



Проблемы генерации структурно разнообразных графов

Ф.С. Великонивцев, ORCID: 0000-0002-1830-8340 <fvelikonivtsev@hse.ru>

Национальный Исследовательский Университет Высшая Школа Экономики,
Россия, 109028, г. Москва, ул. Покровский Бульвар, д. 11.

Аннотация. Для многих задач, связанных с графами, принципиально важно располагать набором структурно разнообразных графов. Например, такие графы могут использоваться для тестирования графовых алгоритмов или их нейросетевых аппроксимаций. Однако, насколько нам известно, задача генерации структурно разнообразных графов ранее не рассматривалась в литературе. В настоящей работе мы восполняем этот пробел. Сначала мы обсуждаем, как определить разнообразие набора графов, почему эта задача нетривиальна и как выбрать подходящую меру разнообразия. Затем для заданной меры разнообразия мы предлагаем и сравниваем несколько оптимизирующих её алгоритмов: рассматриваются подходы на основе стандартных моделей случайных графов, локальной оптимизации графов, генетических алгоритмов и нейронных генеративных моделей. Мы показываем, что разнообразие сгенерированных графов можно существенно повысить по сравнению с базовыми генераторами случайных графов. Кроме того, анализ сгенерированных графов позволяет лучше понять свойства функций расстояний между графами: в зависимости от того, какая функция расстояния используется для оптимизации, получаемые графы могут обладать весьма различными структурными свойствами, что даёт более глубокое понимание расстояния между графами, лежащего в основе меры разнообразия.

Ключевые слова: структурно разнообразные графы; разнообразие графов; расстояния между графами; меры разнообразия; модели случайных графов; генетические алгоритмы; генеративные модели графов; оценка графовых алгоритмов.

Для цитирования: Великонивцев Ф.С. Проблемы генерации структурно разнообразных графов. Труды ИСП РАН, 2026, том 38, вып. 5, с. 217-234. DOI: 10.15514/ISPRAS-2026-38(5)-13

Challenges of Generating Structurally Diverse Graphs

F.S. Velikonivtsev, ORCID: 0000-0002-1830-8340 <fvelikonivtsev@hse.ru>

National Research Institute Higher School of Economics,
10, Pokrovsky Bulvar, Moscow, 109028, Russia.

Abstract. For many graph-related problems, it can be essential to have a set of structurally diverse graphs. For instance, such graphs can be used for testing graph algorithms or their neural approximations. However, to the best of our knowledge, the problem of generating structurally diverse graphs has not been explored in the literature. In this paper, we fill this gap. First, we discuss how to define diversity for a set of graphs, why this task is non-trivial, and how one can choose a proper diversity measure. Then, for a given diversity measure, we propose and compare several algorithms optimizing it: we consider approaches based on standard random graph models, local graph optimization, genetic algorithms, and neural generative models. We show that it is possible to significantly improve diversity over basic random graph generators. Additionally, our analysis of generated graphs allows us to better understand the properties of graph distances: depending on which diversity measure is used for optimization, the obtained graphs may possess very different structural properties which gives a better understanding of the graph distance underlying the diversity measure.

Keywords: structurally diverse graphs; graph diversity; graph distances; diversity measures; random graph models; genetic algorithms; graph generative models; evaluation of graph algorithms.

For citation: Velikonivtsev F.S. Challenges of Generating Structurally Diverse Graphs. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 5, pp. 217-234. DOI: 10.15514/ISPRAS-2026-38(5)-13

1. Введение

Многие объекты реального мира естественно представляются в виде графов: например, биологические и химические сущности (атомы, молекулы, белки, метаболические карты), сети взаимодействий (социальные и цитатные сети, финансовые транзакции), карты дорог, распространение эпидемий. Именно поэтому анализ данных, имеющих графовую структуру, является важной и быстро развивающейся областью исследований.

