variance as (0, 5), (0, 10) and (5, 5) respectively.

3.3.4 CSBM-H results

CSBM-H was initialized with $\mu_0 = [-1, 0]$, $\mu_1 = [0, 1]$, $\sigma_0^2 = 1$, $\sigma_1^2 = 2$, $d_0 = 5$ and $d_1 = 5$. The number of nodes for each class was set to 25 and each node contains 2 features. From Figure 3 we can observe that PBE and D_{NGJ} plot for low pass filtered features is a bell-shaped curve indicating mid-level value of homophily is more detrimental to node distinguishability rather than a low level of homphily. This phenomenon is also known as mid-

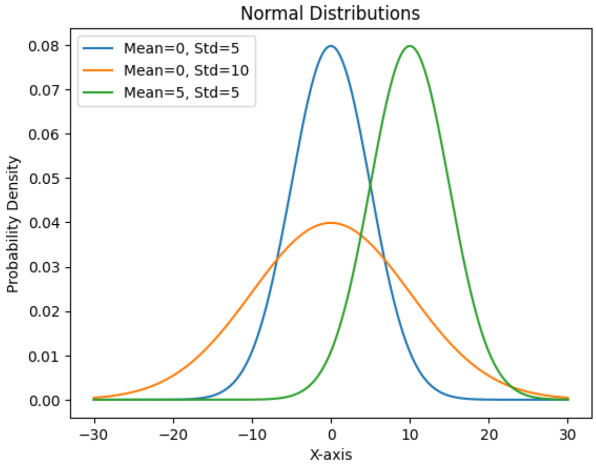

Figure 2: Normal distribution with varying mean and variances

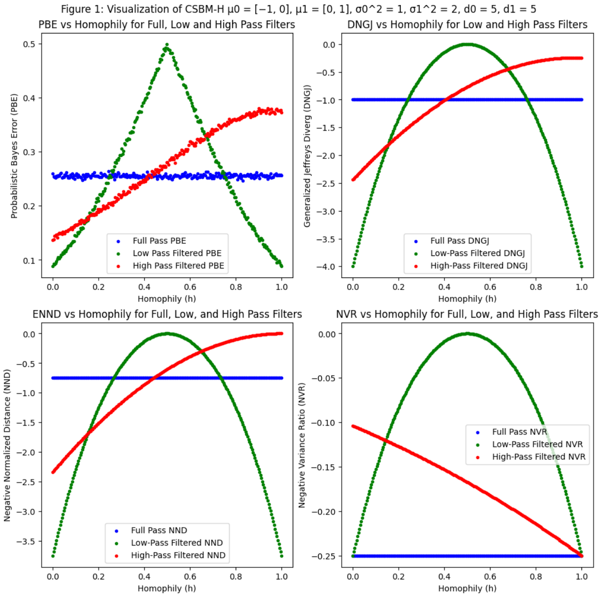

Figure 3: Visualization of CSBM-H $\mu_0 = [-1, 0]$, $\mu_1 = [0, 1]$, $\sigma_0^2 = 1$, $\sigma_1^2 = 2$, $d_0 = 5$, $d_1 = 5$

