DOI: 10.15514/ISPRAS-2026-38(5)-14

Оценка устойчивости и неопределённости графовых моделей при структурных сдвигах распределения данных

Г.В. Баженов, ORCID: 0009-0009-7967-9573 <gvbazhenov@hse.ru>

НИУ Высшая школа экономики,
101978, Россия, г. Москва, ул. Мясницкая, д. 20.

Аннотация. В надёжных системах принятия решений на основе машинного обучения модели должны быть устойчивы к сдвигам распределения или обеспечивать меры неопределённости своих предсказаний. В задачах предсказания на уровне вершин в графовых данных сдвиги распределения особенно сложны ввиду взаимозависимости объектов. Большинство существующих графовых наборов данных разделяют вершины на обучающую и тестовую части случайным образом, не учитывая структурные свойства графа, тогда как именно они существенны для задач обучения на графах. В данной работе предлагается универсальный подход к созданию разнообразных структурных сдвигов распределения, основанный на свойствах вершин: *популярности*, *локальности* и *плотности*. Эмпирическое сравнение показывает, что предложенные сдвиги весьма сложны для существующих графовых методов: сдвиг на основе локальности является наиболее трудным с точки зрения устойчивости предсказания, а сдвиг на основе плотности – наиболее сложным для OOD-детекции. Мы обнаруживаем, что простые методы нередко превосходят более сложные подходы на рассмотренных структурных сдвигах, при этом существует определенный компромисс между качеством предсказания при сдвиге и способностью к OOD-детекции.

Ключевые слова: графовые нейронные сети; сдвиги распределения; OOD-устойчивость; оценка неопределённости; классификация вершин; структурные разбиения данных.

Для цитирования: Баженов Г.В. Оценка устойчивости и неопределённости графовых моделей при структурных сдвигах распределения данных. Труды ИСП РАН, 2026, том 38, вып. 5, с. 235-256. DOI: 10.15514/ISPRAS-2026-38(5)-14

Evaluating Robustness and Uncertainty of Graph Models under Structural Distributional Shifts

G.V. Bazhenov, ORCID: 0009-0009-7967-9573 <gvbazhenov@hse.ru>

National Research University, Higher School of Economics,
20, Myasnitskaya st., Moscow, 101978, Russia.

Abstract. In reliable decision-making systems based on machine learning, models have to be robust to distributional shifts or provide the uncertainty of their predictions. In node-level problems of graph learning, distributional shifts can be especially complex since the samples are interdependent. To evaluate the performance of graph models, it is important to test them on diverse and meaningful distributional shifts. However, most graph benchmarks considering distributional shifts for node-level problems focus mainly on node features, while structural properties are also essential for graph problems. In this work, we propose a general approach for inducing diverse distributional shifts based on graph structure. We use this approach to create data splits according to several structural node properties: *popularity*, *locality*, and *density*. In our experiments, we thoroughly evaluate the proposed distributional shifts and show that they can be quite challenging for existing graph models. We also reveal that simple models often outperform more sophisticated methods on the considered structural shifts. Finally, our experiments provide evidence that there is a trade-off between the quality of learned representations for the base classification task under structural distributional shift and the ability to separate the nodes from different distributions using these representations.

Keywords: graph neural networks; distributional shifts; OOD robustness; uncertainty estimation; node classification; structural data splits.

For citation: Bazhenov G.V. Evaluating Robustness and Uncertainty of Graph Models under Structural Distributional Shifts. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 5, pp. 235-256. DOI: 10.15514/ISPRAS-2026-38(5)-14

1. Введение

В последнее время значительные усилия направляются на создание систем принятия решений на основе машинного обучения для приложений с высоким риском, таких как финансовые операции, медицинская диагностика и автономное вождение. От таких систем требуется ряд важных свойств, позволяющих пользователям доверять их предсказаниям. Одним из ключевых свойств является *устойчивость* (robustness) – способность базовой модели справляться со *сдвигами распределения* (distributional shifts), когда тестовые данные отличаются от тех, что наблюдались на этапе обучения. Если же модель не способна сохранить высокое качество на сдвинутых данных, она должна сигнализировать о проблемах, предоставляя меру своей *неопределённости* (uncertainty). Выполнение этих требований особенно затруднительно в задачах предсказания на вершинах графа, поскольку в графовых данных объекты взаимозависимы, и сдвиги распределения могут быть устроены сложнее, чем в классической постановке с независимыми и одинаково распределёнными наблюдениями.

Оценка качества графовых моделей в задачах на уровне вершин требует их тестирования в условиях разнообразных, сложных и содержательных сдвигов распределения. Однако в большинстве существующих наборов данных вершины делятся на обучающую и тестовую выборки случайным образом. Среди редких исключений – набор Open Graph Benchmark (OGB) [1], где предложены более сложные неслучайные разбиения, использующие предметно-специфичные признаки вершин. Например, в наборе данных ogbn-arxiv статьи разделены по дате публикации, что соответствует реалистичному сценарию. В то же время немногие наборы данных содержат метаинформацию, пригодную для такого разбиения. Отчасти для решения этой проблемы [2] предложили Graph OOD Benchmark (GOOD) – набор задач для оценки графовых моделей в условиях сдвигов распределения. Его авторы выделяют два типа сдвигов: *концептуальный* и *ковариатный*. Однако на практике оба типа сдвигов, как

правило, присутствуют одновременно, а их искусственное создание затруднительно. Помимо этого, стратегии разбиения в GOOD основаны преимущественно на признаках вершин и не учитывают структуру графа.

Мы восполняем этот пробел, предлагая универсальный метод формирования *структурных* сдвигов распределения в графовых данных. Наш подход позволяет создавать разнообразные, сложные и содержательные сдвиги на уровне вершин, применимые к любому набору данных без дополнительной метаинформации. Для этого мы вводим стратегии разбиения, ориентированные на такие свойства вершин, как *популярность*, *локальность* и *плотность*. Предложенный подход гибок: он позволяет как вводить новые структурные сдвиги, так и варьировать долю вершин в обучающей и тестовой выборках. Эмпирически мы показываем, что предложенные сдвиги распределения весьма сложны для существующих графовых методов. В частности, сдвиг на основе локальности оказывается наиболее трудным для *OOD-устойчивости* (out-of-distribution robustness), тогда как сдвиг на основе плотности – для *OOD-детекции* (out-of-distribution detection). Наши эксперименты также показывают, что простые модели нередко превосходят более сложные подходы на структурных сдвигах. Более того, мы исследуем ряд модификаций архитектур графовых моделей, направленных на повышение их OOD-устойчивости и качества OOD-детекции. Результаты указывают на существование определённого компромисса: модификации, повышающие качество представлений для основной задачи классификации при структурном сдвиге, одновременно ухудшают способность модели разделять вершины из разных распределений.

2. Вклад

1. Мы предлагаем универсальный метод создания структурных сдвигов распределения для задач классификации вершин, применимый к любому графовому набору данных без дополнительной метаинформации.
2. Мы вводим три стратегии разбиения данных, основанные на структурных свойствах вершин графа – популярности (на основе характеристики PageRank), локальности (на основе характеристики Personalized PageRank) и плотности (на основе локального коэффициента кластеризации), – что обеспечивает разнообразные и содержательные условия для оценки моделей.
3. Мы проводим всестороннее эмпирическое исследование и показываем, что сдвиги на основе локальности наиболее трудны для OOD-устойчивости (среднее падение точности достигает 14%, в худшем случае – более 30%), а сдвиги на основе плотности – для OOD-детекции (среднее значение метрики AUROC (Area Under the Receiver Operating Characteristic Curve) близко к случайному угадыванию). Также мы обнаруживаем, что простые методы нередко превосходят более сложные, причем архитектурные изменения, повышающие OOD-устойчивость, приводят к снижению качества OOD-детекции, что указывает на компромисс между этими задачами.

3. Предпосылки

3.1 Задачи на графах при сдвигах распределения

В машинном обучении на графах (МО на графах) существует несколько направлений, посвящённых решению задач предсказания на уровне вершин при сдвиге распределения. Их принципиальное различие состоит в том, какую именно проблему они призваны решить. Одно из таких направлений – *состязательная устойчивость* (adversarial robustness). Здесь задача состоит в построении метода, способного противостоять искусственным сдвигам распределения, создаваемым как состязательные атаки на графовые нейронные сети (graph neural network, GNN) в виде искажённых входных данных. Соответствующие подходы часто основаны на различных аугментациях данных, вносящих случайный или обучаемый шум [3-7].

Другое направление – *обобщение на данные вне распределения* (out-of-distribution generalization). Основная задача здесь – разработать метод, способный работать с реальными сдвигами распределения в графах и сохранять высокое качество предсказаний в различных OOD-средах. Для повышения устойчивости графовых моделей к таким сдвигам был предложен ряд методов инвариантного обучения и минимизации риска [8-11].

Наконец, третье направление – *оценка неопределённости* (uncertainty estimation) – охватывает несколько задач. В задаче *обнаружения ошибок* (error detection) модель должна выдавать оценки неопределённости, согласованные с ошибками предсказания, то есть присваивать более высокие значения потенциально неверным предсказаниям. В условиях сдвига распределения оценки неопределённости также могут использоваться для *OOD-детекции* (OOD detection), когда от модели требуется отличить «сдвинутые» OOD-данные от данных из исходного распределения (in-distribution, ID) [12-16].

