**Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики».**

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ
НА ТЕМУ:

# «Characterizing Graph Datasets for Node Classification: Beyond Homophily–Heterophily Dichotomy»

**Студенты** Полокайнен Даниил Андреевич
Попов Илья Иванович
Раева Анастасия Алексеевна
Симонян Дмитрий Арташесович
Солодянкин Андрей Александрович
**Группа** ФТиАД21
**Руководитель** Масютин Алексей Александрович

Москва 2022 г.

# РЕФЕРАТ

Расчетно-пояснительная записка 21 с., 7 рис., 0 табл., 18 источников, 0 прил.

КЛЮЧЕВЫЕ СЛОВА: ГРАФ, ГОМОФИЛИЯ, ГЕТОРОФИЛИЯ, GNN, КЛАССИФИКАЦИЯ, МЕТРИКА КАЧЕСТВА КЛАССИФИКАЦИИ, МЕТРИКА ГОМОФИЛИИ

**Объектом исследования в НИР** являются метрики оценки гомофилии и качества классификации для графовых нейронных сетей для графов, подходящие для датасетов с различными размерами и количеством классов.

**Целью НИРС** является формализация подходящих метрик качества классификации и гомофилии, теоретическое и эмпирическое доказательство их корректности, а также сравнение разных метрик между собой и выявление лучшей.

В результате выполнения НИРС были получены следующие **результаты**:

- Получена метрика Adjusted Homophily, обладающая лучшими свойствами, по сравнению с существующими классическими метриками гомофилии.
- Получена метрика Label Informativeness (LI), подходящая для сравнения различных датасетов
- В результате серии экспериментов получено, что для GNN метрика Label Informativeness лучше, чем рассмтореныые маер гомофильности. Также получено, что LI объясняет, почему GNN иногда могут хорошо работать с гетерофильными наборами данных.

# Содержание

# ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В работе использованы следующие обозначения и сокращения:

GNN – graph neural networks

$LI$ – Label Informativeness

$E$ – Ребра графа

$V$ – Вершины графа

$G$ – Граф

$N(v)$ – Соседи вершины $v$

$d(V)$ – Степень вершины $v$

$X_v$ – Вектор признаков

$Y_v$ – Метки классов вершин графа

$p(.)$ – Эмпирическое распределение меток класса

# ВВЕДЕНИЕ

**Актуальность темы работы** обусловлена тем, что стандартные графовые нейронные сети плохо работают на не-гомофильных графах, однако, на некоторых графах с такого типа они показывают высокую эффективность. Поэтому важно иметь метрику, позволяющую стабильно отличать гомофильные и не-гомофильные графы, так как для них требуются разные подходы к обучению GNN. Также важно иметь метрику качества, показывающую стабильный результат на разных датасетах и позволяющую сделать результаты экспериментов интерпретируемыми и сопоставимыми.

**Целью работы**является проектирование метрики гомофильности графа, не зависящей от особенностей датасета и позволяющей сравнивать с её помощью различные графы между собой, а также - вывод универсальной метрики качества GNN.

Для достижения поставленной цели в работе решаются следующие **основные задачи**:

- Исследование существующих метрик гомофильности и их недостатков
- Определение новой метрики гомофильности, которая закрывает недостатки классических метрик, и её теоретическое обоснование
- Определение новой метрики качества GNN, выходящей за пределы подхода гомофилии графов, и её теоретическое обоснование
- Эмпирическое сравнение полученных метрик

Решение поставленных задач осуществляется с использованием следующих **методов и подходов**: для определения метрик исследованы существующие подходы оценки гомофилии:

1. Гомофилия рёбер
2. Гомофилия вершин
3. Гомофилия классов

А также определены необходимые свойства:

1. Maximal agreement
2. Minimal agreement
3. Constant baseline
4. Empty class tolerance

И введены основные понятия:

1. Монотонность
2. Модель конфигурации
3. Асимптотическое константное предсказание

При выполнении работы использованы следующие исходные данные:

1. Синтетические данные, полученные с помощью stochastic block model.
2. Полусинтетические данные, полученные путем добавления ребер между классами разными способами к нескольким реальным графам, что даёт несколько наборов графов с разной степенью гомофильности.
3. Синтетические данные, сгенерированные согласно подходу, описанному в [2].