Для генерации реалистичных структур графов в литературе предложено множество моделей случайных графов [1]. Такие модели стремятся воспроизвести свойства, типично наблюдаемые в природных структурах: степенно-закономерное распределение степеней, малый диаметр, структуру сообществ и другие. Каждая модель случайных графов воспроизводит часть этих свойств, вследствие чего генерируемые ею графы неизбежно схожи в определённых аспектах.

С другой стороны, для ряда приложений важна возможность генерировать набор *структурно разнообразных* графов. Например, при автоматической проверке корректности графового алгоритма, оценке качества аппроксимации эвристического алгоритма для графовых задач, тестировании нейронных аппроксимаций графовых алгоритмов или оценке структурного разнообразия и покрытия моделью генерации молекулярных графов, алгоритмы и модели необходимо тестировать на как можно более разнообразных экземплярах графов – иначе результаты могут быть смещены в пользу специфических свойств тестовой выборки [2]. Особенно ясно эта проблема проявляется в нейронном алгоритмическом рассуждении (Neural algorithmic reasoning, NAR) [3] – направлении исследований, изучающем способность нейронных сетей обучаться выполнению классических алгоритмов. Многие алгоритмы в тестах NAR определены на графах; стандартный тест CLRS [4] оценивает все задачи на графах Эрдёша–Рени – структурно узким распределением. Недавние работы по NAR [5, 6] показали, что для улучшения обобщения в данной постановке требуется тщательный учёт структурных свойств оценочных графов, что непосредственно мотивирует использование структурно разнообразных наборов. Наборы разнообразных графов также полезны для

оценки выразительной мощности графовых нейронных сетей. В этом направлении работа [7] предлагает тест GraphWorld из графов с различными статистическими свойствами, однако структуры в нём ограничены степенно-откорректированной стохастической блочной моделью [8] и не покрывают всё многообразие паттернов связности.

Задача построения набора данных, в котором графы максимально разнообразны, не была сформулирована до нашей работы. Настоящая работа заполняет этот пробел. Для этого необходимо прежде всего определить разнообразие набора графов – само по себе нетривиальная задача. Чтобы измерить разнообразие, мы определяем несходство (расстояние) для пары графов, а затем агрегируем попарные расстояния в общую меру разнообразия. Мы показываем, что популярные методы измерения разнообразия имеют существенные недостатки, и предлагаем использовать меру Energy, удовлетворяющую ряду важных желательных свойств. Расстояния между графами, используемые в настоящей работе, описаны в разделе 2.2.

Определив целевую функцию, мы разрабатываем и анализируем следующие стратегии её оптимизации: жадный метод на основе разнообразных генераторов случайных графов, метод локальной оптимизации графов, адаптацию генетического алгоритма к нашей задаче и метод на основе нейронного генеративного моделирования. Для предлагаемых алгоритмов, не основанных на нейронных сетях (Greedy, Genetic, LocalOpt) проведён анализ алгоритмической сложности. Для жадного алгоритма дополнительно доказана гарантия аппроксимации: выход достигает не менее половины разнообразия оптимального подмножества той же размерности из исходной популяции.

Мы эмпирически исследуем предложенные стратегии и показываем, что разнообразие сгенерированных графов можно существенно улучшить по сравнению с базовыми моделями случайных графов. Помимо численного исследования мер разнообразия, мы анализируем распределение структурных характеристик графов и связи между ними. Более того, поскольку мы рассматриваем меры разнообразия на основе нескольких расстояний между графами, наши результаты проливают свет на свойства этих расстояний: в зависимости от оптимизируемой функции структурные свойства сгенерированных графов варьируются, что даёт лучшее понимание того, к каким характеристикам чувствительно то или иное расстояние.

2. Определение разнообразия набора графов

2.1 Постановка задачи

Как уже отмечалось, для различных задач, связанных с графами, принципиально важно иметь репрезентативные структурно разнообразные графы. Интуитивно, такие графы должны «покрывать» (в некотором смысле) пространство всех графов.¹ В данном разделе обсуждается, как определить разнообразие и почему это нетривиально.