Предлагаемые нами структурные сдвиги распределения применимы для оценки и обобщаемости на OOD-данные, и OOD-детекции, и обнаружения ошибок, поскольку они воспроизводят свойства реалистичных графовых данных.

3.2 Методы оценки неопределённости

В зависимости от источника неопределённость принято делить на *неопределённость данных*, описывающую присущий данным шум вследствие ошибок разметки или перекрытия классов, и *неопределённость знаний*, обусловленную недостатком информации для точных предсказаний, когда распределение тестовых данных отличается от обучающих [12, 17-18].

Методы общего назначения. Простейшими подходами являются стандартные классификационные модели, предсказывающие параметры распределения softmax. Для таких моделей мерой неопределённости служит энтропия предсказанного категориального распределения. Однако этот подход не позволяет разделить неопределённость данных и неопределённость знаний, поэтому обычно уступает методам, которые мы рассмотрим далее.

Методы ансамблирования – это мощные, но ресурсоёмкие подходы, которые обеспечивают хорошее качество как предсказаний, так и оценок неопределённости. Наиболее распространённый пример – подход *Deep Ensemble* [19], где эмпирическое распределение параметров модели получают, обучая несколько её экземпляров с разной случайной инициализацией. Другой способ построения ансамбля, метод *Monte Carlo Dropout* [20], традиционно считается базовым методом в литературе по оценке неопределённости. Однако было показано, что эта техника, как правило, уступает глубоким ансамблям, поскольку её предсказания зависимы и недостаточно разнообразны. Важно отметить, что ансамблевые методы позволяют естественным образом разложить общую неопределённость на неопределённость данных и неопределённость знаний [18].

Помимо этого, существует семейство *методов на основе распределения Дирихле*. Их ключевая идея состоит в моделировании поточечного распределения Дирихле через предсказание его параметров для каждого входного примера с помощью отдельной модели. Чтобы получить параметры категориального распределения, параметры распределения Дирихле нормируют на их сумму. Эта нормировочная константа называется *свидетельством* (evidence) и может служить мерой общей уверенности модели. Подобно ансамблевым методам, подходы на основе распределения Дирихле способны разделять неопределённость данных и неопределённость знаний. Существует множество таких методов. Один из первых – *Prior Network* [12], который моделирует распределение Дирихле с помощью *контрастного обучения* (constrastive learning) на OOD-примерах. Несмотря на теоретическую обоснованность, этот метод требует доступа к OOD-данным на этапе обучения, что является существенным ограничением. Другой подход, *Posterior Network* [14], контролирует поведение распределения Дирихле с помощью *нормализующих потоков* (normalizing flows)

[21-22]. Потоки оценивают плотность распределения в скрытом пространстве и снижают *свидетельство Дирихле* в областях низкой плотности, не требуя OOD-примеров.

Методы, специализированные для графов. Недавно был разработан ряд методов оценки неопределённости, специально предназначенных для задач на уровне вершин. Так, *Graph Posterior Network* [23] расширяет известный подход *Posterior Network*. Для моделирования распределения Дирихле он сначала кодирует признаки вершин в латентные представления с помощью инвариантной к графу модели, а затем использует по одному нормализующему потоку на класс для оценки плотности. После этого к параметрам Дирихле применяется схема *персонализированного распространения* (personalized propagation) [24] для учёта сетевых эффектов. Ещё один метод на основе Дирихле – *Graph-Kernel Dirichlet Estimation* [25]. В отличие от *Posterior Network*, его главная особенность – составная целевая функция, которая одновременно оптимизирует качество классификации и моделирует распределение Дирихле. Первое достигается *дистилляцией знаний* (knowledge distillation) из GNN-учителя, а второе – регуляризацией относительно априорного распределения Дирихле, вычисляемого через оценку графового ядра. Однако, как показано в работе [23], этот подход уступает *Graph Posterior Network*, требуя при этом более сложной процедуры обучения и больших вычислительных ресурсов.

3.3 Методы повышения устойчивости

Повысить OOD-устойчивость можно с разных точек зрения, но общая идея состоит в обучении представлений, инвариантных к нежелательным изменениям во входных данных. В *доменной адаптации* (domain adaptation) ключевой тезис заключается в том, что для эффективного переноса между доменами и устойчивости к сдвигам предсказания должны опираться на признаки, общие для исходного и целевого доменов. В свою очередь, в *инвариантном обучении* (invariant learning) методы разделяют входное пространство на различные среды и нацелены на построение представлений, устойчивых к переходу между этими средами.

Методы общего назначения. Классический подход в доменной адаптации – *Domain-Adversarial Neural Network* [26]. Он способствует формированию признаков, которые являются дискриминативными для основной задачи в исходном домене, но не позволяют отличить один домен от другого. Это достигается совместной оптимизацией двух классификаторов, работающих поверх общих *скрытых представлений* (latent representations): основной решает задачу классификации, а доменный классификатор обучается различать исходный и целевой домены. Другой простой метод из класса *ненаблюдаемой доменной адаптации* (unsupervised domain adaptation) – *Deep Correlation Alignment* [27]. Он согласовывает статистики второго порядка (ковариации) в исходном и целевом распределениях, вычисленные по активациям слоёв нейронной сети.

Типичный метод инвариантного обучения – *Invariant Risk Minimization* [8]. Он ищет такие представления данных, для которых оптимальный классификатор является одним и тем же во всех средах и повсюду обеспечивает приемлемое качество. Ещё один метод, *Risk Extrapolation* [10], нацелен на устойчивость к ковариатным сдвигам и инвариантность предсказаний, уделяя особое внимание тем сдвигам, которые сильнее всего влияют на качество в обучающих средах.

Методы, специализированные для графов. Недавно был предложен и ряд методов повышения OOD-устойчивости, специализированных для графов. Один из них – метод инвариантного обучения *Explore-to-Extrapolate Risk Minimization* [11]. Он использует несколько «агентов-исследователей», которые представлены редакторами структуры графа и состязательно обучаются для максимизации дисперсии рисков между различными созданными средами. Существует также простой метод аугментации данных *Mixup* [28], который обучает нейронные сети на выпуклых комбинациях пар примеров и их меток.

Разработать аналог этого метода для задач на уровне вершин непросто из-за взаимосвязей между объектами. Основываясь на этой идее, [29] предложили адаптацию *Mixup* для графовых данных. Вместо обучения на комбинациях исходных признаков вершин их метод использует промежуточные представления, полученные графовыми свёртками для самих вершин и их соседей.

4. Структурные сдвиги и распределения

4.1 Общий подход

Как отмечалось ранее, существующие наборы данных для оценки устойчивости и неопределённости в задачах на уровне вершин ориентированы преимущественно на сдвигах в пространстве признаков. Мы предлагаем универсальный подход, который порождает нетривиальные, но содержательные структурные сдвиги распределения. Для этого мы вводим некоторую скалярную характеристику вершины σ_i и вычисляем её для каждой вершины $i \in \mathcal{V}$. Затем все вершины сортируются по возрастанию этой характеристики: вершины с наименьшими значениями σ_i считаются внутри распределения (*in-distribution*, ID), а остальные – за пределами распределения (*out-of-distribution*, OOD). В результате получается сдвиг распределения, основанный на структуре графа, при котором ID- и OOD-вершины обладают различными структурными свойствами. Конкретный тип сдвига определяется выбором σ_i; несколько вариантов этой характеристики мы рассматриваем в разд. 4.2.

Далее ID-вершины равномерно случайным образом делятся на следующие части:

- Обучающая часть разбиения (далее Train) содержит вершины $\mathcal{V}_{\text{train}}$, используемые для обычного обучения моделей; только они участвуют в вычислении градиентов.
- Валидационная ID часть (далее Valid-In) позволяет отслеживать наилучшую модель на этапе обучения путём вычисления функции потерь на валидации для вершин $\mathcal{V}_{\text{valid-in}}$ и выбора лучшего контрольного состояния.
- Тестовая ID часть (далее Test-In) используется для тестирования на оставшихся ID-вершинах $\mathcal{V}_{\text{test-in}}$ и представляет наиболее простую постановку, требующую от модели воспроизведения зависимостей в пределах распределения.

Оставшиеся OOD-вершины делятся на два подмножества в порядке возрастания σ_i:

- Тестовая OOD часть (далее Test-Out) используется для оценки устойчивости моделей к сдвигам распределения. Оно состоит из вершин с наибольшими значениями σ_i и таким образом представляет наиболее сдвинутую часть $\mathcal{V}_{\text{test-out}}$.
- Валидационная OOD часть (далее Valid-Out) содержит OOD-вершины с меньшими значениями σ_i и потому сдвинуто менее, чем Test-Out. Это подмножество $\mathcal{V}_{\text{valid-out}}$ может использоваться для мониторинга качества модели на сдвинутом распределении. В наших экспериментах предполагается более сложная постановка, при которой такие сдвинутые данные недоступны в процессе обучения. Вместе с тем наличие этого подмножества позволяет дополнительно отдалить $\mathcal{V}_{\text{test-out}}$ от $\mathcal{V}_{\text{train}}$: чем больше $\mathcal{V}_{\text{valid-out}}$, тем более значительный сдвиг распределения создаётся между обучающими и OOD-тестовыми вершинами.

Предложенный подход достаточно гибок и позволяет легко варьировать как размер обучающей выборки, так и тип структурного сдвига. Далее мы рассмотрим конкретные сдвиги, используемые в нашей работе.