# 1 Обзор литературы и направлений исследователей

## 1.1 Обозначения

Возьмем простой ненаправленный граф $G = (V,E)$ без петель и множественных ребер между одной парой вершин со множеством вершин $V$ и множеством ребер $E$ соответственно. Для каждой вершины $v$ определим вектор признаков $x_v$ и метка класса $y_v$. Пусть $nk$ обозначает размер класса k. $N(v)$ – соседи вершины $v$ в $G$, $d(v)$ – степень вершины $v$. За $p(.)$ примем эмпирическое распределение меток класса.

## 1.2 Классические меры гомофильности

Свойство гомофилии определяет то, насколько похожие вершины графа связаны. Сходство оценивается в терминах меток вершин или их признаков.

Классическими мерами гомофильности являются:

1. Гомофилия рёбер определяется как доля ребер, связывающих вершины одного класса, от общего количества ребер в графе:

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

2. Гомофилия вершин определяется как усредненная по всем вершинам доля соседей, принадлежащих одному классу, в общем количестве вершин

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

3. Гомофилия классов измеряется как отношение гомофильности текущей модели к гомофильности нулевой модели, где ребра не зависят от меток

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

Меры 1 и 2 имеют существенный недостаток – они чувствительны к балансу классов и их количеству. Этих недостатков нет у метрики гомофильности классов, однако, она не учитывает изменения степеней вершин.

## 1.3 Критерии для меры гомофильности

В статье авторы опирались на работы [3], [4]. В ней описаны критерии для метрик похожести кластеров, однако не все критерии метрик применимы для меры гомофильности. Например, критерий симметричности не применим для гомофильности, так как сравниваются разные сущности объектов - граф и класс объекта. В работе авторы отобрали 4 критерия, которые подходят для меры гомофильности:

- Maximal agreement
- Minimal agreement
- Constant baseline
- Empty class tolerance baseline

**Maximal agreement**

Это ограничение требуется, чтобы ввести максимальное значение метрики для графа, у которого все ребра соединяют только вершины своего класса (максимальная мера гомофильности).

**Minimal Agreement**

По аналогии с maximal agreement - для графа, у которого все ребра соединяют вершины разных классов (минимальная мера гомофиьльности). Максимальное и минимальное ограничение могут показаться несущественными критериями, но важно, чтобы метрика вела себя предсказуемо, что сделает ее интерпретируемой.

**Constant baseline**

Это требование как раз подсвечивает недостатки гомофилии ребер и гомофилии вершин. Метрика не должна зависеть от количества и размера класса. Ожидается, что хорошая мера не должна изменяться при добавлении похожих классов, а также при увеличении выборки внутри классов.

**Empty class tolerance baseline**

Уточнение предыдущего ограничения: хорошая мера должна позволить сравнить различные графы, но в графах может быть различное количество классов. Стандартное решение - добавить пустой класс, чтобы уравнять количество, при этом метрика должна быть нечувствительна к добавлению пустых классов.

## 1.4 Метрики гомофильности по аналогии с метриками классификации

Основываясь на работе “Good Classification Measures and How to Find Them”, авторы статьи заявляют о сопоставимости edge-wise метрик гомофилии и метрик качества классификации. Данное сравнение позволяет им развить метрику гомофилии по аналогии с классификационными метриками. Авторы приводят следующие рассуждения:

При условии, что у ненаправленного графа каждое ребро $u, v$ дает две упорядоченные пары узлов $(u, v)$ и $(v, u)$, мы можем допустить, что для всех таких пар $(u, v)$ существуют лэйбл правдивых значений $y_u$ и лэйбл предсказаний $y_v$. При этом авторы замечают, что метрики классификации будут отражать соотношения этих двух лэйблов, а $edge - wise$ метрики в свою очередь демонстрируют как часто узлы, связанные ребром, принадлежат одному и тому же лэйблу. Из этого наблюдения авторы статьи выводят матрицу ошибок C с элементами $c_{ij}$, обозначающих количество ребер так, что $y_u = i$ и $y_v = j$, при этом отмечают, что такая матрица будет симметричной, так как граф не является направленным.