Начнём с мотивирующего примера. Простейшим генератором случайных графов является модель Эрдёша–Рени: в ней ребро между любыми двумя вершинами добавляется с вероятностью p независимо от других рёбер. При фиксированном N и $p = 0.5$ каждый простой граф на N вершинах может быть сгенерирован с равной вероятностью, и поэтому на первый взгляд кажется, что данная модель порождает репрезентативные графы.

Однако известно, что с высокой вероятностью графы модели Эрдёша–Рени имеют типичные свойства и потому весьма похожи друг на друга [9]. Например, как показано на рис. 1, для ER-0.5 средняя степень вершины и средний коэффициент кластеризации сконцентрированы вблизи $(N - 1)/2$ и 0.5 соответственно. Варьирование p (ER-mix) позволяет получить все возможные значения этих характеристик по отдельности, но в матожидании они линейно

¹ В настоящей работе термины «разнообразие» и «покрытие» используются взаимозаменяемо.

зависимы. Таким образом, пространство всех возможных комбинаций характеристик не покрывается моделью Эрдёша–Рени.

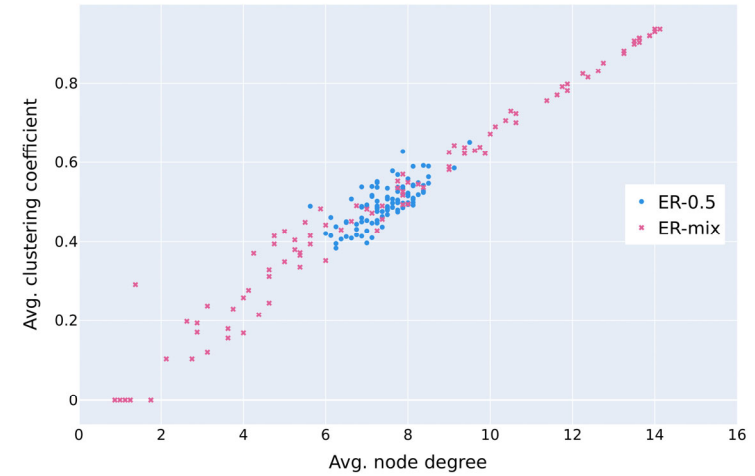


Рис. 1. Средняя степень вершины и средний коэффициент кластеризации в модели Эрдёша–Рени при $N = 16$ и $p = 0,5$ (ER-0.5) или переменном p (ER-mix). Несмотря на охват всех вероятностей рёбер, модель не способна воспроизводить всё многообразие совместных структурных характеристик.
Fig. 1. Average vertex degree and average clustering coefficient in the Erdős–Rényi model with $N = 16$ and $p = 0,5$ (ER-0.5) or variable p (ER-mix). Despite covering all edge probabilities, the model is unable to reproduce the full diversity of joint structural characteristics.

Пусть теперь задано мультимножество $S = G_1, \dots, G_{|S|}$. На протяжении работы мы будем рассматривать неориентированные графы без петель и кратных рёбер. Пусть также задана мера расстояния $D(G, G')$, оценивающая различие двух графов. Тогда разнообразие определим как

$$\text{Diversity}(S) = F(\{D(G, G') : G, G' \in S\}),$$

где F – некоторая функция, вычисляющая разнообразие по набору попарных расстояний. Выбор F также нетривиален; возможные подходы обсуждаются в разделе 2.3.