4.2 Предлагаемые сдвиги распределения

Выбор свойства вершины σ_i определяет итоговое разбиение данных. Мы рассматриваем разнообразные характеристики графа, отвечающие различным типам структурных сдвигов,

которые могут возникать на практике. В стандартной постановке машинного обучения (без графов) сдвиги, как правило, затрагивают только пространство признаков (или, в общем случае, совместное распределение признаков и целевой переменной). Однако в задачах на графах возможны и сдвиги, непосредственно связанные со структурой графа.

Сдвиг на основе популярности. Первая стратегия воспроизводит возможное смещение в сторону популярности. В ряде приложений естественно ожидать, что обучающая выборка состоит из более популярных объектов. Например, в веб-поиске важность страниц в интернет-графе можно измерить с помощью PageRank [30]. Для этого приложения разметку страниц логично начинать с важных, поскольку они посещаются чаще. Аналогичные ситуации возникают в социальных сетях, где естественно начинать разметку с наиболее влиятельных пользователей, или в сетях цитирований, где в первую очередь размечаются наиболее цитируемые статьи. Однако при применении графовой модели принципиально важно делать точные предсказания и для менее популярных объектов. Исходя из этого, мы вводим разбиение на основе популярности, опираясь на PageRank. Вектор значений PageRank π_i описывает стационарное распределение случайного блуждания с перезапусками, для которого матрица переходов определена как нормализованная матрица смежности $\mathbf{AD}^{-1}$, а вероятность перезапуска равна α:

$$\boldsymbol{\pi} = (1-\alpha)\mathbf{AD}^{-1}\boldsymbol{\pi} + \alpha\boldsymbol{p}. \tag{1}$$

Вектор вероятностей перезапуска $\mathbf{p}$ называется вектором персонализации; по умолчанию $p_i = 1/n$, что соответствует равномерному распределению по вершинам.

Для построения разбиения на основе популярности мы используем известную структурную характеристику PageRank (PR) как меру важности вершин. Мы вычисляем PR для каждой вершины i и полагаем $\sigma_i = -\pi_i$, то есть вершины с меньшими значениями PR (менее важные) относятся к OOD-подмножествам. Вместо PR потенциально можно использовать любую другую меру важности вершины, например степень вершины или центральность по посредничеству.

Рис. 1 (слева) показывает, что предложенный сдвиг разделяет наиболее важные вершины, принадлежащие ядрам крупных кластеров, и структурную периферию, состоящую из менее важных вершин по значению PR. Мы также наблюдаем, что такое разбиение на основе популярности хорошо согласуется с некоторыми естественными сдвигами распределения. Распределение PageRank в обучающей и тестовой частях набора данных ogbn-arxiv [1] подтверждает это: статьи разделены по дате публикации, а более старые статьи закономерно имеют больше цитирований и, следовательно, более высокое значение характеристики PageRank. Предложенное нами разбиение позволяет воспроизводить такое поведение для наборов данных, где временные метки отсутствуют.

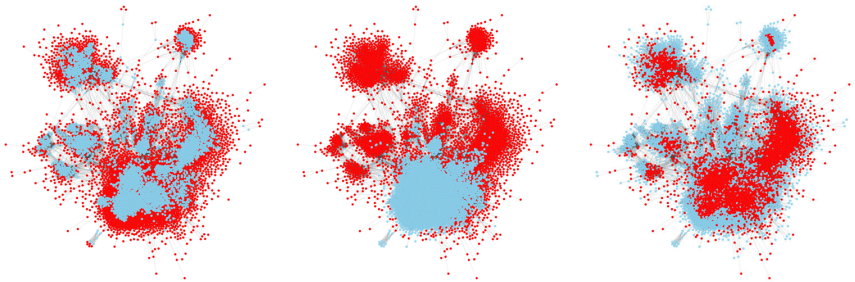

Рис. 1. Визуализация структурных сдвигов для набора данных amazon-photo: ID – синий, OOD – красный.
Fig. 1. Visualization of structural shifts for amazon-photo dataset: ID – blue, OOD – red.

Сдвиг на основе локальности. Следующая стратегия ориентирована на возможное смещение в сторону локальности, которое может возникать при разметке путём обхода графа начиная с некоторой вершины. Например, в задачах веб-поиска краулер должен обходить веб-граф, следуя по ссылкам. Аналогично, информацию о пользователях социальной сети обычно можно получить через прикладной интерфейс, обнаруживая новых пользователей через друзей уже известных. Для моделирования такой ситуации можно было бы разделить вершины по расстоянию кратчайшего пути до заданной вершины. Однако расстояния в графе дискретны, а число вершин на заданном расстоянии может расти экспоненциально с расстоянием. Поэтому такой подход не даёт желаемой гибкости при варьировании размера обучающей выборки. Вместо этого мы используем концепцию персонализированного страничного ранга PageRank (PPR) [30] для определения локальной окрестности вершины. PPR – это стационарное распределение случайного блуждания, всегда перезапускающегося из некоторой фиксированной вершины j. Таким образом, вектор персонализации $\mathbf{p}$ в формуле (1) полагается равным единичному вектору вершины j. Соответствующий сдвиг распределения естественным образом фиксирует локальность, поскольку случайное блуждание всегда перезапускается из одной и той же вершины.

В наших экспериментах в качестве вершины перезапуска мы выбираем вершину j с наибольшим значением PR. Затем вычисляем значения PPR π_i для каждой вершины i и определяем меру $\sigma_i = -\pi_i$. Вершины с высоким PPR, входящие в ID-подмножество, оказываются близкими к вершине перезапуска, тогда как далёкие вершины попадают в OOD-подмножество. Рис. 1 (по центру) подтверждает, что локальность действительно сохраняется: ID-часть состоит из одной компактной области вокруг вершины перезапуска. В OOD-подмножество, таким образом, попадают как периферийные вершины, так и вершины, отмеченные как важные при разбиении по популярности, но расположенные далеко от вершины перезапуска. Разбиение на основе PPR существенно влияет на распределение попарных расстояний внутри подмножеств ID и OOD: смещение ID-части в сторону локальности делает OOD-вершины более удалёнными друг от друга.

Хотя сдвиги распределения на основе локальности вполне естественны, нам не известны публично доступные наборы данных, специально на них ориентированные. Мы полагаем, что наш подход будет полезен для оценки устойчивости GNN к подобным сдвигам. Эмпирические результаты в разд. 6 показывают, что разбиения на основе локальности наиболее сложны для графовых моделей и потому требуют особого внимания.

Сдвиг на основе плотности. Следующий предлагаемый сдвиг распределения основан на *плотности*. Одной из наиболее простых характеристик вершины, описывающих локальную плотность в графе, является локальный коэффициент кластеризации. Для вершины i пусть d_i – её степень, а γ_i – число рёбер, соединяющих соседей i. Тогда локальный коэффициент кластеризации определяется как плотность рёбер в одношаговой окрестности:

$$c_i = \frac{2\gamma_i}{d_i(d_i-1)}.$$

В наших экспериментах вершины с наибольшим коэффициентом кластеризации считаются принадлежащими к ID, что означает $\sigma_i = -c_i$.

Это структурное свойство представляет особый интерес для формирования сдвигов распределения, поскольку определяется через число треугольников – подструктур, которые большинство стандартных GNN не способны различать и подсчитывать [31]. Поэтому интересно изучить, как изменение коэффициента кластеризации влияет на качество предсказания и оценку неопределённости.

Рис. 1 (справа) визуализирует разбиение на основе плотности. Видно, что OOD-часть включает как высокостепенные *центральные* вершины, так и *периферийные* вершины степени один. Действительно, вершины степени один естественным образом имеют нулевой коэффициент кластеризации. С другой стороны, для высокостепенных вершин число рёбер

между их соседями обычно растёт медленнее, чем квадратично. Поэтому такие вершины, как правило, имеют стремящийся к нулю коэффициент кластеризации.

Наконец, отметим, что в существующих наборах данных с реалистичными разбиениями локальный коэффициент кластеризации нередко сдвинут между обучающей и тестовой частями. В зависимости от набора данных обучающее подмножество может быть смещено в сторону вершин с более высоким или более низким коэффициентом кластеризации. В наших экспериментах мы сосредотачиваемся на первом сценарии.

На протяжении всей данной работы мы используем однобуквенные обозначения **P** (популярность), **L** (локальность) и **D** (плотность) для трех структурных сдвигов распределения и используем их в подписях к рисункам и заголовках таблиц для краткости.

5. Условия эксперимента

Наборы данных. Наш подход применим к любому набору данных для предсказания на вершинах, но в экспериментах мы используем следующие семь гомофильных наборов данных, широко применяемых в литературе: три сети цитирований – cora-ml, citeseer [32], [33-35] и pubmed [36], – два графа соавторства – coauthor-physics и coauthor-cs [37], – и два набора данных о совместных покупках [37-38] – amazon-photo и amazon-computers. Кроме того, мы рассматриваем ogbn-products – крупномасштабный набор данных из тестового набора задач OGB. Некоторые из рассматриваемых методов не способны обрабатывать столь большой набор данных, поэтому мы проводим только анализ структурных сдвигов на нём и не используем его для сравнения методов.

Характеристики используемых в данной работе наборов данных приведены в табл. 1.

Табл. 1. Описание рассматриваемых графовых наборов данных для задачи классификации вершин.

Table 1. Description of the considered graph datasets for node classification task.