Кроме того, отталкиваясь от заданной матрицы ошибок, авторы вводят новое определение монотонности метрики гомофилии, позволяющее сравнивать графы с различным количеством ребер и классов. Согласно этому определению, монотонной метрику можно считать при условии, что выполнена empty class tolerant, метрика возрастает при увеличении диагонального элемента на 2 и снижается при росте $_{i,j}$ и $c_{j,i}$ на 1 для $i! = j$.

Тем не менее, вернемся к рассмотрению эволюции метрики по аналогии с задачами классификации. Авторы, описывая усложнение показателей, приводят пары парамаетров точность - гомофилия ребер, сбалансированная точность - сбалансированная гомофилия. Переход от одной пары к другой подвержен одной логике -подобно тому, как точность может демонстрировать некорректные для принятия решений результаты на несбалансированной по классам выборке, гомофилия ребер ($h_{edge}$) также приводит к ошибочным выводам на таких выборках. С целью нивелировать данный эффект вводится мера сбалансированной гомофилии:

$$h_{bal} = \frac{1}{C} \sum_{k=1}^{C} \frac{|(u,v) : y_u = y_v = k|}{D_k}$$

Однако для такой метрики не выполняются критерии монотонности maximal, minimal agreement и asymptotic constant baseline. Для решения большей части проблемы можно получить скорректированную сбалансированную меру гомофилии:

$$h_{bal}^{abj} := \frac{C}{C-1}\left(h_{bal} - \frac{1}{C}\right)$$

Для данной метрики будут выполнены критерии maximal agreement и constant baseline, но она по-прежнему будет немонотонной, а также не будет удовлетворять критерию emtpy class tolerance.

Любопытно заметить, что скорректированная гомофилия ребер (без учета разбалансировки классов) в свою очередь удовлетворяет условиямmaximal agreement, asymptotic constant baseline и empty class tolerance, частично удовлетворяет условиям монотонности, доказательство чего авторы приводят в приложении.

Рассмотрев, достоинства и ограничения различных метрик и их сходство с метриками классификации авторы приходят к 2 значимым выводам. Во-первых, они рекомендуют в качестве ключевой метрики гомофилии использовать скорректированную меру гомофилии вне зависимости от количества классов и их баланса. Во-вторых, исследователи замечают, что метрики гомофилии могут напрямую относиться к метрикам выявления групп (community detection evaluation), сопоставляю гомофилию с метрикой модульности (modularity).

# 2 Описание методов, предлагаемых авторами статьи, постановка задачи, новые термины

## 2.1 Получение метрики adjusted homophily

Как уже было показано ранее, две классические метрики (гомофилия ребер и гомофилия вершин) не удовлетворяют свойству constant baseline, что делает их непригодными для оценки степени гомо-/гетерофилии на разных датасетах.

Проиллюстрировать это можно на простом примере графа, где каждая вершина соединена с одной другой вершиной из каждого класса. Для графов такого вида получаем:

$$h_{edge} = \frac{|u, v \in E : y_u = y_v|}{|E|} = \frac{|V|}{C \cdot |V|} = \frac{1}{C}$$

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

Значения обоих метрик зависят только от количества классов C, однако вне зависимости от количества классов графы будут являться в одинаковой степени негомофильными. Аналогичный вывод был приведен в работе [5].

Чтобы удовлетворить свойству constant baseline авторы работы модифицировали классический вариант гомофилии ребер, нормировав его вычитанием из $h_{edge}$ его мат. ожидания.