Определив расстояние $D(\cdot, \cdot)$ и меру разнообразия, формулируем основную цель – нужно найти мультимножество графов $\bar{S}$, максимизирующее его разнообразие:

$$\bar{S} = \arg \max_{G_1, G_2, \dots, G_{|S|} \in \mathcal{G}_N} \text{Diversity}(\{G_1, G_2, \dots, G_{|S|}\}),$$

где $\mathcal{G}_N$ – множество всех графов с N вершинами.²

2.2 Расстояния между графами

Измерение сходства между графами является предпосылкой оптимизации разнообразия. Принципиальная трудность состоит в том, что графы не имеют канонического упорядочения вершин: два изоморфных графа с переставленными индексами вершин должны иметь нулевое расстояние, однако вычисление этого требует решения задачи изоморфизма графов, которая не является полиномиальной в общем случае [10]. Поэтому практические расстояния

² Для простоты предполагается, что число вершин N одинаково для всех графов. Заметим, что N естественно может служить верхней границей числа вершин: меньшие графы получаются, если некоторые вершины имеют нулевую степень.

между графами жертвуют точностью ради приемлемой вычислительной сложности; разные меры чувствительны к различным структурным свойствам. В настоящей работе используются три представительных расстояния: NetLSD, Graphlet Correlation Distance (GCD) и Portrait Divergence [10-13]. В отличие от упомянутых выше методов, *расстояние редактирования графа* (Graph Edit Distance, GED) определяет расстояние между двумя графами как минимальное число элементарных операций редактирования – вставок, удалений и замен вершин и рёбер, – необходимых для преобразования одного графа в другой. GED теоретически привлекателен, однако его точное вычисление является NP-трудным [10], поэтому в экспериментах он не используется.

NetLSD [11] представляет каждый граф через его тепловой след – компактную спектральную сигнатуру, получаемую моделированием диффузии тепла на граф-лапласиане при различных временных масштабах. Поскольку спектр лапласиана кодирует глобальную связность, структуру сообществ и паттерны циклов, NetLSD чувствителен к крупномасштабной топологии и при этом инвариантен к перемаркировке вершин.

Graphlet Correlation Distance (GCD) [12] работает с локальной статистикой подграфов. Каждая вершина описывается частотой участия в каждом из 11 неизбыточных паттернов графлетов (связных подграфов из не более чем четырёх вершин). Поузловые векторы суммируются в матрицу корреляций графлетов; расстояние между двумя графами равно евклидову расстоянию между их матрицами корреляций. GCD особенно чувствителен к различиям в плотности треугольников, кластеризации и структуре малых циклов.

Portrait Divergence [13] кодирует полное распределение длин кратчайших путей в графе: для каждого возможного расстояния l и числа вершин k фиксируется, сколько вершин имеют ровно k других вершин на расстоянии l . Расстояние между двумя графами равно дивергенции Йенсена–Шеннона между их портретными распределениями. Данная мера одновременно захватывает локальную структуру окрестностей и глобальную статистику длин путей.

В настоящей работе мы не стремимся определить, какое расстояние между графами является универсально наилучшим. Каждое расстояние выявляет конкретные структурные свойства, и полученные разнообразные графы существенно отличаются в зависимости от того, какое расстояние оптимизируется. Наши эксперименты освещают эти различия и дают дополнительное понимание того, к каким структурным характеристикам каждое расстояние наиболее чувствительно.

2.3 Желательные свойства и существующие меры разнообразия

Зафиксировав попарное расстояние D , необходимо выбрать функцию агрегации F . Рассмотрим, почему распространённые варианты менее подходят для нашей задачи, и сформулируем свойства, которым должна удовлетворять хорошая мера разнообразия.

Наиболее естественной и широко используемой мерой является *среднее попарное расстояние* между элементами. Однако данная мера не подходит, поскольку её оптимизация может приводить к вырожденным конфигурациям. Как показано на рис. 2 (слева), оптимизация среднего попарного расстояния вынуждает многие точки коллапсировать в одну (на концах отрезка). Это происходит потому, что мера Average не учитывает уникальность и изолированность элементов набора данных.

Ещё одной возможной мерой является *минимальное попарное расстояние* (Bottleneck). Однако эта мера нечувствительна ко всем расстояниям, кроме минимального, и не способна различать принципиально разные конфигурации.