Набор данных	# Вершин	# Рёбер	# Классов	# Признаков
amazon-computers	13 381	259 159	10	767
amazon-photo	7 484	126 530	8	745
coauthor-cs	18 333	163 788	15	6 805
coauthor-physics	34 493	495 924	5	8 415
cora-ml	2 995	16 316	7	2 879
citeseer	3 327	4 732	6	3 703
pubmed	19 717	44 338	3	8 415
ogbn-products	2 449 029	61 859 140	47	100

Для любого сдвига распределения каждый набор данных разбивается следующим образом. Половина вершин с наименьшими значениями σ_i относится к ID-подмножеству и делится на Train, Valid-In и Test-In равномерно случайным образом в пропорции 30% : 10% : 10%. Вторая половина содержит оставшиеся OOD-вершины и делится на Valid-Out и Test-Out в порядке возрастания σ_i в пропорции 10% : 40%. Таким образом, в базовой постановке соотношение ID и OOD составляет 50% : 50%. Также мы провели эксперименты с другими соотношениями разбиений, предполагающими меньший размер OOD-подмножеств; подробности см. в Разделе 6.5.

Модели. В наших экспериментах мы применяем предложенный набор задач для оценки OOD-устойчивости и неопределённости различных графовых моделей. Для повышения качества обобщения на OOD-данные в задаче классификации вершин мы рассматриваем следующие методы:

- **ERM** – метод *минимизации эмпирического риска* (Empirical Risk Minimization), обучающий простую GNN-модель путём оптимизации стандартной функции потерь в задаче классификации;

- **DANN** – подход *Domain Adversarial Neural Network* [26], обучающий регулярный и доменный классификаторы таким образом, чтобы признаки были неразличимы между различными доменами;
- **CORAL** – техника *Deep Correlation Alignment* [27], направленная на повышение сходства представлений вершин из разных доменов;
- **EERM** – метод *Explore-to-Extrapolate Risk Minimization* [11] на основе редакторов структуры графа, создающих виртуальные среды в процессе обучения;
- **Mixup** – реализация *Mixup* из работы [29], простая техника аугментации данных, адаптированная для задач обучения на графах;
- **DE** – подход *Deep Ensemble* [19] из графовых моделей как мощный, но ресурсоёмкий метод повышения качества предсказания.

Для OOD-детекции мы рассматриваем следующие методы оценки неопределённости:

- **SE** – простая GNN-модель, используемая в методе **ERM**, для которой мера неопределённости – *энтропия softmax* (энтропия предсказательного распределения);
- **GPN** – реализация *Graph Posterior Network* [23] – метода оценки неопределённости на уровне вершин;
- **NatPN** – *Natural Posterior Network* [39], в котором энкодер имеет ту же архитектуру, что и в методе **SE**;
- **DE** – *Deep Ensemble* [19] над несколькими GNN-моделями, позволяющий выделить неопределённость знаний для OOD-детекции;
- **GPE** и **NatPE** – *байесовские комбинации* **GPN** и **NatPN** в качестве примеров подхода к построению ансамбля моделей Дирихле [39].

Конфигурации методов. В основной серии экспериментов с графовыми моделями в Разделах 6.1 и 6.2 мы используем графовый энкодер на основе трёх слоёв графовой свёрточной сети (GCN) [40]. Поверх энкодера располагается единственный линейный слой, который служит для предсказания параметров категориального распределения в задачах классификации, параметров распределения Дирихле для моделирования неопределённости и т. д. Скрытая размерность базовой архитектуры равна 256, а коэффициент дропаута между скрытыми слоями составляет $p = 0{,}2$. Мы используем стандартный оптимизатор Adam [41] со скоростью обучения 0,0003 и весовой регуляризацией 0,00001. В дополнительной серии экспериментов с улучшенной архитектурой GNN, обсуждаемой в Разделе 6.3, число слоёв графовой свёртки уменьшается с 3 до 2: первый заменяется шагом предобработки на основе линейного слоя, между слоями графовой свёртки добавляются пропускные соединения, а свёртка GCN заменяется на SAGE [42].

Некоторые из рассматриваемых методов повышения OOD-устойчивости, такие как DANN и CORAL, оперируют понятием *доменов* (областей) и требуют знания о принадлежности каждой обучающей вершины к тому или иному домену. Поскольку наши структурные характеристики вершин вещественнозначны, а не дискретны, определение доменов нетривиально. Для этого мы дискретизируем значения характеристик на $k = 10$ непересекающихся интервалов (доменов) и присваиваем каждой вершине индекс её домена. Эти индексы и используются как метки доменов в DANN и CORAL в процессе обучения.

Для экспериментов с методами DANN, CORAL, EERM и Mixup используется их техническая реализация, представленная в работе [2], тогда как для остальных методов применяются наши собственные реализации.

Задачи предсказания и метрики оценки. Для оценки OOD-устойчивости в задаче классификации вершин мы используем стандартную долю правильных ответов (Accuracy). Для измерения качества оценок неопределённости задача OOD-детекции рассматривается как бинарная классификация с положительными событиями, соответствующими наблюдениям из OOD-подмножества; качество измеряется метрикой AUROC.

6. Эмпирическое исследование

В данном разделе мы показываем, как предложенный подход к созданию сдвигов распределения может использоваться для оценки устойчивости и неопределённости графовых моделей. В частности, мы сравниваем типы введённых сдвигов и обсуждаем, какие из них более сложны. Также рассматривается их влияние на качество предсказания методов повышения OOD-устойчивости и на способность к OOD-детекции методов оценки неопределённости.

6.1 Анализ структурных сдвигов распределения

В данном разделе мы анализируем и сравниваем предложенные структурные сдвиги распределения.

OOD-устойчивость. Для изучения влияния предложенных сдвигов на качество предсказания графовых моделей мы используем наиболее простой метод ERM и приводим падение точности между ID- и OOD-тестовыми подмножествами в таблице 2 (столбцы OOD-устойчивости). Видно, что результаты классификации вершин на рассматриваемых наборах данных стабильно ниже при измерении на OOD-части, причём это падение может достигать десятков процентов в некоторых случаях. Наибольшее снижение качества наблюдается на разбиениях на основе локальности: в среднем оно достигает 14%, а в худшем случае – более 30%. Это согласуется с нашей интуицией о том, что обучение на локальных регионах графа может мешать обобщению на OOD-данные и создаёт серьёзное препятствие для повышения OOD-устойчивости. Хотя сдвиг на основе плотности по всей видимости менее трудный, он всё же сложнее, чем сдвиг на основе популярности, приводя к падению качества в среднем на 5,5% и в худшем случае на 17%.

OOD-детекция. Для анализа способности графовых моделей обнаруживать сдвиги распределения, предоставляя более высокую неопределённость на сдвинутых входных данных. В табл. 2 (столбцы OOD-детекции) приводится качество наиболее простого метода **SE** для каждого предложенного сдвига и набора данных.

Табл. 2. Сравнение структурных сдвигов распределения по OOD-устойчивости (падение качества предсказаний метода ERM, %) и OOD-детекции (SE, AUROC).
Table 2. Comparison of structural distributional shifts in terms of OOD robustness (drop in predictive performance of ERM, %) and OOD detection (SE, AUROC).

Набор данных	**P** ↓Acc	**L** ↓Acc	**D** ↓Acc	**P** AUROC	**L** AUROC	**D** AUROC
amazon-computers	−10,01%	−21,95%	−7,91%	88,52	86,48	44,24
amazon-photo	−8,54%	−30,93%	−3,58%	92,05	93,29	41,08
coauthor-cs	−3,13%	−1,22%	−4,86%	83,25	85,74	50,91
coauthor-physics	−3,42%	−4,75%	−1,41%	86,60	87,73	37,50
cora-ml	−3,94%	−14,61%	−17,09%	75,67	87,13	81,55
citeseer	−0,02%	−26,51%	−8,39%	68,01	89,89	66,90
pubmed	−3,23%	−5,71%	−0,78%	68,60	66,34	58,60
ogbn-products	−2,86%	−2,83%	−0,12%	88,50	88,56	36,00
Среднее	**−4,39%**	**−13,56%**	**−5,52%**	**81,40**	**85,65**	**52,10**

Видно, что сдвиги на основе популярности и локальности эффективно детектируются данным методом, что подтверждается средними значениями метрик. В частности, значения AUROC варьируются от 68 до 92 на разбиениях на основе популярности и примерно в том же диапазоне на разбиениях на основе локальности. Относительно сдвигов на основе плотности видно, что качество OOD-детекции в среднем практически совпадает со случайным предсказанием. Только для сетей цитирований AUROC превышает 58, достигая пика в 81. Это согласуется с предшествующими работами, показывающими, что стандартные

GNN не способны подсчитывать подструктуры типа треугольников. В нашем случае это приводит к тому, что графовые модели не способны обнаруживать изменения плотности, измеренной как число треугольников вокруг центральной вершины.

Таким образом, сдвиг на основе популярности по всей видимости является наиболее простым для OOD-детекции, тогда как сдвиг на основе плотности – наиболее сложным. Это наглядно показывает, как наш подход к созданию разбиений данных позволяет варьировать сложность сдвигов распределения, используя различные структурные свойства в качестве факторов разбиения.

6.2 Сравнение существующих методов

В этом разделе мы сравниваем несколько существующих методов для повышения OOD-устойчивости и OOD-детекции на предложенных структурных сдвигах. Для наглядного отображения общего качества моделей мы сначала ранжируем их по данной метрике на конкретном наборе данных, а затем усредняем результаты по наборам данных.