Положим, что дан случайный граф, в котором каждая вершина $v$ из $n$ вершин соединена с другими $d(v)$ вершин ребром вне зависимости от их принадлежности классу. Тогда для такого графа вероятность того, что произвольное ребро имеет на одном конце вершину класса $k$ будет равна $\frac{\sum_{v:y_v=k} d(v)}{2|E|}$, и формула имеет вид:

$$\frac{|u, v \in E : y_u = y_v|}{|E|} - \sum_{k=1}^{c} \frac{\left(\sum_{v:y_v=k}\right)^2 d(v)}{(2|E|)^2} = h_{edge} - \sum_{k=1}^{c} \frac{D_k^2}{4|E|^2}$$

.

Чтобы значение было ограничено сверху 1 в случае абсолютно гомофильного графа, как и значение исходной метрики, нормируем на максимально возможное значение полученной выше величины, когда $h_{edge} = 1$:

$$h_{adj} = \frac{h_{edge} - \sum_{k=1}^{c} \frac{D_k^2}{(2|E|)^2}}{1 - \sum_{k=1}^{c} \frac{D_k^2}{(2|E|)^2}} = \frac{h_{edge} - \sum_{k=1}^{c} \bar{p}(k)^2}{1 - \sum_{k=1}^{c} \bar{p}(k)^2}$$

Полученная метрика:

- Удовлетворяет maximal agreement из последнего пункта построения;
- Удовлетворяет empty class tolerance, так как количество классов $C$ не участвует ни в числителе метрики, ни в знаменателе;
- Удовлетворяет constant baseline, доказательство чего авторы приводят в Appendix 3 работы [6];
- Не удовлетворяет minimal agreement, так как в случае абсолютной гетерофильности (при $h_{edge} = 0$) числитель $-\sum_{k=1}^{c} \frac{D_k^2}{(2|E|)^2}$ никак не ограничен и может принимать разные значения для разных графов.

Полученная метрика далее используется в работе как adjusted homophily ($h_{adj}$).

## 2.2 Получение метрики label informativeness

Однако, не полная гомофильность или её отсутствие являются главными блокерами для применения GNN к графу. Ранее разработанные подходы уже способны справляться с гетерофильными графами (MixHop [7], Geom-GCN [8], H2GCN [9], CPGNN [10], GPR-GNN [11])

Например, среди гетерофильных графов есть те, в которых ребрами соединены лишь определенные классы, а между другими никогда не бывает ребер.

Описанная ранее метрика adjusted homophily способна лишь определить, что указанные выше графы не являются гомофильными, но не способна определить тип гетерофильности.

Определим конкретную характеристику, измеряющую информативность метки соседа для метки узла. Например, для графа 1 (см. рисунок 2.1) метка соседа однозначно определяет метку узла. Таким образом, задача классификации узлов на этом наборе данных тривиальна, и мы хотим, чтобы наша информативность была максимальной для таких графов.

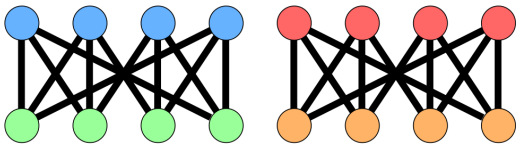

Рис. 2.1. Граф 1

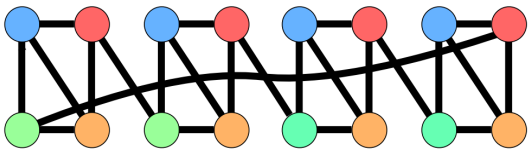

Рис. 2.2. Граф 2

Возьмем произвольное ребро $(\xi, \eta) \in E$. Тогда классы вершин, соединяемых этим ребром, $y_\xi$ и $y_\eta$ соответственно. Задача метрики состоит в том, чтобы показать, сколько информации о классе $y_\eta$ дает знание $y_\xi$. Энтропия $H(y_\xi)$ измеряет «трудность» предсказания метки $\xi$ без знания $y_\eta$. При заданном $y_\eta$ это значение сводится к условной энтропии $H(y_\xi|y_\eta)$. Другими словами, $y_\eta$ раскрывает $I(y_\xi,y_{eta}) = H(y_\xi) - H(y_\xi|y_\eta)$ информации о метке. Из этого можно посчитать label informativeness для отдельного класса:

