Characterizing Graph Datasets for Node Classification: Homophily–Heterophily Dichotomy and Beyond

Oleg Platonov[*]
HSE University, Yandex Research
olegplatonov@yandex-team.ru

Denis Kuznedelev
Yandex Research, Skoltech
dkuznedelev@yandex-team.ru

Artem Babenko
Yandex Research
artem.babenko@phystech.edu

Liudmila Prokhorenkova[*]
Yandex Research
ostroumova-la@yandex-team.ru

Abstract

Homophily is a graph property describing the tendency of edges to connect similar nodes; the opposite is called *heterophily*. Much effort has been put into developing efficient methods for learning on heterophilous graphs. However, there is no universally agreed-upon measure of homophily in the literature. In this work, we show that commonly used homophily measures have critical drawbacks preventing the comparison of homophily levels across different datasets. For this, we formalize desirable properties for a proper homophily measure and verify which measures satisfy which properties. In particular, we show that a measure that we call *adjusted homophily* satisfies more desirable properties than other popular homophily measures while being rarely used in graph machine learning literature. Then, we go beyond the homophily–heterophily dichotomy and propose a new characteristic allowing one to further distinguish different sorts of heterophily. The proposed *label informativeness* (LI) characterizes how much information a neighbor's label provides about a node's label. We prove that this measure satisfies important desirable properties and also observe empirically that LI better agrees with GNN performance compared to homophily measures.

1 Introduction

In many real-world networks, edges tend to connect similar nodes, this property is usually called *homophily*. The opposite is *heterophily*: e.g., in social networks, fraudsters rarely connect to other fraudsters. Early research on Graph Neural Networks (GNNs) mainly focused on homophilous graphs, and recently there have been discussions whether specialized models are needed for the heterophilous setting [1–6].

To measure the level of homophily, several *homophily measures* are used in the literature [1–3, 7], but these measures may significantly disagree with each other. In this work, we address the problem of how to properly measure the homophily level of a graph. Motivated by recent studies of clustering and classification performance measures [8, 9], we formalize some desirable properties of a reasonable homophily measure and check which measures satisfy which properties. Our analysis reveals that commonly used homophily measures do not satisfy some important properties and cannot be compared across datasets with different number of classes and class size balance. In contrast, a measure that we call *adjusted homophily* (a.k.a. *assortativity coefficient*) satisfies most of the desirable properties while being rarely used in graph machine learning literature.

Then, we propose a new graph property called *label informativeness* (LI) that allows one to distinguish different heterophily patterns. LI characterizes how much information the neighbor's label provides

[*]Equal contribution.

O. Platonov et al., Characterizing Graph Datasets for Node Classification: Homophily–Heterophily Dichotomy and Beyond (Extended Abstract). Presented at the Second Learning on Graphs Conference (LoG 2023), Virtual Event, November 27–30, 2023.

about the node's label. We analyze this measure via the same theoretical framework and show that it satisfies important properties and thus can be compared across different datasets. Moreover, our experiments on synthetic and semi-synthetic datasets show that LI better agrees with GNN performance compared to homophily measures. Thus, LI is a useful graph characteristic.

This paper is an extended abstract that briefly describes the main ideas of our work. More details, proofs, and experiments can be found in the full paper [10].

2 Homophily measures

2.1 Desirable properties for homophily measures

Let us start with the necessary notation. Assume that we are given a simple (without self-loops and multiple edges) and undirected graph $G = (V, E)$ with nodes V, $|V| = n$, and edges E. Each node $v \in V$ has a class label $y_v \in \{1, \dots, C\}$. Let $n_k = |\{v : y_v = k\}|$ denote the size of k-th class. By $N(v)$ we denote the neighbors of v in G and by $d(v) = |N(v)|$ the degree of v. Also, let $D_k := \sum_{v\,:\,y_v=k} d(v)$. Let $p(\cdot)$ denote the empirical distribution of class labels, i.e., $p(k) = \frac{n_k}{n}$. We also define degree-weighted distribution as $\bar{p}(k) = \frac{\sum_{v\,:\,y_v=k} d(v)}{2|E|} = \frac{D_k}{2|E|}$.