OOD-устойчивость. Для каждой модели мы измеряем как абсолютные значения точности, так и её падение между ID- и OOD-подмножествами в табл. 3. Видно, что простая техника аугментации данных Mixup нередко показывает наилучшее качество. Лишь на сдвиге на основе плотности дорогостоящий DE в среднем превосходит её при тестировании на OOD-подмножестве. Это подтверждает практическую полезность аугментации данных: она предотвращает переобучение к структурным закономерностям на ID-подмножестве и повышает OOD-устойчивость. Из остальных методов повышения OOD-устойчивости специализированный для графов метод EERM превосходит DANN и CORAL на сдвигах на основе популярности и локальности. Однако методы доменной адаптации превосходят EERM на сдвигах на основе плотности, обеспечивая в среднем лучшее качество как на ID-, так и на OOD-подмножествах.

В итоге мы обнаруживаем, что наиболее сложные методы, генерирующие виртуальные среды и предсказывающие принадлежность вершин к областям для повышения OOD-устойчивости, нередко уступают более простым методам, таким как аугментация данных.

OOD-детекция. При сравнении качества оценок неопределённости в табл. 4 можно наблюдать, что методы на основе энтропии предсказательного распределения, как правило, превосходят методы на основе Дирихле. В частности, естественное выделение неопределённости знаний в DE позволяет ему производить оценки неопределённости, наиболее согласующиеся с OOD-входными данными в среднем, особенно при применении к сдвигам на основе локальности и плотности. В целом GPN и его ансамблевый вариант GPE обеспечивают более высокое качество OOD-детекции, чем аналоги на основе NatPN, при тестировании на разбиениях на основе популярности и локальности.

Табл. 3. Сравнение нескольких графовых методов для повышения OOD-устойчивости. Представлены ранги методов, усреднённые по различным наборам данных (меньше – лучше).
Table 3. Comparison of several graph methods for improving the OOD robustness. Ranks averaged across different datasets are reported for methods (lower is better).

Метод	**P** ID	**P** OOD	**L** ID	**L** OOD	**D** ID	**D** OOD
ERM	4,0	4,0	3,3	4,1	3,9	4,0
Mixup	1,4	2,1	1,4	2,4	1,9	3,3
EERM	3,6	3,9	4,4	3,3	5,0	4,3
DANN	4,3	4,3	5,0	4,1	3,0	3,6
CORAL	4,7	4,1	4,3	4,7	4,1	3,9
DE	3,0	2,6	2,6	2,3	3,1	2,0

Табл. 4. Сравнение нескольких графовых методов для OOD-детекции посредством оценки неопределённости. Ранги методов, усреднённые по различным наборам данных (меньше – лучше).
Table 4. Comparison of several graph methods for improving the OOD detection. Ranks averaged across different datasets are reported for methods (lower is better).

Метод	**P**	**L**	**D**
SE	1,4	2,1	4,0
GPN	3,3	3,9	4,3
NatPN	5,3	4,1	2,9
DE	2,1	1,1	2,7
GPE	3,1	4,3	3,4
NatPE	5,7	5,4	3,7

6.3 Влияние улучшений архитектуры GNN

В данном разделе мы рассматриваем ряд модификаций базовой архитектуры GNN для таких методов, как ERM (используемый для оценки OOD-устойчивости) и SE (обеспечивающий оценки неопределённости для OOD-детекции). В частности, число слоёв графовой свёртки уменьшается с 3 до 2: первый заменяется шагом предобработки на основе многослойного перцептрона (MLP), между слоями графовой свёртки добавляются *пропускные соединения* (residual connections), а свёртка GCN [40] заменяется на SAGE [42].

Эти изменения направлены на ослабление ограничений на обмен информацией между центральной вершиной и её соседями и обеспечение большей независимости в обработке представлений вершин в нейронных слоях. Ожидается, что такая модификация поможет GNN-модели обучать структурные паттерны, более успешно переносимые на сдвинутое OOD-подмножество. Далее мы исследуем, как эти изменения в архитектуре модели влияют на качество предсказания и оценок неопределённости соответствующих методов при тестировании на предложенных структурных сдвигах распределения.

Для этого мы используем счётчики выигрыш/ничья/проигрыш, отражающие, сколько раз модифицированная архитектура превзошла, дала статистически незначимое отличие или проиграла соответствующему методу соответственно. Как видно из табл. 5, метод ERM с предложенными модификациями, как правило, превосходит базовую архитектуру как на ID, так и на OOD-подмножествах, что подтверждается высокими счётчиками выигрыш. Однако как видно из табл. 6, та же модификация в соответствующем методе SE приводит к устойчивому снижению качества, что отражается в высоких счётчиках проигрыш.

Табл. 5. Сравнение предложенных модификаций архитектуры в методе ERM для OOD-устойчивости. Счётчики выигрыш/ничья/проигрыш по наборам данных для модифицированной GNN против базовой.
Table 5. Comparison of the proposed architecture modifications that are used in the ERM method for OOD robustness. Win/tie/loss counts are reported across graph datasets.

Метод	**P** ID	**P** OOD	**L** ID	**L** OOD	**D** ID	**D** OOD
ERM + mod	4/2/1	5/2/0	4/2/1	3/2/2	4/2/1	5/2/0

Табл. 6. Сравнение предложенных модификаций архитектуры в методе SE для OOD-детекции. Счётчики выигрыш/ничья/проигрыш по наборам данных.
Table 6. Comparison of the proposed architecture modifications that are used in the ERM method for OOD detection. Win/tie/loss counts across graph datasets across graph datasets.

Метод	**P**	**L**	**D**
SE + mod	0/0/7	0/1/6	4/0/3

Это означает, что при достижении более высокого качества предсказания на сдвинутом подмножестве становится труднее обнаруживать входные данные из этого подмножества как OOD, поскольку они становятся менее различимы для GNN-модели в контексте базовой

задачи классификации вершин. Наше наблюдение весьма схоже с тем, что требуется от методов *инвариантного обучения*, стремящихся производить представления вершин, инвариантные к различным областям или средам. Это может указывать на компромисс между качеством выученных представлений для решения целевой задачи классификации вершин в условиях структурного сдвига распределения и способностью разделять вершины из разных распределений на основе этих представлений.

6.4 Свойства сдвигов распределения

В данном разделе приводится детальный анализ и сравнение предложенных сдвигов распределения. Мы рассматриваем три репрезентативных реальных набора данных – amazon-computers, coauthor-cs и cora-ml – и обсуждаем, как различные сдвиги влияют на базовые свойства данных: *баланс классов*, *распределение степеней* и *расстояния в графе* между вершинами внутри ID- и OOD-подмножеств.

Баланс классов. Баланс классов непосредственно влияет на объём информации, накапливаемой моделью обработки графов и используемой для оценки неопределённости и принятия предсказаний. Это особенно важно для оценки моделей на основе Дирихле, использующих нормализующие потоки: их оценки плотности могут стать нерелевантными при значительном изменении баланса классов.

Из рис. 2 видно, что разбиение на основе популярности не создаёт существенного различия в балансе классов между ID- и OOD-подмножествами (для рассматриваемых наборов данных). Таким образом, более важные и менее важные вершины в среднем имеют одинаковую вероятность принадлежности к тому или иному классу. Более заметные различия создаются при разбиении на основе плотности. Разбиение на основе локальности приводит к наиболее значительным изменениям для некоторых классов. Это показывает, что стратегии разбиения, основанные на структурной локальности в графе, могут быть весьма сложными, поскольку влияют и на такие ключевые статистики, как баланс классов.

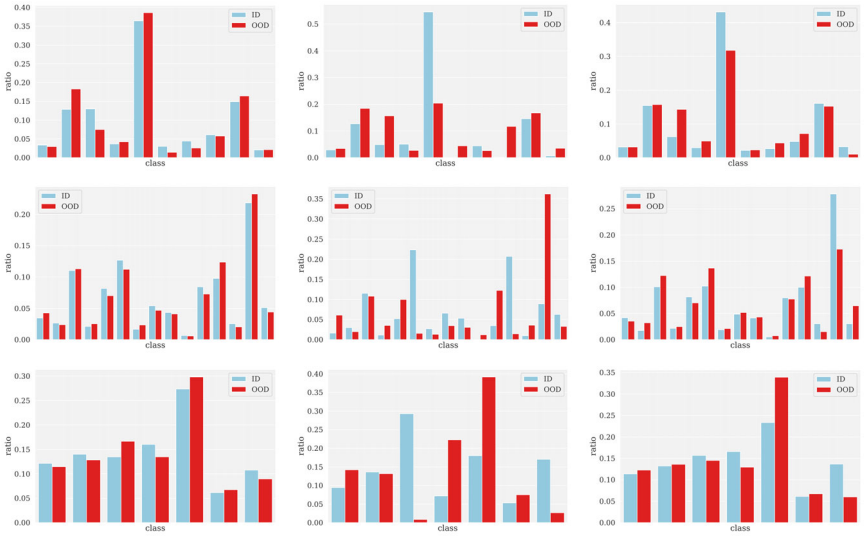

Рис. 2. Баланс классов при предложенных структурных сдвигах для amazon-computers (верхний ряд), coauthor-cs (средний ряд) и cora-ml (нижний ряд).
Fig. 2. Class balance across different types of structural shifts for amazon-computers (top row), coauthor-cs (middle row) and cora-ml (bottom row).