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

Полученная метрика LI принадлежит интервалу [0, 1], при $LI = 1$ знание класса $y_\xi$ позволяет однозначно узнать класс $y_\eta$, а при $LI = 0$ метки классов не зависят друг от друга. При этом, метрика в зависимости от выбранного распределения может принимать различный вид. Например, если ребра выбираются из равномерного распределения, мы получаем формулу:

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

где

$$p(c_1, c_2) = \sum_{u,v \in E} \frac{[y_u = c_1, y_v = c_2]}{2|E|}$$

Полученная метрика:

- Удовлетворяет maximal agreement, так как принимает значение 1 в случае, когда знание класс одного из конец ребра позволяет однозначно определить класс второго;
- Удовлетворяет constant baseline, доказательство чего авторы приводят в Aappendix C работы [6].

## 2.3 Генерация синтетических данных

Для того, чтобы контролировать характеристики создаваемого графа предлагается использовать стохастическую блочную модель (SBM) [12]. Вершины в этой модели разделены на C классов. Любую пару вершин $i$,$j$ можно соединить ребром с вероятностью $p_{c(i)c(j)}$ независимо от других вершин. Здесь $c(i)$ соответствует классу вершины $i$ или искомой метке вершины.

Возьмем количество классов равное $C = 4$, размеры классов одинаковые $l = n/4$, тогда вероятность соединения вершин $p_{i,j}$ можно записать следующим образом:

$$p_{i,j} = \begin{cases} p_0 K, & \text{если } i = j, \\ p_1 K, & \text{если } i + j = 5, \\ p_2 K, & \text{в остальных случаях,} \end{cases}$$

где $p_0 + p_1 + 2p_2 = 1$ и $K$ положительное число. При этом среднее ожидаемая степень любой вершины будет равна $p_0 Kl + p_1 Kl + 2p_2 Kl = Kl$.

Данная модель позволяет получить граф с различными характеристиками. Коэффициент $p_0$ позволяет контролировать степень гомофилии в графе, а соотношение между $p_1$ и $p_2$ позволяет влиять на LI графа.

Манипуляцию характеристиками графа можно объяснить следующим образом: есть соотношение $i + j = 5$, по нему ребрами соединяются пары вершин с классами (1, 4) и (2, 3). Тогда при $p_2 = 0$ и $p_1 > 0$, зная метки классов соседей можно точно предсказать метку текущей вершины. Если же есть отношение $p_1 = p_2$, то сложно дать какую-либо точную информацию о метке текущей точки. Во время манипуляций с $p_1$ и $p_2$ на $p_0$ не накладывались дополнительные ограничения, хотя именно эта величина характеризует уровень гомофилии в создаваемом графе.

Для того, чтобы определить границы характеристик, которые могут быть получены при помощи данной модели необходимо вычислить асимптотические значения.

При $n \to \infty$:

$$h_{adj} = \frac{4}{3} p_0 - \frac{1}{3}$$

$$LI = 1 - \frac{H(p_0,p_1,p_2,p_2)}{\log 4}$$

где $H(X) = -\sum_i x_i \log(x_i)$

Получается, $h_adj$ может принимать значения от $-1/3$ до $1$, $LI$ может быть от $0$ до $1$. Если $LI = 0$, тогда всегда $h_{adj} = 0$; если $h_{adj} = 1$ то и $LI = 1$. Но если $LI = 1$, то $h_{adj} = 1$ или $h_{adj} = -1/3$.