Now, let us propose a list of properties desirable for a good homophily measure.

Maximal agreement. This property requires that perfectly homophilous graphs achieve a constant upper bound of the measure. Formally, we say that a homophily measure h satisfies maximal agreement if for any graph G in which $y_u = y_v$ for all $\{u, v\} \in E$ we have $h(G) = c_{\max}$. For all other graphs G, we require $h(G) < c_{\max}$.

Minimal agreement. We say that a homophily measure h satisfies minimal agreement if $h(G) = c_{\min}$ for any graph G in which $y_u \neq y_v$ for all $\{u, v\} \in E$. For all other graphs G, we require $h(G) > c_{\min}$. In other words, if all edges connect nodes of different classes, we expect to observe a constant minimal value.

Constant baseline. This property ensures that homophily is not biased towards particular class size distributions. Intuitively, if the graph structure is independent of labels, we would expect a low homophily value. Moreover, if we want a measure to be comparable across datasets, we expect the same low value in all such cases. We formalize this intuition via the so-called *configuration model*: take n nodes, assign each node v degree $d(v)$, and randomly pair edge endpoints to obtain a graph.

Definition 1. A homophily measure h has *asymptotic constant baseline* if for G generated according to the configuration model and for any $\varepsilon > 0$ with probability $1 - o(1)$ we have $|h(G) - c_{base}| < \varepsilon$ for some constant c_{base} as $n \to \infty$.

Empty class tolerance. This condition is required if a homophily measure has to be comparable across datasets with varying numbers of classes. A measure is tolerant to empty classes if it is defined and it does not change when we introduce an additional dummy label that is not present in the data.

Monotonicity. We define this property as follows.

Definition 2. A homophily measure is *monotone* if it is empty class tolerant, increases when we add an edge between two nodes of the same class (except for perfectly homophilous graphs), and decreases when we add an edge between two nodes of differen classes (except for perfectly heterophilous graphs).

As we discuss in [10], alternative definitions of monotonicity can be considered for general homophily measures. However, Definition 2 naturally aligns with *edge-wise* measures that we now define. First, we define a *class adjacency matrix* $\mathcal{C}$: each element c_{ij} denotes the number of edges (u, v) such that $y_u = i$ and $y_v = j$. Since the graph is undirected, the matrix $\mathcal{C}$ is symmetric. A homophily measure is *edge-wise* if it is a function of the class adjacency matrix. For edge-wise measures, monotonicity requires an increase when we increment a diagonal element by two and a decrease when we increment c_{ij} and c_{ji} by one for $i \neq j$.

2.2 Properties of existing homophily measures

We first discuss the properties of three popular homophily measures. Then, we introduce adjusted homophily and show that it satisfies many desirable properties.

Edge homophily. Edge homophily [1, 3] is the fraction of edges that connect nodes of the same class: $h_{edge} = \frac{|\{\{u,v\} \in E : y_u = y_v\}|}{|E|}$. It satisfies maximal and minimal agreement and is empty class tolerant and monotone. However, it does not satisfy asymptotic constant baseline, which is a critical drawback: one can get misleading results in settings with imbalanced classes.

Node homophily. Node homophily [2] computes the fraction of neighbors that have the same class for all nodes and then averages these values across the nodes: $h_{node} = \frac{1}{n} \sum_{v \in V} \frac{|\{u \in N(v) : y_u = y_v\}|}{d(v)}$. It satisfies maximal and minimal agreement. It is empty class tolerant, but monotoncity is violated since adding an edge between two perfectly homophilous nodes does not change the value. Also, node homophily does not satisfy the asymptotic constant baseline.

Class homophily. Class homophily [7] is another recently proposed measure:

$$h_{class} = \frac{1}{C-1} \sum_{k=1}^{C} \left[\frac{\sum_{v : y_v = k} |\{u \in N(v) : y_u = y_v\}|}{\sum_{v : y_v = k} d(v)} - \frac{n_k}{n} \right]_+ ,$$

where $[x]_+ = \max\{x, 0\}$. It satisfies maximal agreement with $h_{class} = 1$, but minimal agreement is not satisfied. Class homophily is not empty class tolerant and thus is not monotone. Additionally, it does not have the asymptotic constant baseline.