Распределение степеней. Распределение степеней вершин – одна из базовых структурных характеристик графа, описывающих локальную важность вершин. Степени особенно важны для GNN, поскольку определяют число каналов вокруг рассматриваемой вершины, используемых для передачи сообщений и агрегации информации по соседям.

Из рис. 3 видно, что наиболее значительное изменение в распределении степеней происходит при разделении ID- и OOD-подмножеств по характеристике PageRank: ID-часть содержит больше высокостепенных вершин. Это ожидаемо, поскольку PageRank – характеристика важности (центральности) вершины, а степень вершины является наиболее простой мерой центральности, коррелирующей с характеристикой PageRank. При разбиениях на основе локальности разница в распределении степеней меньше, но всё ещё значима: характеристика PPR отбирает вершины по их относительной важности для конкретной вершины, поэтому некоторые высокостепенные вершины могут быть менее важными. Наконец, при разбиениях на основе плотности распределение степеней также изменяется, однако степень сдвига обычно менее значительна, а для некоторых наборов данных (например, coauthor-cs) высокостепенные вершины оказываются в OOD-подмножестве.

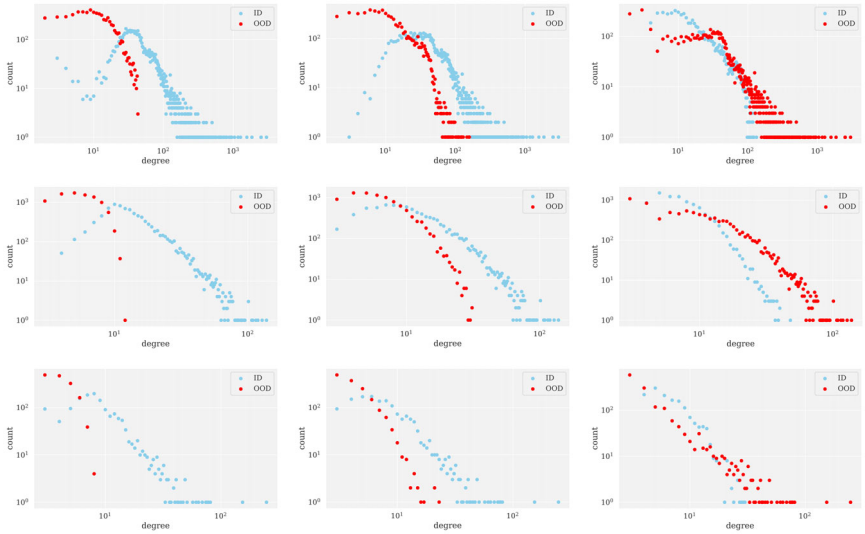

Рис. 3. Распределение степеней при предложенных структурных сдвигах для amazon-computers (верхний ряд), coauthor-cs (средний ряд) и cora-ml (нижний ряд).
Fig. 3. Distribution of node degrees across different types of structural shifts for amazon-computers (top row), coauthor-cs (middle row) и cora-ml (bottom row).

Распределение расстояний в графе. Расстояние между двумя вершинами в графе определяется как длина кратчайшего пути между ними. Здесь мы вычисляем расстояния между вершинами в ID- или OOD-подмножестве в рамках исходного графа, то есть при поиске кратчайшего пути используется полный граф. Распределение расстояний показывает, насколько локально сосредоточены части набора данных.

Из рис. 4 видно, что разбиение на основе локальности приводит к наиболее значительным изменениям в расстояниях: OOD-вершины удалены друг от друга почти вдвое дальше, чем ID-вершины. При этом разбиение на основе популярности не создаёт такой разницы: распределения на ID- и OOD-подмножествах практически совпадают. Это означает, что смещение в сторону популярности в графе не препятствует охвату менее популярных *периферийных* вершин, поскольку наиболее популярные вершины могут быть широко

распределены. Аналогично разбиение на основе плотности не индуцирует значительной разницы в попарных расстояниях.

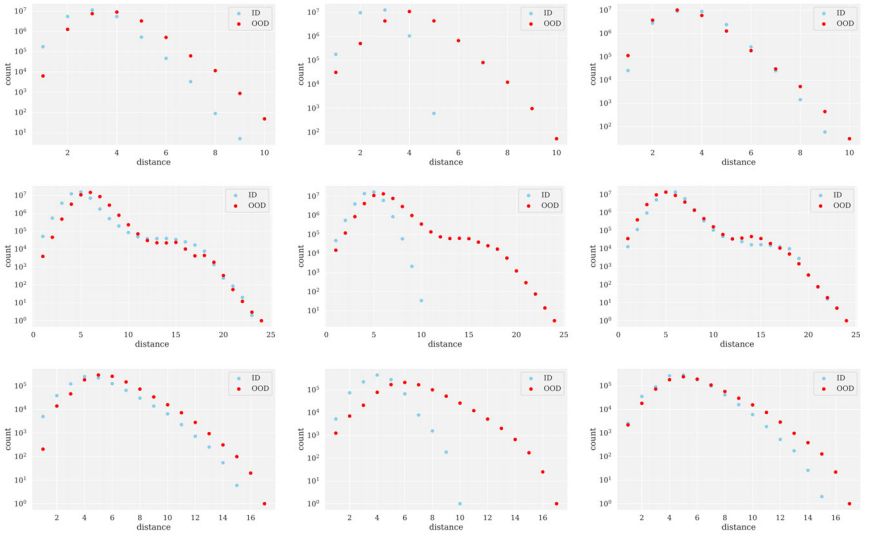

Рис. 4. Распределение расстояний в графе при предложенных структурных сдвигах для amazon-computers (верхний ряд), coauthor-cs (средний ряд) и cora-ml (нижний ряд).
Fig. 4. Distribution of graph distances across different types of structural shifts for amazon-computers (top row), coauthor-cs (middle row) и cora-ml (bottom row).

6.5 Результаты при других соотношениях разбиений

Основная часть нашего эмпирического исследования сосредоточена на соотношении разбиений 50% : 50%. Здесь мы расширяем обсуждение, рассматривая переменные соотношения ID- и OOD-выборок и изучая две постановки, в которых доля OOD-вершин меньше: 70% : 30% и 90% : 10%.

Сначала рассмотрим анализ предложенных структурных сдвигов при соотношении 70% : 30%. Как видно из табл. 7, результаты как для OOD-устойчивости (в задаче классификации вершин), так и для OOD-детекции (через оценку неопределённости) изменились несущественно по сравнению с базовой постановкой: рассматриваемые графовые модели показывают среднее падение качества в 3% на сдвиге по популярности и 6% на сдвиге по плотности, тогда как качество OOD-детекции остаётся на уровне 81 и 54 пунктов AUROC соответственно. Несмотря на то что теперь модели имеют доступ к более разнообразным структурам с точки зрения популярности и локальной плотности, делать верные предсказания для неважных и слабо связанных вершин по-прежнему так же сложно.

В то же время падение качества предсказания на сдвиге по локальности оказывается менее значительным, чем в исходной постановке, и достигает лишь 5% в среднем вместо прежних 14%. При этом качество OOD-детекции на этом структурном сдвиге снижается до 79 пунктов (в исходной постановке оно составляло 85 пунктов). Эти результаты согласуются с нашей интуицией: по мере уменьшения расстояния между ID- и OOD-вершинами OOD-примеры становятся менее отличимы от ID-примеров. Это приводит к снижению качества OOD-детекции в среднем, тогда как разрыв между ID- и OOD-метриками в задаче классификации вершин сокращается.

Табл. 7. Сравнение структурных сдвигов по OOD-устойчивости (ERM, ↓Acc) и OOD-детекции (SE, AUROC). Соотношение ID- и OOD-подмножеств составляет 70%:30%.
Table 7. Comparison of structural distributional shifts in terms of OOD robustness (ERM, ↓Acc) and OOD detection (SE, AUROC). The ID to OOD ratio is 70%:30%.

Набор данных	**P** ↓Acc	**L** ↓Acc	**D** ↓Acc	**P** AUROC	**L** AUROC	**D** AUROC
amazon-computers	−3,80%	−12,66%	−14,04%	87,83	85,77	50,46
amazon-photo	−7,19%	−3,03%	−6,79%	91,80	85,78	47,27
citeseer	+3,24%	+2,50%	−4,13%	70,56	64,55	58,04
coauthor-cs	−6,51%	−1,89%	−6,64%	83,68	81,13	63,24
coauthor-physics	−2,58%	−5,50%	−2,47%	86,70	82,06	39,68
cora-ml	−7,56%	−13,63%	−9,49%	82,06	84,63	76,82
pubmed	+0,46%	−0,67%	−1,76%	66,64	63,34	55,79
ogbn-products	−2,35%	−2,36%	−4,02%	82,35	82,44	43,50
Среднее	**−3,28%**	**−4,65%**	**−6,17%**	**81,45**	**78,71**	**54,35**

При сравнении существующих графовых моделей по результатам в табл. 8 видно, что их ранжирование остаётся практически таким же, как и в исходной постановке. В частности, простейшая техника аугментации данных Mixup нередко показывает наилучшую OOD-устойчивость, а DE обеспечивает вторые по качеству результаты на большинстве структурных сдвигов. Что касается OOD-детекции (табл. 9), методы на основе энтропии предсказательного распределения снова превосходят методы на основе распределения Дирихле, а оценки неопределённости DE лучше всего коррелируют с OOD-примерами. Таким образом, наши выводы о качестве графовых моделей согласуются с ранее сделанными в работе.