# 3 Описание данных для экспериментов, результаты применения на данных, потенциал и области прикладного применения методов

В данном разделе исследуется взаимосвязь между характеристиками графа и качеством GNN. В предыдущих исследованиях было показано, что GNN способны показывать хорошее качество на не гомофильных наборах данных. Предполагается, что GNN могут выделять из графов более сложные отношения, чем просто гомофилию. При этом ожидается, что GNN будет работать до тех пор, пока окружение вершины содержит некоторую информацию о ней. Предлагается использовать LI для измерения информативности окружения вершин графа.

Для проверки данной гипотезы были собраны различные наборы реальных данных, а также получен способ для генерации графов на основе стохастической блочной модели с предсказуемыми коэффициентами гомофилии и LI.

Для генерации графов был составлено 208 различных комбинаций $p_0$,$p_1$,$p_2$ для получения различных характеристик LI и гомофилии. Для каждой комбинации генерировалось 10 графов, в каждом из которых по 1000 вершин со степенью 10. Признаки вершин брались из набора данных cora [13], [14], [15] и [16]. Полученные графы делились на обучающую/валидационную/тестовую выборки в соотношении 50%/25%/25%.

В качестве GNN моделей использовались GCN [17] и GraphSAGE [18].

## 3.1 Результаты работы с синтетическим наборами данных на основе стохастической блочной модели

На графиках (см. рисунки 3.1 и 3.2) представлено качество обученных GNN сетей. Каждая точка соответствует сгенерированному набору данных, при этом по оси X откладывается метрика $h_a dj$, по оси Y LI, цвет отражает Accuracy полученной модели.

Разберем график с изображением качества модели GraphSAGE (см. рисунок 3.1). По нему видно, что качество модели больше коррелирует с LI, чем с гомофилией. Действительно, коэффициент корреляции Спирмена между accuracy и LI равен 0.93, а между accuracy и гомофилией 0.05. Если показатель LI высокий, то accuracy

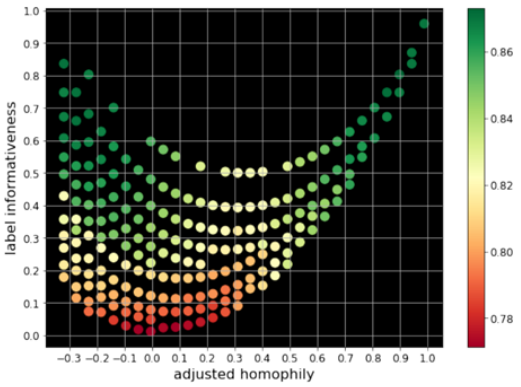


Рис. 3.1. Accuracy GraphSAGE модели на синтетических графах

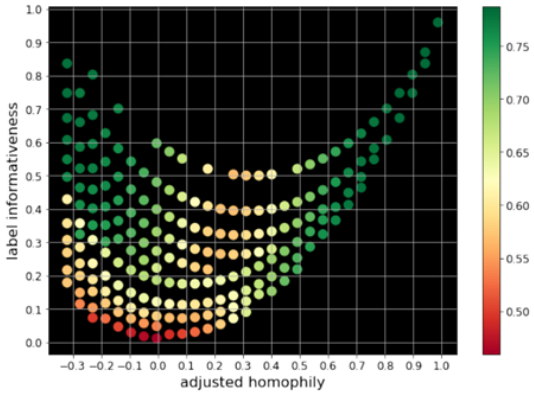


Рис. 3.2. Accuracy GCN модели на синтетических графах

GraphSAGE остается также высоким несмотря на отрицательный коэффициент гомофилии. Для модели GCN наблюдается аналогичная ситуация (см. рисунок 3.2).

## 3.2 Полусинтетические данные из [1]

В [1] авторы выявили, что GNN показывает хорошие результаты на некоторых гетерофильных наборах данных. Были получены полусинтетические графы при помощи добавления межклассовых ребер согласно паттернам из реальных графов. Это позволило получить набор данных с графами с разными уровнями гомофилии.