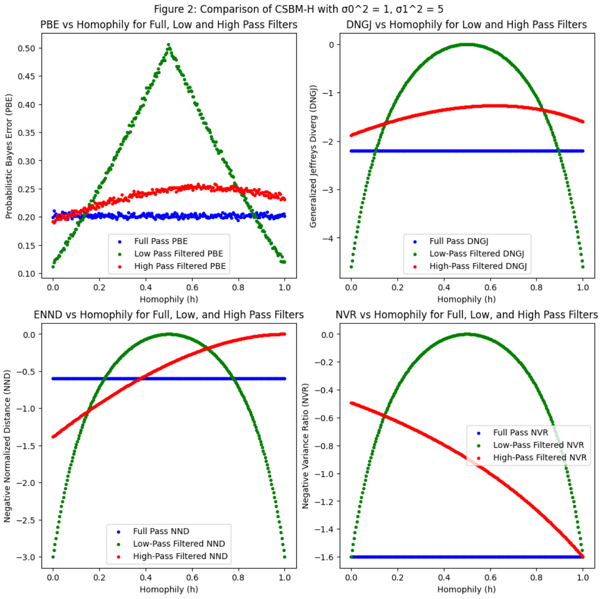

Figure 4: Comparison of CSBM-H with $\sigma_0^2 = 1$, $\sigma_1^2 = 5$

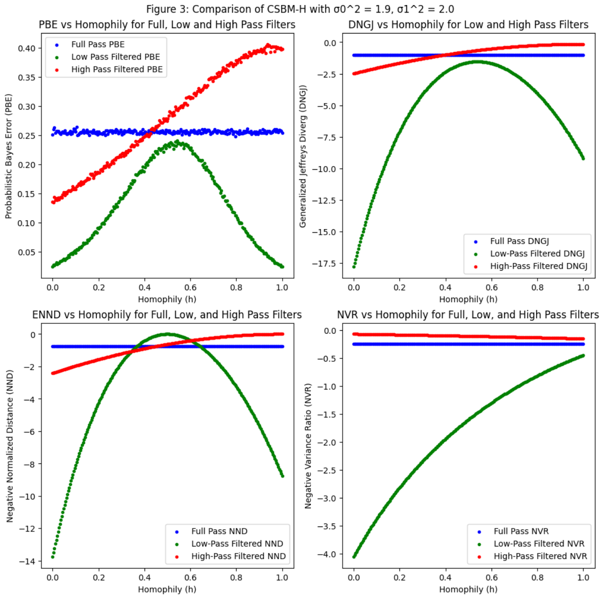

Figure 5: Comparison of CSBM-H with $\sigma_0^2 = 1.9$, $\sigma_1^2 = 2$

homophily pitfall. It can also be observed that LP, HP and FP-filtered features attain the lowest point in different homophily regions, dividing the entire range into various regimes. It can be observed that LP works better for very high and very low values of homophily, HP filter works better for low to medium values of homophily and FP filter works better for medium to high Values of homophily.

3.3.5 Ablation results

1. **Increase the Variance of High-variation Class :** Increasing the variance of class 1 from 2.0 to 5.0 makes the variance between both classes imbalanced. PBE and D_{NGJ} values for all three filtered features go up indicating embeddings have become less distinguishable. We can also observe shrinkage of HP regime and expansion of FP regime suggesting that original features are more robust to imbalanced variances.

2. **Increase the Variance of Low-variation Class :** Increasing the variance of class 0 from 1.0 to 1.9 makes the variance between both classes less imbalanced. PBE and D_{NGJ} values for all three filtered features go up indicating embeddings have become less distinguishable and LP, HP and FP regime almost stays the same

3. **Increase the Node Degree of High-variation Class :** Increasing d_1 from 5 to 25 causes the LP-filtered feature to go down leading to expansion of the LP regime and shrinkage of FP and HP regimes

4. **Increase the Node Degree of Low-variation Class :** Increasing d_0 from 5 to 25 leads to the same result as above case but the expansion and shrinkages are not as significant as before

4 References

Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, Doina Precup: When Do Graph Neural Networks Help with Node Classification? Investigating the Impact of Homophily Principle on Node Distinguishability

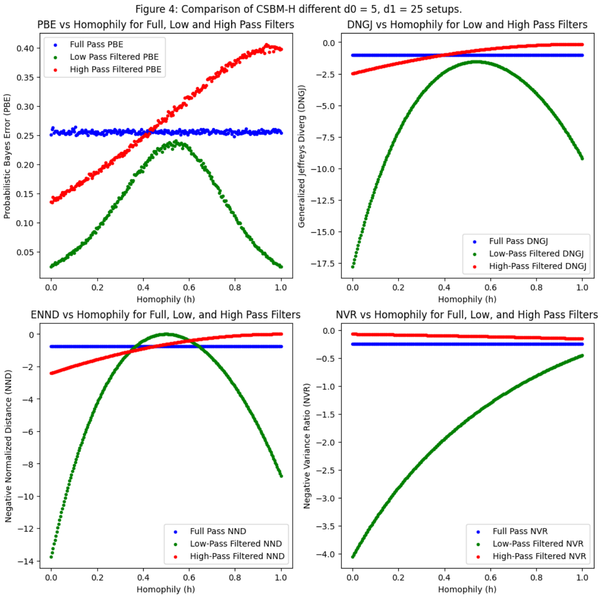

Figure 6: Comparison of CSBM-H different d_0 = 5, d_1 = 25 setups

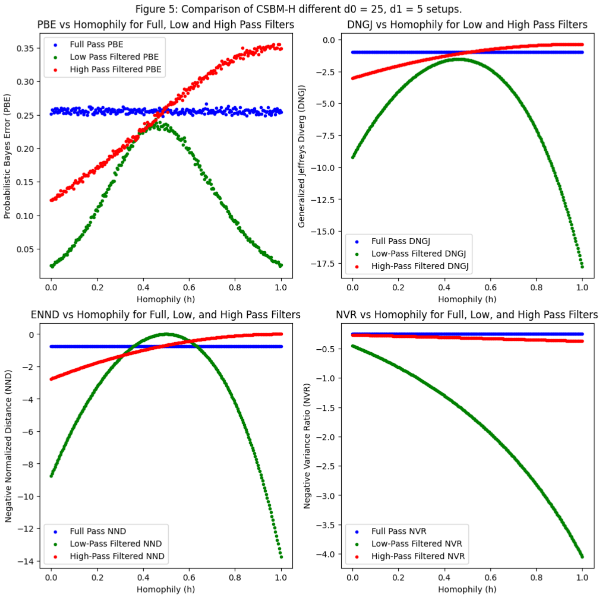

Figure 7: Comparison of CSBM-H different d_0 = 25, d_1 = 5 setups