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

Sikta Mohanty

June 10 2024

Table of Contents

1. Abstract
2. Introduction
3. Homophily Measures
4. Popular homophily measures
5. Desirable Properties for homophily measures
6. Properties of popular homophily measures
7. Adjusted Homophily
8. Label Informativeness (LI)
9. Empirical Illustrations
10. Correlation of LI with GNN performance
11. Conclusion
12. Limitations

Abstract

- Homophily: Graph property where edges tend to connect similar nodes.
- Heterophily: Opposite of homophily, where edges connect dissimilar nodes.
- Issue: Lack of universally agreed-upon measure for homophily.
- Label Informativeness (LI): Proposed to further distinguish different types of heterophily based on neighbor label information.

Introduction

- **Graphs in Machine Learning:**
  - Graphs represent data from various domains like social networks, citation networks, etc.
  - Graph Neural Networks (GNNs) have shown strong results in machine learning on graph-structured data.
- **Importance of Homophily:**
  - Homophily: Nodes tend to connect with similar nodes, a common property in real-world networks.
  - Early GNNs focused on homophilous graphs, but heterophilous graphs pose challenges.
- **Problem Statement:**
  - Existing homophily measures vary and may disagree with each other.
  - Need for a proper measure to quantify the level of homophily in graphs.
- **Proposed Solution:**
  - Formalization of desirable properties for a homophily measure.
  - Introduction of adjusted homophily (assortativity coefficient) as a better alternative.
  - Proposal of a new measure, label informativeness (LI), to distinguish heterophily patterns.

- **Theoretical Framework and Analysis:**
  - Analysis of adjusted homophily and LI against desirable properties.
  - Recommendation to use adjusted homophily for measuring homophily and LI for characterizing heterophily.
- **Empirical Observations:**
  - LI better aligns with GNN performance compared to traditional homophily measures.
  - Illustrates why GNNs perform well on heterophilous datasets.
  - Proposed theoretical framework for choosing suitable characteristics in node classification tasks.
  - Recommendation to use adjusted homophily and LI for measuring and characterizing graph connectivity patterns.

Homophily Measures

Assume that we are given a graph $G = (V, E)$ with nodes $V, |V| = n$, and edges E. Throughout the paper, it is assumed that the graph is simple (without self-loops and multiple edges) and undirected. Each node $v \in V$ has a class label $y_v \in \{1, \ldots, C\}$. Let n_k denote the size of the k-th class, i.e., $n_k = |\{v : y_v = k\}|$. By $N(v)$ we denote the neighbors of v in G and by $d(v) = |N(v)|$ the degree of v.

Popular homophily measures

- **Edge homophily** is the fraction of edges that connect nodes of the same class:

$$h_{\text{edge}} = \frac{|\{(u,v) \in E : y_u = y_v\}|}{|E|}$$

- **Node homophily** computes the fraction of neighbors that have the same class for all nodes and then averages these values across the nodes:

$$h_{\text{node}} = \frac{1}{n} \sum_{v \in V} \frac{|\{u \in N(v) : y_u = y_v\}|}{d(v)}$$

- **Class homophily** used to understand how nodes in a graph are connected based on their labels or classes.

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

However, there are still some issues with class homophily,especially when nodes of different classes have different degrees

Desirable Properties for homophily measures

A list of properties desirable for a good homophily measure:

- **Maximal Agreement:** 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 $h(G) = c_{\max}$.
- **Minimal Agreement:** 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 $\{u, v\} \in E$.
- **Asymptotic constant Baseline:** This property ensures that homophily is not biased towards particular class size distributions. This is done using configuration model.

- **Empty Class Tolerance:** Homophily measures need to be comparable across datasets with varying numbers of classes.
- **Monotonicity:** A homophily measure is monotone if it is empty class tolerant and increases when we add an edge between two nodes of the same class and decreases when we add an edge between two nodes of different classes

Table 1: Properties of Homophily Measures

Measure	Maximal Agreement	Minimal Agreement	Asymptotic Constant Baseline	Empty Class Tolerance	Monotonicity
Edge homophily	✓	✓	×	✓	✓
Node homophily	✓	✓	×	✓	×
Class homophily	✓	×	×	×	×
Adjusted homophily	✓	×	✓	✓	×*

Adjusted Homophily

Adjusted homophily is a less common way to measure homophily, which is how much nodes of the same type tend to connect in a graph.
Here's how it's calculated:
Basic Measure: Edge homophily is the fraction of edges in a graph that connect nodes of the same class.
Configuration Model: The probability that a given edge endpoint will be connected to a node with class k is approximately:

$$\frac{\sum_{v:y_v=k} d(v)}{2|E|}$$

where:

- $\sum_{v:y_v=k} d(v)$ is the total degree of all nodes in class k.
- $|E|$ is the number of edges.