Недавняя работа [14] предлагает измерять разнообразие через Vendi Score (VS). Данная оценка требует задания ядерной функции на парах элементов; матрица схожести из попарных значений ядра строится, а Vendi Score определяется как экспонента энтропии Шеннона собственных значений нормализованной матрицы схожести.

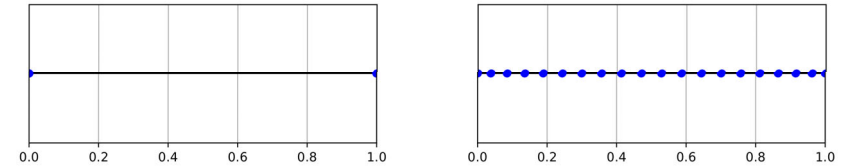


Рис. 2. Точки, оптимизированные по мерам разнообразия Average (слева) или Energy (справа) на отрезке. Оптимизация по Average сжимает большинство точек к концам отрезка, тогда как Energy порождает равномерно распределённую конфигурацию.

Fig. 2. Points optimized by the diversity measures Average (left) or Energy (right) on a segment. Optimization by Average compresses most points toward the ends of the segment, while Energy produces a uniformly distributed configuration.

Другая недавняя работа [15] исследует задачу измерения покрытия (разнообразия) набора S в контексте генерации молекул. Авторы работы [15] предлагают три аксиомы, которым должна удовлетворять хорошая мера покрытия: монотонность, субаддитивность и несходство. Показывается, что ни одна из рассмотренных мер не удовлетворяет всем трём аксиомам одновременно, и предлагается новая мера – #Circles. Хотя #Circles удовлетворяет всем аксиомам, она обладает двумя недостатками: экспоненциальная сложность вычисления и зависимость от параметра радиуса, хорошее значение которого зависит от набора данных. В нашей постановке, где набор графов динамически меняется в процессе оптимизации, данная мера неприменима.

Мотивированные этими наблюдениями, мы формулируем два свойства, которым должна удовлетворять хорошая мера разнообразия. Предположим, что задано мультимножество S из $|S|$ элементов и для каждой пары $G, G' \in S$ известно расстояние $D(G, G')$.

Монотонность. Пусть заданы два мультимножества S, S' из $|S|$ различных элементов и биекция $g: S \rightarrow S'$. Предположим, что для любых $G_i, G_j \in S$ выполняется $D(G_i, G_j) \leq D(g(G_i), g(G_j))$ при строгом неравенстве хотя бы для одной пары. Тогда требуем $\text{Diversity}(S) < \text{Diversity}(S')$. Данное свойство отражает суть мер разнообразия: большие попарные расстояния должны приводить к более высоким значениям разнообразия.

Уникальность. Пусть S состоит из $|S|$ различных элементов $G_1, \dots, G_{|S|}$. Обозначим через S_{ij} мультимножество, полученное из S удалением G_i и добавлением второй копии G_j ($j \neq i$), то есть $S_{ij} := (S \setminus \{G_i\}) \cup \{G_j\}$. Тогда требуем $\text{Diversity}(S) > \text{Diversity}(S_{ij})$. Иными словами, данное свойство требует, чтобы добавление уникального элемента при фиксированных остальных повышало разнообразие больше, чем дублирование уже существующего.

Оказывается, ни одна из рассмотренных мер не удовлетворяет одновременно монотонности и уникальности; это формально показано в [16].

2.4 Мера Energy

Мотивированные энергией системы одинаково заряженных частиц, мы определяем *Energy* набора графов S как

$$-\frac{1}{|S|(|S|-1)} \sum_{i \neq j} \frac{1}{D(G_i, G_j)^\gamma}$$

Параметр $\gamma > 0$ управляет силой штрафа за малые попарные расстояния. В предельном случае $\gamma \rightarrow \infty$ данная мера становится эквивалентной Bottleneck. В экспериментах используется $\gamma = 1$, так что данное выражение естественно интерпретируется как среднее попарное взаимодействие для системы равно заряженных частиц (умножение на -1 обеспечивает большие значения для более разнообразных наборов).