Табл. 8. Сравнение методов для повышения OOD-устойчивости. Ранги, усреднённые по различным наборам данных (меньше – лучше). Соотношение ID- и OOD-подмножеств составляет 70%:30%.
Table 8. Comparison of methods for improving OOD robustness. Ranks averaged across different datasets are reported for methods (lower is better). The ID to OOD ratio is 70%:30%.

Метод	**P** ID	**P** OOD	**L** ID	**L** OOD	**D** ID	**D** OOD
ERM	3,0	3,4	2,9	3,1	4,3	4,3
Mixup	1,9	2,6	1,4	2,7	1,4	3,3
EERM	4,9	3,9	5,0	4,0	4,4	4,1
DANN	4,1	4,4	4,3	4,3	3,7	2,6
CORAL	4,1	4,0	4,4	4,9	4,1	3,7
DE	3,0	2,7	3,0	2,0	3,0	3,0

Таблица 9. Сравнение методов для OOD-детекции. Ранги, усреднённые по различным наборам данных (меньше – лучше). Соотношение ID- и OOD-подмножеств составляет 70%:30%.
Table 9. Comparison of methods for improving OOD detection. Ranks averaged across different datasets are reported for methods (lower is better). The ID to OOD ratio is 70%:30%.

Метод	**P**	**L**	**D**
SE	1,3	2,3	3,7
GPN	3,3	3,9	3,9
NatPN	5,4	4,3	3,7
DE	3,0	1,4	1,9
GPE	3,1	3,9	3,4
NatPE	4,9	5,3	4,4

Теперь рассмотрим постановку с более экстремальным соотношением 90%:10%. Согласно табл. 10, падение качества предсказаний при сдвигах по популярности и плотности становится более выраженным: более 6% (вместо 3%) для популярности и почти 11% (вместо 6%) для плотности. Это ожидаемый результат: графовые модели теперь тестируются на вершинах с самыми экстремальными структурными свойствами (то есть с наименьшими значениями PageRank или коэффициента кластеризации), и их неточные предсказания уже не компенсируются более точными на вершинах с менее аномальными свойствами.

При этом падение качества на сдвиге по локальности составляет в среднем лишь около 1%. Этот результат также закономерен: графовые модели теперь имеют доступ ко всему многообразию подструктур и признаков вершин, что позволяет им одинаково точно делать предсказания для любого региона графа, независимо от расстояния до конкретной вершины. Наблюдаемые эффекты важны для учёта при использовании нашего подхода для оценки графовых моделей.

Табл. 10. Сравнение структурных сдвигов по OOD-устойчивости (ERM, ↓Acc) и OOD-детекции (SE, AUROC). Соотношение ID- и OOD-подмножеств составляет 90%:10%.
Table 10. Comparison of structural distributional shifts in terms of OOD robustness (ERM, ↓Acc) and OOD detection (SE, AUROC). The ID to OOD ratio is 90%:10%.

Набор данных	**P** ↓Acc	**L** ↓Acc	**D** ↓Acc	**P** AUROC	**L** AUROC	**D** AUROC
amazon-computers	−8,45%	−7,79%	−30,98%	83,42	76,53	68,97
amazon-photo	−8,00%	+0,12%	−16,35%	89,01	82,57	64,41
citeseer	−5,14%	+1,35%	+1,86%	75,90	65,11	55,48
coauthor-cs	−9,79%	−4,42%	−7,76%	83,22	76,02	70,63
coauthor-physics	−3,77%	−4,91%	−5,32%	84,68	78,60	52,97
cora-ml	−16,10%	+5,52%	−8,03%	79,36	69,64	69,78
pubmed	+3,33%	+4,70%	−2,63%	61,05	57,14	55,11
ogbn-products	−2,92%	−2,75%	−15,61%	77,92	78,02	68,33
Среднее	**−6,36%**	**−1,02%**	**−10,60%**	**79,32**	**72,95**	**63,21**

Что касается качества OOD-детекции, результаты на сдвиге по локальности продолжают снижаться и достигают 73 пункта AUROC по сравнению с 79 пунктов в предыдущей постановке. В то же время метрики детекции на остальных структурных сдвигах либо остаются почти на том же уровне (79 пунктов вместо 81 на сдвиге по популярности), либо даже улучшаются в среднем (63 пункта вместо 54 на сдвиге по плотности), что согласуется с описанными выше изменениями в качестве предсказаний. Ранжирование различных моделей по OOD-устойчивости (табл. 11) и OOD-детекции (табл. 12) остаётся практически неизменным, причём простейшие методы показывают наилучшее качество на большинстве задач.

6.6 Сравнение с набором задач GOOD

Наша работа дополняет и расширяет набор задач GOOD, предложенный в работе [2]. Однако между нашими подходами существует ряд важных отличий, которые мы обсуждаем далее.

Табл. 11. Сравнение методов для повышения OOD-устойчивости. Ранги, усреднённые по различным наборам данных (меньше – лучше). Соотношение ID- и OOD-подмножеств составляет 90%:10%.
Table 11. Comparison of methods for improving OOD robustness. Ranks averaged across different datasets are reported for methods (lower is better). The ID to OOD ratio is 90%:10%.

Метод	**P** ID	**P** OOD	**L** ID	**L** OOD	**D** ID	**D** OOD
ERM	3,3	3,7	3,4	3,0	3,3	4,4
Mixup	1,7	2,4	1,7	2,9	1,7	3,0
EERM	5,0	4,1	5,0	3,7	4,9	3,9
DANN	4,3	4,0	4,0	4,3	3,4	3,0
CORAL	3,9	3,6	4,1	5,1	4,3	3,4
DE	2,9	3,1	2,7	2,0	3,4	3,3

Таблица 12. Сравнение методов для OOD-детекции. Ранги, усреднённые по различным наборам данных (меньше – лучше). Соотношение ID- и OOD-подмножеств составляет 90%:10%.
Table 12. Comparison of methods for improving OOD detection. Ranks averaged across different datasets are reported for methods (lower is better). The ID to OOD ratio is 90%:10%.

Метод	**P**	**L**	**D**
SE	1,6	2,6	3,3
GPN	3,1	3,0	3,9
NatPN	5,0	4,4	3,7
DE	2,6	1,7	1,6
GPE	3,6	4,0	4,3
NatPE	5,1	5,3	4,3

Одним из ключевых свойств набора задач GOOD является теоретическое разграничение двух типов сдвигов распределения, представленных через графическую модель. В частности, авторы рассматривают *ковариатные сдвиги*, при которых распределение признаков изменяется, а условное распределение целевой переменной при данных признаках остаётся прежним, и *концептуальные сдвиги*, где происходит обратное: условное распределение целевой переменной изменяется, а распределение признаков – нет. Хотя такое разграничение полезно для понимания свойств конкретных GNN-моделей, на практике чисто ковариатные или концептуальные сдвиги встречаются редко – как правило, оба типа сдвигов присутствуют одновременно.

Для создания «чистых» ковариатных или концептуальных сдвигов [2] вводят различные подмножества переменных, которые либо полностью определяют целевую переменную, либо создают с ней смешивающие зависимости, либо полностью независимы от неё. Это усложняет создание сдвигов распределения на новых наборах данных с помощью такого подхода. Действительно, сдвиги распределения в наборе задач GOOD могут быть корректно реализованы только для синтетических наборов данных, либо путём добавления синтетических признаков, которые описывают различные области как полностью независимые переменные или создают необходимые концепты через ложную корреляцию с целевой переменной. Более того, авторы утверждают, что для реальных наборов данных необходимо проводить отбор признаков вершин, чтобы создать требуемую постановку сдвига. Всё это накладывает многочисленные ограничения на подготовку разбиений данных.

Наш метод, напротив, не разграничивает ковариатные и концептуальные сдвиги и потому универсально применим к любому набору данных без каких-либо модификаций. При этом тип сдвига распределения и размеры всех частей разбиения легко контролируются. Именно эта гибкость является главным преимуществом нашего подхода. Подробные таблицы с результатами по каждому набору данных доступны в дополнительных материалах к работе.

Наконец, авторы [2] подтверждают важность использования как признаков вершин, так и структуры графа. Тем не менее предлагаемые ими сдвиги на уровне вершин основаны

преимущественно на признаках, таких как число слов, год публикации в сети цитирований, язык пользователей в социальной сети или названия организаций в сети веб-страниц. Из графовых свойств для создания сдвигов в некоторых сетях цитирований используются только степени вершин. Мы же, в свою очередь, сосредотачиваемся на структуре графа и предлагаем разнообразные структурные сдвиги вместе с подходом, который позволяет легко создавать и другие разбиения на основе структурных свойств.

7. Заключение

В данной работе мы предлагаем и анализируем структурные сдвиги распределения для оценки устойчивости и неопределённости моделей в задачах предсказания на вершинах графа. Наш подход позволяет создавать реалистичные, сложные и разнообразные сдвиги для произвольного набора графовых данных. В экспериментах мы оцениваем предложенные структурные сдвиги и показываем, что они могут быть весьма сложными для существующих графовых моделей. Мы также обнаруживаем, что простые модели нередко превосходят более сложные на этих трудных сдвигах. Более того, применяя различные модификации архитектур графовых моделей, мы демонстрируем наличие компромисса между качеством представлений, выученных для целевой задачи классификации в условиях структурного сдвига, и способностью обнаруживать этот сдвиг по тем же представлениям.