Expected Edge Homophily: The expected value of edge homophily under random connections is:

$$\sum_{k=1}^{C} \left(\frac{D_k}{2|E|} \right)^2$$

where $D_k = \sum_{v:y_v=k} d(v)$.

Adjusted Edge Homophily: Subtract the expected value from the observed edge homophily to enforce the constant baseline property:

$$h_{\text{adjusted}} = h_{\text{edge}} - \sum_{k=1}^{C} \left(\frac{D_k}{2|E|} \right)^2$$

Normalization: To ensure the measure reaches a constant upper bound in perfectly homophilous graphs, we normalize the adjusted measure:

$$h_{\text{adj}} = \frac{h_{\text{edge}} - \sum_{k=1}^{C} \left(\frac{D_k}{2|E|}\right)^2}{1 - \sum_{k=1}^{C} \left(\frac{D_k}{2|E|}\right)^2}$$

The denominator $1 - \sum_{k=1}^{C} \left(\frac{D_k}{2|E|}\right)^2$ scales the measure to ensure it is within a consistent range, achieving maximal agreement.
The formula for adjusted homophily looks like this:

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

where we use the notation $\bar{p}(k) = \frac{D_k}{2|E|}$.

Label Informativeness (LI)

What is Label Informativeness?

- Label Informativeness (LI) measures how much information a neighboring node's label provides about the label of a node itself.
- Particularly useful in heterophilous graphs where similar nodes may not be directly connected.

Why is Label Informativeness Important?

- Understanding the pattern of connections in heterophilous graphs can improve models using graph structure for tasks like node classification.

How is Label Informativeness Calculated?

- Pick an edge (connection) in the graph randomly.Let y_ξ and y_η be the class labels of the two nodes connected by the edge.
- Entropy without Neighbor Information $H(y_\xi)$: Measures the difficulty of predicting a node's label without knowing its neighbor's label.
- Entropy with Neighbor Information $H(y_\xi \mid y_\eta)$: Measures the difficulty of predicting a node's label given its neighbor's label.
- Mutual Information: Difference between the two entropies, indicating how much information the neighbor's label provides.

$$I(y_\xi, y_\eta) = H(y_\xi) - H(y_\xi \mid y_\eta)$$

- Normalized Mutual Information: Normalizes mutual information by dividing by the entropy of the node's label.

$$\mathrm{LI} = \frac{I(y_\xi, y_\eta)}{H(y_\xi)}$$

- Range of LI: Ranges from 0 to 1, where 1 implies the neighbor's label fully determines the node's label and 0 implies no information.

Empirical Illustrations

Table 2: Characteristics of some real graph datasets,

Dataset	C	hedge	hadj	LI
lastfm-ais	18	0.87	0.86	0.75
cora	7	0.81	0.77	0.59
ogbn-arxiv	40	0.65	0.59	0.45
twitter-hate	2	0.78	0.55	0.23
wiki	5	0.38	0.15	0.06
twitch-gamers	2	0.55	0.09	0.01
actor	5	0.22	0.00	0.00
questions	2	0.84	0.02	0.00
roman-empire	18	0.05	-0.05	0.11

Correlation of LI with GNN performance

Hypothesis The authors hypothesize that GNNs can learn complex relationships beyond just homophily. Therefore, GNN performance should correlate more strongly with Label Informativeness (LI) than with homophily.
Synthetic Data Based on SBM Model To test their hypothesis, the authors generate synthetic graphs using a variant of the Stochastic Block Model (SBM).
Synthetic Data Generation Cluster Setup: Nodes are divided into $C = 4$ classes and class size in n/4.
Edge Probabilities:

- p_0: Probability of an edge within the same class(i=j).
- p_1: Probability of an edge between specific pairs of classes(i+j=5).
- p_2: Probability of an edge between other pairs of classes.