Утверждение 1. Мера Energy удовлетворяет как монотонности, так и уникальности.

Монотонность вытекает непосредственно из формы Energy: каждый член $1/D(G_i, G_j)^\gamma$ убывает при увеличении соответствующего попарного расстояния, поэтому увеличение любого расстояния строго уменьшает сумму и строго увеличивает Energy (рис. 3). Уникальность выполняется, поскольку любая дублированная пара вносит член $1/D(G_i, G_j)^\gamma \rightarrow \infty$ при $D \rightarrow 0$, сводя значение Energy к $-\infty$, что всегда хуже любой конфигурации из различных элементов. Формальное доказательство приведено в [16].

Несмотря на желательные свойства, Energy может быть неограниченно велика, когда два элемента оказываются очень близкими. Однако, как показывает недавняя работа [17], лучших альтернатив пока нет: все меры, удовлетворяющие монотонности и уникальности даже при совпадении элементов, являются NP-трудными для вычисления. Для численной устойчивости в экспериментах к знаменателю Energy добавляется малая константа ϵ .

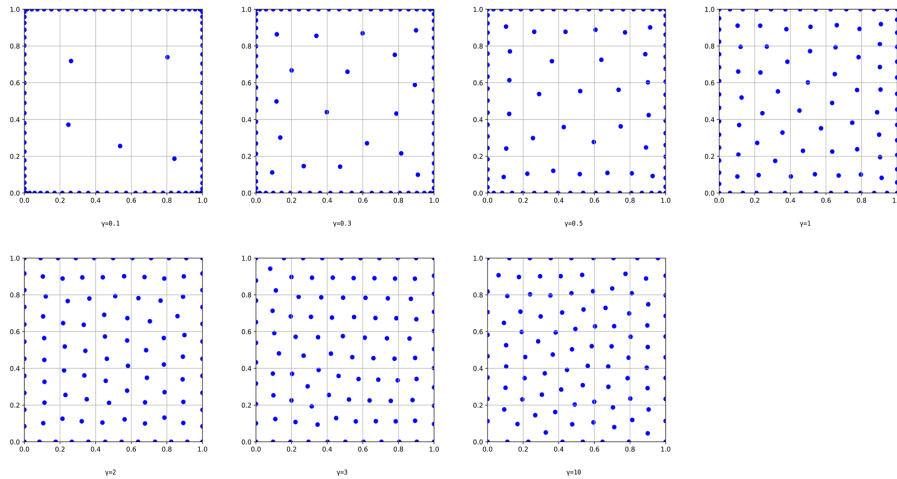


Рис. 3. Влияние параметра γ меры Energy на конфигурацию точек в квадрате (слева направо – $\gamma \in \{0,1; 0,3; 0,5; 1; 2; 3; 10\}$). При малых γ покрытие внутренней области недостаточно; при $\gamma \geq 1$ распределение хорошо равномерно.

Fig. 3. The influence of the parameter γ of the Energy measure on the configuration of points in a square (from left to right $\gamma \in \{0,1; 0,3; 0,5; 1; 2; 3; 10\}$). For small γ the coverage of the interior region is insufficient; for $\gamma \geq 1$, the distribution is well uniform..

В табл. 1 перечислены меры разнообразия, обсуждаемые в предшествующих работах [14,15]. Оценки монотонности и уникальности для всех мер доказаны в [16].

Теорема 1. Среди мер, перечисленных выше строки Energy в табл. 1, монотонности удовлетворяют только Average и SumAverage, а уникальности – только Bottleneck.

Доказательство приведено в работе [16].

3. Алгоритмы оптимизации разнообразия графов

В данном разделе предлагается несколько подходов к оптимизации разнообразия – от базовых методов на основе генераторов случайных графов до более продвинутых, основанных на нейронном генеративном моделировании.