Реализация. Исходный код, включая генераторы структурных сдвигов и экспериментальную среду, доступен в репозитории GitHub [43]. Помимо этой реализации, предложенные в работе структурные разбиения были интегрированы в две основные библиотеки для глубокого обучения на графах в качестве встроенного инструмента обработки данных: см. реализацию [44] в программном пакете PyTorch Geometric [45] и реализацию [46] в программном пакете Deep Graph Library [47]. Это наиболее простой способ применить предложенные структурные сдвиги к произвольному набору графовых данных.

Ограничения. Хотя предложенные нами методы создания структурных сдвигов мотивированы реальными сдвигами распределения, они генерируются синтетически. Для конкретных приложений естественные сдвиги были бы предпочтительнее, однако наша цель – предложить решение для тех случаев, когда такие сдвиги недоступны. Поэтому мы выбрали подход, универсально применимый к любому набору данных. Важно, что структура графа – единственная общая модальность для различных наборов графовых данных, которую можно единообразно использовать для моделирования разнообразных и сложных сдвигов распределения.

Перспективные направления. Наша работа открывает два ключевых направления для дальнейших исследований. Во-первых, необходимо разрабатывать новые методы для повышения устойчивости и оценки неопределённости на предложенных нами структурных сдвигах. Предложенный подход может содействовать достижению этой цели, предоставляя инструмент для тестирования и оценки таких решений. Во-вторых, следует создавать новые графовые наборы данных, которые точно отражают свойства данных из реальных приложений. Это подразумевает замену синтетических сдвигов на реалистичные. Такой подход позволит выявлять проблемы и сложности, возникающие на практике, и тем самым будет способствовать разработке более эффективных и применимых графовых моделей.

Список литературы / References

[1]. Hu W. et al. Open graph benchmark: Datasets for machine learning on graphs. Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 22118-22133.
[2]. Gui S., Li X., Wang L., Ji S. GOOD: A graph out-of-distribution benchmark. Advances in Neural Information Processing Systems, 2022, vol. 35, pp. 2059-2073.
[3]. Rong Y., Huang W., Xu T., Huang J., DropEdge: Towards deep graph convolutional networks on node classification. In International conference on learning representations, 2020.

[4]. Zügner D., Günnemann S. Adversarial attacks on Graph Neural Networks via meta learning. In International conference on learning representations, 2019.
[5]. Zhang X., Zitnik M. GNNGuard: Defending Graph Neural Networks against adversarial attacks. Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 9263-9275.
[6]. Sun Y., Wang S., Tang X., Hsieh T.-Y., Honavar V. Adversarial attacks on Graph Neural Networks via node injections: A hierarchical reinforcement learning approach. In Proceedings of the web conference 2020, 2020, pp. 673-683.
[7]. Ma J., Ding S., Mei Q. Towards more practical adversarial attacks on Graph Neural Networks. Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 4756-4766.
[8]. Arjovsky M., Bottou L., Gulrajani I., Lopez-Paz D. Invariant risk minimization, 2019. DOI: 10.48550/arXiv.1907.02893.
[9]. Koyama M., Yamaguchi S. When is invariance useful in an out-of-distribution generalization problem? 2020. DOI: 10.48550/arXiv.2008.01883.
[10]. Krueger D. et al. Out-of-Distribution Generalization via Risk Extrapolation (REx). In International conference on machine learning, 2021, pp. 5815-5826.
[11]. Wu Q., Zhang H., Yan J., Wipf D. Handling Distribution Shifts on Graphs: An Invariance Perspective. In International conference on learning representations, 2022.
[12]. Malinin A., Gales M. Predictive uncertainty estimation via prior networks. Advances in Neural Information Processing Systems, 2018, vol. 31.
[13]. Wang B. et al. Deep uncertainty quantification: A machine learning approach for weather forecasting. In Proceedings of the ACM SIGKDD conference on knowledge discovery and data mining, 2019, pp. 2087-2095.
[14]. Charpentier B., Zügner D., Günnemann S. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In Advances in neural information processing systems, 2020, pp. 1356-1367.
[15]. Mukhoti J., Kirsch A., van Amersfoort J., Torr P.H., Gal Y. Deep deterministic uncertainty: A simple baseline. arXiv preprint, 2021. DOI: 10.48550/arXiv.2102.11582.
[16]. Kotelevskii N. et al. Nonparametric Uncertainty Quantification for deterministic neural networks. Advances in Neural Information Processing Systems, 2022, vol. 35.
[17]. Gal Y. Uncertainty in deep learning. University of Cambridge, 2016.
[18]. Malinin A. Uncertainty estimation in deep learning with application to spoken language assessment. PhD thesis, University of Cambridge, 2019.
[19]. Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in neural information processing systems, 2017.
[20]. Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In International conference on machine learning, 2016, pp. 1050-1059.
[21]. Kingma D.P., Salimans T., Jozefowicz R., Chen X., Sutskever I., Welling M. Improved variational inference with inverse autoregressive flow. Advances in Neural Information Processing Systems, 2016, vol. 29.
[22]. Huang C.-W., Krueger D., Lacoste A., Courville A. Neural autoregressive flows. In International conference on machine learning, 2018, pp. 2078-2087.
[23]. Stadler M., Charpentier B., Geisler S., Zügner D., Günnemann S. Graph Posterior Network: Bayesian predictive uncertainty for node classification. Advances in Neural Information Processing Systems, 2021, vol. 34, pp. 18033-18048.
[24]. Klicpera J., Bojchevski A., Günnemann S. Predict then propagate: Graph Neural Networks meet personalized PageRank. In International conference on learning representations, 2019.
[25]. Zhao X., Chen F., Hu S., Cho J.-H. Uncertainty Aware Semi-Supervised Learning on graph data. Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 12827-12836.
[26]. Ganin Y. et al. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 2016, vol. 17, no. 1, pp. 2096-2030.
[27]. Sun B., Saenko K. Deep CORAL: Correlation alignment for deep domain adaptation. In Computer vision – ECCV 2016 workshops, 2016, pp. 443-450.
[28]. Zhang H., Cisse M., Dauphin Y.N., Lopez-Paz D. Mixup: Beyond empirical risk minimization. In International conference on learning representations, 2018.
[29]. Wang Y., Wang W., Liang Y., Cai Y., Hooi B. Mixup for node and graph classification. In Proceedings of the web conference 2021, 2021, pp. 3663-3674.

[30]. Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.
[31]. Chen Z., Chen L., Villar S., Bruna J. Can graph neural networks count substructures? In Advances in neural information processing systems, 2020, pp. 10383-10395.
[32]. McCallum A.K., Nigam K., Rennie J., Seymore K. Automating the construction of internet portals with machine learning. Information Retrieval, 2000, vol. 3, no. 2, pp. 127-163.
[33]. Giles C.L., Bollacker K.D., Lawrence S. CiteSeer: An automatic citation indexing system. In Proceedings of the third ACM conference on digital libraries, 1998, pp. 89-98.
[34]. Getoor L. Link-based classification. In Advanced methods for knowledge discovery from complex data. Springer, 2005, pp. 189-207.
[35]. Sen P., Namata G., Bilgic M., Getoor L., Galligher B., Eliassi-Rad T. Collective classification in network data. AI Magazine, 2008, vol. 29, no. 3, pp. 93-93.
[36]. Namata G., London B., Getoor L., Huang B. Query-driven active surveying for collective classification. In Proceedings of the workshop on mining and learning with graphs (MLG-2012), 2012.
[37]. Shchur O., Mumme M., Bojchevski A., Günnemann S. Pitfalls of graph neural network evaluation, 2018. DOI: 10.48550/arXiv.1811.05868.
[38]. McAuley J., Targett C., Shi Q., Van Den Hengel A. Image-based recommendations on styles and substitutes. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp. 43-52.
[39]. Charpentier B., Borchert O., Zügner D., Geisler S., Günnemann S., Natural posterior network: Deep Bayesian predictive uncertainty for exponential family distributions. In International conference on learning representations, 2022.
[40]. Kipf T.N., Welling M. Semi-supervised classification with Graph Convolutional Networks. In International conference on learning representations, 2017.
[41]. Kingma D.P., Ba J. Adam: A method for stochastic optimization. In International conference on learning representations, 2015.
[42]. Hamilton W., Ying Z., Leskovec J. Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 2017, vol. 30.
[43]. GitHub: structural-graph-shifts. Available at: https://github.com/yandex-research/structural-graph-shifts, accessed 29.08.2026.
[44]. PyG Team. NodePropertySplit. Available at: https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.transforms.NodePropertySplit.html, accessed 29.08.2026.
[45]. Fey M., Lenssen J.E. Fast graph representation learning with PyTorch Geometric. Representation Learning on Graphs and Manifolds Workshop at The Seventh International Conference on Learning Representations, 2019.
[46]. DGL Team, 2018. add_node_property_split. Available at: https://www.dgl.ai/dgl_docs/generated/dgl.data.utils.add_node_property_split.html, accessed 29.08.2026.
[47]. Wang M. et al. Deep Graph Library: A graph-centric, highly-performant package for graph neural networks. Representation Learning on Graphs and Manifolds Workshop at The Seventh International Conference on Learning Representations, 2019.

Информация об авторах / Information about authors

Глеб Владимирович БАЖЕНОВ – аспирант факультета компьютерных наук Национального исследовательского университета «Высшая школа экономики». Область научных интересов: машинное обучение на табличных данных, графах и реляционных базах данных.

Gleb Vladimirovich BAZHENOV – postgraduate student at the Faculty of Computer Science, HSE University. Research interests: tabular machine learning, graph machine learning, relational machine learning.