Experiment Setup

- Generated graphs have an expected node degree of 10.
- Features are taken from the four largest classes in the Cora dataset.
- Four GNN models are tested: GCN, GraphSAGE, GAT, and Graph Transformer (GT).
- Over 20,000 experiments are run for each model.

Results

Model	h_{adj}	LI
GCN	0.19	0.76
GraphSAGE	0.05	0.93
GAT	0.17	0.77
Graph Transformer	0.17	0.77

Table 3: Spearman Correlation Between Model Accuracy and Characteristics of Synthetic SBM Datasets

Observations

- LI has a much higher correlation with GNN performance than h_{adj}.
- For GraphSAGE, the correlation between accuracy and LI is 0.93, while between accuracy and h_{adj} it is only 0.05.

Semi-Synthetic Data Setup: Real-world graphs modified to have varying levels of homophily by adding inter-class edges in different patterns.

Results: Standard GNNs perform well on heterophilous graphs if they have high LI.

Model	h_{edge}	h_{node}	h_{class}	h_{adj}	LI
Cora					
GCN	-0.31	-0.31	-0.31	-0.31	0.72
GraphSAGE	-0.24	-0.24	-0.24	-0.24	0.78
GAT	-0.24	-0.25	-0.24	-0.24	0.77
Graph Transformer	-0.23	-0.24	-0.23	-0.23	0.79
Citeseer					
GCN	-0.24	-0.24	-0.24	-0.24	0.76
GraphSAGE	-0.53	-0.53	-0.54	-0.54	0.51
GAT	-0.27	-0.27	-0.27	-0.27	0.75
Graph Transformer	-0.19	-0.19	-0.19	-0.19	0.80

Table 4: Spearman Correlation Between Model Accuracy and Dataset Characteristics (Semi-Synthetic Datasets)

Key Findings

- The correlation between GNN performance and LI is positive and significant, while the correlation with homophily measures is negative.
- This indicates that LI is a better predictor of GNN performance than homophily.

Conclusion

In this paper, it is explored how to describe and understand graph node classification datasets.

- **Understanding Homophily:**
  - **Homophily:** This is the idea that similar nodes (like friends in a social network) are more likely to be connected.
  - **Problems with Existing Measures:** It is found that current ways to measure homophily have flaws, making it hard to compare homophily levels across different datasets.
  - **Adjusted Homophily:** It is suggested to use "adjusted homophily" as a better measure because it has properties that make comparisons more reliable.
- **Importance for GNN Development:**
  - **Heterophily-Suited GNNs:** For developing graph neural networks (GNNs) that work well with both homophilous and heterophilous graphs, it's crucial to accurately estimate homophily levels. Heterophilous graphs have nodes that connect to different types of nodes (e.g., different social circles).

- **Introducing Label Informativeness (LI):**
  - **Label Informativeness (LI):** A new concept called LI is proposed, which measures how much knowing a neighbor's label helps predict a node's label. This is important for understanding heterophilous graphs with varied structural patterns.
  - **Comparison Across Datasets:** LI, like adjusted homophily, has properties that allow it to be used to compare different datasets, even those with different class sizes and numbers of classes.
- **Experimental Findings:**
  - **Correlation with GNN Performance:** This experiments shows that LI is strongly related to how well GNNs perform, more so than traditional homophily measures.
- **Future Implications:**
  - **Usefulness:** Adjusted homophily and LI can help researchers and practitioners describe the connectivity patterns in graph datasets.
  - **Need for New Datasets:** New, realistic datasets will be created to explore different combinations of adjusted homophily and LI.
  - **Framework for Future Research:** This theoretical framework can aid in developing more reliable graph characteristics in the future.

Better homophily measures and LI enhance understanding of graph data and improve GNN development.

Limitations

- **Adjusted Homophily**:
  - Recommended for its constant baseline and desirable properties.
  - **Limitations**:
    - Violates minimal agreement criterion.
    - Guarantees monotonicity only for large values of h_{adj}.
- **Label Informativeness (LI)**:
  - Designed to be informative, simple to compute, and interpret.
  - **Limitations**:
    - Considers edges individually, ignoring the overall neighborhood.
    - Not a universal predictor of GNN performance, but correlates better than homophily.
- **Future Directions**:
  - Current analysis is limited to graph-label interactions.
  - Important to analyze node features and their interactions with graphs and labels.
  - Understanding feature-based homophily and informativeness could improve GNN performance insights.