Наши алгоритмы применимы к произвольным расстояниям между графами и мерам разнообразия. Однако из соображений масштабируемости мы ограничиваемся мерами, представимыми в следующем виде. Пусть задан набор S и элемент $G \in S$; обозначим $S' = S \setminus \{G\}$.

Табл. 1. Сравнение известных мер разнообразия. Energy удовлетворяет как монотонности, так и уникальности.

Table 1. Comparison of known diversity measures. Energy satisfies both monotonicity and uniqueness.

Мера	Формула	Монотонность	Уникальность
Average	$\frac{2}{ S (S -1)} \sum_{i \neq j} D(G_i, G_j)$	✓	✗
SumAverage	$\frac{1}{ S } \sum_{i \neq j} D(G_i, G_j)$	✓	✗
Diameter	$\max_{i \neq j} D(G_i, G_j)$	✗	✗
SumDiameter	$\sum_i \max_{j \neq i} D(G_i, G_j)$	✗	✗
Bottleneck	$\min_{i \neq j} D(G_i, G_j)$	✗	✓
SumBottleneck	$\sum_i \min_{j \neq i} D(G_i, G_j)$	✗	✗
#Circles(t)	$\max_C C $ при $D(G_i, G_j) > t \forall i \neq j \in C$	✗	✗
Vendi Score	$\exp(-\sum_i \lambda_i \log \lambda_i)$	✗	✗
Energy (наша)	$-\frac{1}{ S (S -1)} \sum_{i \neq j} D(G_i, G_j)^{-\gamma}$	✓	✓

Тогда разнообразие исходного набора выражается как

$$\text{Diversity}(\{G\} \cup S) = g(f(G, S), c(S)),$$

где g монотонна по обоим аргументам, $f(G, S)$ зависит только от расстояний $\{D(G, G_i) : G_i \in S\}$, а $c(S)$ – только от S . Назовём $f(G, S)$ *приспособленностью* графа G относительно набора S . Меры Energy и Bottleneck удовлетворяют данному разложению.

3.1 Жадный алгоритм

Основная идея жадного алгоритма – итеративное построение набора разнообразных графов путём добавления на каждом шаге наиболее подходящего графа из predetermined множества $\hat{S}$ значительно большего размера (см. Алгоритм 1). Это множество может быть подано пользователем, получено другим алгоритмом или сгенерировано моделями случайных графов. Процесс начинается со случайного выбора графа из $\hat{S}$. На каждом шаге из $\hat{S}$ выбирается наиболее подходящий граф по приспособленности $f(G, S)$ относительно текущего набора S .

Несмотря на простоту, жадный алгоритм оказывается весьма эффективным при достаточно разнообразном начальном множестве $\hat{S}$. Следующая теорема даёт теоретическую гарантию.

Теорема 2. Предположим, что функция разнообразия – Energy, Average или Bottleneck. Пусть $\hat{S}$ – некоторый набор графов. Если жадный алгоритм выбрал подмножество S из $\hat{S}$, то $\text{Diversity}(S) \geq 1/2 \text{ Diversity}(\bar{S})$ для Average и Bottleneck и $\text{Diversity}(S) \geq 2^\gamma \text{ Diversity}(\bar{S})$ для Energy (γ), где $\bar{S}$ – максимально разнообразное подмножество $\hat{S}$ размера $|S|$.

Доказательства для Average, Bottleneck и Energy приведены в [16]. Данный результат даёт гарантию аппроксимации с коэффициентом два для жадного алгоритма. Для Average и Bottleneck жадный результат составляет не менее половины от оптимального подмножества той же размерности из $\hat{S}$. Для Energy с параметром γ граница $\text{Diversity}(S) \geq 2^\gamma \text{ Diversity}(\bar{S})$ при $\gamma = 1$ также является гарантией с коэффициентом два.

Утверждение 2. Временная сложность жадного алгоритма равна $O((a + b|S|)|\hat{S}|)$, где $|S|$ – размер желаемой разнообразной популяции, $|\hat{S}|$ – размер исходной популяции, a – стоимость вычисления представления графа, b – стоимость вычисления расстояния между двумя представлениями.