Adjusted homophily. Adjusted homophily is obtained by taking edge homophily, subtracting its expected value, and normalizing by the maximum value:

$$h_{adj} = \frac{h_{edge} - \sum_{k=1}^{C} \bar{p}(k)^2}{1 - \sum_{k=1}^{C} \bar{p}(k)^2}. \tag{1}$$

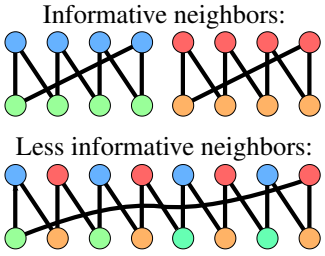
This measure is rarely used in graph ML literature, but is known in graph analysis literature as *assortativity coefficient* [11]. Our theoretical analysis shows that when used as a homophily measure, h_{adj} satisfies important desirable properties.

Theorem 1. *Adjusted homophily satisfies maximal agreement, asymptotic constant baseline, and empty class tolerance. The minimal agreement is not satisfied. Moreover, this measure is monotone if* $h_{adj} > \frac{\sum_i \bar{p}(i)^2}{\sum_i \bar{p}(i)^2 + 1}$ *and we note that the bound* $\frac{\sum_i \bar{p}(i)^2}{\sum_i \bar{p}(i)^2 + 1}$ *is always smaller than 0.5. When* h_{adj} *is small, counterexamples to monotonicity exist.*

While adjusted homophily violates some properties, it still dominates all other measures and is comparable across different datasets with varying numbers of classes and class size balance. We recommend using it as a more reliable measure of homophily in further works.

3 Label informativeness

Since heterophily is the negation of homophily, heterophilous graphs may have very different connectivity patterns (as shown on the right). To distinguish such patterns, we define a characteristic measuring the informativeness of a neighbor's label for a node's label.



Formally, assume that we sample an edge $(\xi, \eta) \in E$ (from some distribution). The class labels of nodes ξ and η are then random variables y_ξ and y_η. We want to measure the amount of knowledge the label y_η gives for predicting y_ξ. This can be measured via the mutual information, which is the reduction of entropy $H(y_\xi)$ if we know the value y_η: $I(y_\xi, y_\eta) = H(y_\xi) - H(y_\xi | y_\eta)$. To make the obtained quantity comparable across different datasets, we say that *label informativeness* is the normalized mutual information of y_ξ and y_η:

$$\mathrm{LI} := I(y_\xi, y_\eta) / H(y_\xi)\,. \tag{2}$$

We have $\mathrm{LI} \in [0, 1]$. If the label y_η allows for unique reconstruction of y_ξ, then $\mathrm{LI} = 1$. If y_ξ and y_η are independent, $\mathrm{LI} = 0$.

If the edges are sampled uniformly at random, (2) becomes:

$$\mathrm{LI}_{edge} = -\frac{\sum_{c_1,c_2} p(c_1,c_2) \log \frac{p(c_1,c_2)}{\bar{p}(c_1)\bar{p}(c_2)}}{\sum_c \bar{p}(c) \log \bar{p}(c)} = 2 - \frac{\sum_{c_1,c_2} p(c_1,c_2) \log p(c_1,c_2)}{\sum_c \bar{p}(c) \log \bar{p}(c)},$$

where $p(c_1, c_2) = \sum_{(u,v)\in E} \frac{\mathbb{1}\{y_u=c_1, y_v=c_2\}}{2|E|}$. For brevity, we further denote LI_{edge} by LI.

To claim that LI is a suitable graph characteristic, we need to show that it is comparable across different datasets. For this, we need to verify maximal agreement and asymptotic constant baseline. Recall that LI is upper bounded by one and equals one if and only if the neighbor's class uniquely reveals the node's class. This property can be considered as a direct analog of the maximal agreement defined in Section 2.1. The following proposition shows that LI satisfies the asymptotic constant baseline.