Для выявления причины выявленного изменения качества была обучена GraphSAGE модель на полусинтетических данных (как в [1]) основанных на cora графе. Было выявлено, что модель получает хорошее качество, когда граф имеет высокий уровень LI. Зависимость между LI, $h_{adj}$ и accuracy показана на рисунке 4. Коэффициент корреляции Спирмена между accuracy и LI равен 0.78, а между accuracy и гомофилией -0.24.

Аналогичный опыт был проведен для графа citeseer, график результатами приведен на рисунке 3.4. Коэффициент корреляции Спирмена между accuracy и LI равен 0.51, а между accuracy и гомофилией -0.54.

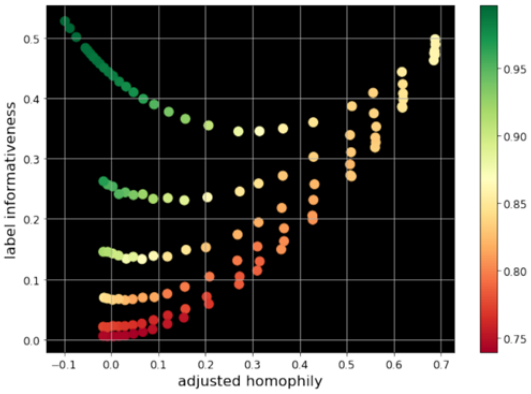


Рис. 3.3. Accuracy GraphSAGE модели на полусинтетическом cora наборе данных [1]

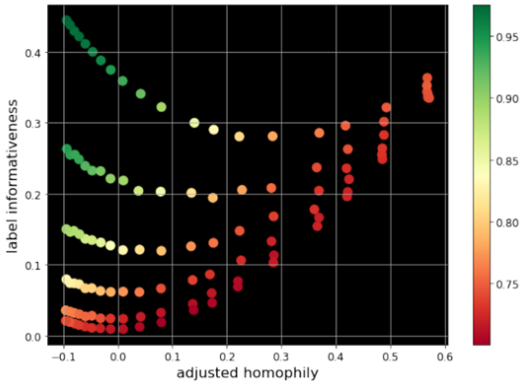


Рис. 3.4. Accuracy GraphSAGE модели на полусинтетическом citeseer наборе данных [1]

## 3.3 Синтетические данные из [2]

В статье [2] авторы также приходят к выводу, что GNN показывает хорошее качество на некоторых гетерофильных графах. Авторы исследовали зависимость качества GNN сетей на различных уровнях гомофилии. Было выявлено, что график этой зависимости имеет U форму, т.е. наблюдается высокое качество GNN модели как при высоких значениях гомофилии, так и при низких.

Были сгенерированы данные аналогично [2]. Для каждого полученного графа был рассчитан уровень LI (см. рисунок 3.5). График зависимости LI от edge homophily также имеет U форму. Отсюда можно сделать вывод, что GNN модели из [2] хорошо работают при высоких значениях LI.

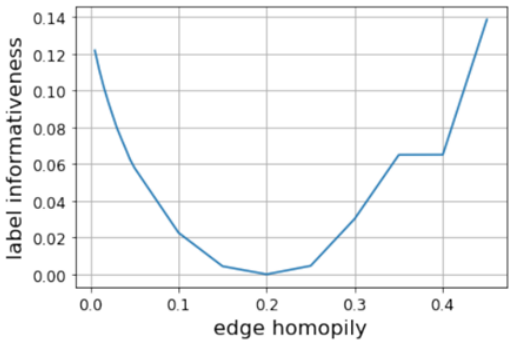


Рис. 3.5. LI в зависимости от $h_{edge}$ на синтетическом графе [2]

# ЗАКЛЮЧЕНИЕ

В данной статье авторы рассмотрели проблему метрик качества для не-гомофильных графов. Проблема заключается в том, что существующие меры гомофилии - гемофилия ребер, вершин чувствительны к балансу классов и их количеству.