Proposition 1. *Assume that $|E| \to \infty$ as $n \to \infty$ and that the entropy of $\bar{p}(\cdot)$ is bounded from below by some constant. Let $\bar{p}_{min} = \min_k \bar{p}(k)$ and assume that $\bar{p}_{min} \gg C/\sqrt{|E|}$ as $n \to \infty$. Then, for the random configuration model, we have $\mathrm{LI} = o(1)$ with high probability.*

Additionally, we empirically observe that LI better correlates with GNN performance than homophily. For this, we generate synthetic graphs via a variant of the *stochastic block model* (SBM) [12]. We carefully design parameter combinations (community-to-community edge probabilities) to obtain various combinations of dataset characteristics — LI and homophily. Given the class labels, the features are taken from the four largest classes in the `cora` dataset [13–16]. For each obtained graph, we train the **GraphSAGE** model [17] and measure its classification accuracy. Figure 1 shows that the model performance is much more correlated with LI than with homophily. In particular, when LI is high, GraphSAGE achieves good performance even on strongly heterophilous graphs with negative homophily. We refer to the full paper [10] for the details and more experiments on various datasets.

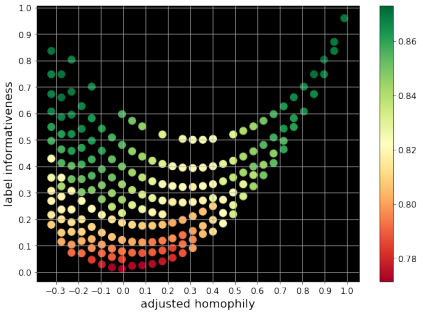


Figure 1: Accuracy of GraphSAGE on synthetic SBM graphs

In summary, our results show that LI is a meaningful graph characteristic that complements homophily measures and can be useful for both graph analysis and graph machine learning.

4 Conclusion

In this paper, we discuss how to characterize graph node classification datasets. First, we revisit the concept of homophily and show that commonly used homophily measures have significant drawbacks. For this, we formalize properties desirable for a good homophily measure and show which measures satisfy which properties. We conclude that *adjusted homophily* is a better measure of homophily than the ones commonly used in the literature. We recommend using it to estimate and compare homophily levels of various graphs in future works.

Then, we argue that heterophilous graphs may have very different structural patterns and propose a new property called *label informativeness* (LI) that allows one to distinguish them. LI characterizes how much information a neighbor's label provides about a node's label. Similarly to adjusted homophily, this measure satisfies important properties and thus can be used to compare datasets with different numbers of classes and class size balance. Through a series of experiments, we show that LI correlates well with the performance of GNNs.

To conclude, we believe that adjusted homophily and label informativeness will be helpful for researchers and practitioners as they allow one to easily characterize the connectivity patterns of graph datasets.

References

[1] Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. MixHop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *International Conference on Machine Learning*, pages 21–29. PMLR, 2019. 1, 3

[2] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020. 3

[3] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in Neural Information Processing Systems*, 33:7793–7804, 2020. 1, 3

[4] Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? In *International Conference on Learning Representations*, 2022.

[5] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[6] Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress? In *The Eleventh International Conference on Learning Representations*, 2023. 1

[7] Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3

[8] Martijn Gösgens, Anton Zhiyanov, Aleksey Tikhonov, and Liudmila Prokhorenkova. Good classification measures and how to find them. *Advances in Neural Information Processing Systems*, 34:17136–17147, 2021. 1

[9] Martijn M Gösgens, Alexey Tikhonov, and Liudmila Prokhorenkova. Systematic analysis of cluster similarity indices: How to validate validation measures. In *International Conference on Machine Learning*, pages 3799–3808. PMLR, 2021. 1

[10] Oleg Platonov, Denis Kuznedelev, Artem Babenko, and Liudmila Prokhorenkova. Characterizing graph datasets for node classification: Homophily-heterophily dichotomy and beyond. *Advances in Neural Information Processing Systems*, 2023. 2, 4

[11] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2), 2003. 3

[12] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983. 4

[13] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–93, 2008. 4

[14] Galileo Namata, Ben London, Lise Getoor, Bert Huang, and UMD EDU. Query-driven active surveying for collective classification. In *10th International Workshop on Mining and Learning with Graphs*, volume 8, page 1, 2012.

[15] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning*, pages 40–48. PMLR, 2016.

[16] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000. 4

[17] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017. 4