Поэтому были предложены критерии для подходящей меры гомофильности, такие как максимальное и минимальное ограничение метрики гомофилии (maximal / minimal agreement), стабильность к масштабу (constant baseline) при добавлении новых классов и расширении выборки, а также возможность сравнения метрики для различных графов (empty class tolerance baseline).

Была предложена скорректированная мера гомофильности. При этом строго доказано, что эта мера удовлетворяет требуемым критериям. Но скорректированная мера гомофильности не способна определить тип гомофильности. Она определяет насколько гомофильным является граф. Но GNN модели могут определять некоторые типы гетерофильных графов, поэтому авторы хотели определить метрику, которая покажет не только степень гетерофильности, но и тип гетерофильности, чтобы понять, с какими типами графов не справляются GNN модели.

Так авторы вывели метрику информативности узла (label informativeness), а также проверили ее на синтетических и частично реальных данных. Эксперименты показали, что метрика информативности лучше объясняет результат GNN моделей для гетерофильных графов, чем метрика гомофилии.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Is homophily a necessity for graph neural networks? / Yao Ma [et al.] // arXiv preprint arXiv:2106.06134. — 2021.

2. Is Heterophily A Real Nightmare For Graph Neural Networks To Do Node Classification? / Sitao Luan [et al.] // arXiv preprint arXiv:2109.05641. — 2021.

3. Good Classification Measures and How to Find Them / Martijn Gösgens [et al.] // Advances in Neural Information Processing Systems. — 2021. — Vol. 34. — P. 17136–17147.

4. Gösgens Martijn M, Tikhonov Alexey, Prokhorenkova Liudmila. Systematic analysis of cluster similarity indices: How to validate validation measures // International Conference on Machine Learning / PMLR. — 2021. — P. 3799–3808.

5. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods / Derek Lim [et al.] // Advances in Neural Information Processing Systems. — 2021. — Vol. 34. — P. 20887–20902.

6. Characterizing Graph Datasets for Node Classification: Beyond Homophily-Heterophily Dichotomy / Oleg Platonov [et al.] // arXiv preprint arXiv:2209.06177. — 2022.

7. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing / Sami Abu-El-Haija [et al.] // international conference on machine learning / PMLR. — 2019. — P. 21–29.

8. Geom-gcn: Geometric graph convolutional networks / Hongbin Pei [et al.] // arXiv preprint arXiv:2002.05287. — 2020.

9. Beyond homophily in graph neural networks: Current limitations and effective designs / Jiong Zhu [et al.] // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 7793–7804.

10. Graph neural networks with heterophily / Jiong Zhu [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. — Vol. 35. — 2021. — P. 11168–11176.

11. Adaptive universal generalized pagerank graph neural network / Eli Chien [et al.] // arXiv preprint arXiv:2006.07988. — 2020.

12. Holland Paul W, Laskey Kathryn Blackmond, Leinhardt Samuel. Stochastic blockmodels: First steps // Social networks. — 1983. — Vol. 5, no. 2. — P. 109–137.

13. Automating the construction of internet portals with machine learning / Andrew Kachites McCallum [et al.] // Information Retrieval. — 2000. — Vol. 3, no. 2. — P. 127–163.

14. Query-driven active surveying for collective classification / Galileo Namata [et al.] // 10th International Workshop on Mining and Learning with Graphs. — Vol. 8. — 2012. — P. 1.

15. Collective classification in network data / Prithviraj Sen [et al.] // AI magazine. — 2008. — Vol. 29, no. 3. — P. 93–93.

16. Yang Zhilin, Cohen William, Salakhudinov Ruslan. Revisiting semi-supervised learning with graph embeddings // International conference on machine learning / PMLR. — 2016. — P. 40–48.

17. Kipf Thomas N, Welling Max. Semi-supervised classification with graph convolutional networks // arXiv preprint arXiv:1609.02907. — 2016.

18. Hamilton Will, Ying Zhitao, Leskovec Jure. Inductive representation learning on large graphs // Advances in neural information processing systems. — 2017. — Vol. 30.