Алгоритм 3 Алгоритм локальной оптимизации разнообразия

Вход: Начальная популяция графов S размера $|S|$, лимит обновлений L , окно стагнации K

Выход: Разнообразная популяция графов S размера $|S|$

- 1: Вычислить приспособленность $f(G, S \setminus \{G\})$ для всех $G \in S$
- 2: **для** $\ell = 1, \dots, L$ **выполнить**
- 3: $U \leftarrow$ выбрать граф из S с вероятностью $\propto 1/f$ (или $\propto -f$ для Energy)
- 4: Выбрать две различные вершины $u, v \in V(U)$ равновероятно
- 5: **если** ребро $(u, v) \in U$ **то** удалить (u, v) из U , получив U'
- 6: **иначе** добавить (u, v) в U , получив U'
- 7: **конец если**
- 8: Вычислить $f(U', S \setminus \{U\})$
- 9: **если** $f(U', S \setminus \{U\}) > f(U, S \setminus \{U\})$ **то**
- 10: $S \leftarrow (S \setminus \{U\}) \cup \{U'\}$
- 11: **иначе**
- 12: Записать неудачную попытку; если последние K попыток неудачны, принять наименее ухудшающий U' среди них
- 13: **конец если**
- 14: **конец для**
- 15: **Вернуть** S

Алгоритм 4 Итеративное генеративное моделирование графов (IGGM)

Вход: Размер батча генерации K , общий бюджет L (кратен K), размер удержания $R < K$, желаемый размер выхода $|S| \leq R$

Выход: Множество S из $|S|$ структурно разнообразных графов

- 1: Сгенерировать $10 \cdot R$ графов из ER-mix; применить Алгоритм 1 для получения S_0 размера R // Инициализация
- 2: **для** $t = 1, \dots, L/K$ **выполнить** // Итеративное уточнение
- 3: Обучить генеративную модель g_{θ_t} на S_{t-1} , инициализируя весами $g_{\theta_{t-1}}$
- 4: Сгенерировать K графов с помощью g_{θ_t}
- 5: Применить Алгоритм 1 к K сгенерированным графам для получения S_t размера R
- 6: **конец для**
- 7: Применить Алгоритм 1 к $S_{L/K}$ для выбора $|S|$ разнообразных графов
- 8: **Вернуть** выбранные $|S|$ графов

Алгоритм IGGM не зависит от выбора генеративной модели: любая модель, способная генерировать графы фиксированного размера, подходит в качестве g_{θ} . В экспериментах используется DiGress [20] – модель на основе дискретной шумовой диффузии, которая на момент работы была актуальной моделью генерации графов и распространяется под лицензией MIT. В экспериментах установлено $R = 10^3$, $K = 10^5$, $L = 10^6$.

4. Эксперименты

4.1 Описание протокола экспериментов

Рассматриваются четыре расстояния между графами: NetLSD-heat, NetLSD-wave, GCD и Portrait Divergence. Согласно анализу из раздела 2.4, в качестве меры разнообразия выбрана Energy. Формально оптимизируется и отображается следующая мера:

$$\frac{1}{|S|(|S| - 1)} \sum_{i \neq j} \frac{1}{D(G_i, G_j) + \epsilon}$$

где ϵ – малая константа для численной устойчивости.

Оцениваются все четыре предложенных алгоритма. В качестве базовых методов используются графы Эрдёша–Рени с различными p (ER-mix) и выборка из разнообразных генераторов случайных графов. Для иллюстрации также включена выборка из тестового набора данных GraphWorld [7]. В большинстве экспериментов генерируется $|S| = 100$ графов с $N = 16$ вершинами; дополнительно проводятся эксперименты с нейронными алгоритмами на графах с $N = 64$ вершинами. Все алгоритмы, кроме IGGM, используют суммарный лимит в 3 млн сгенерированных графов; для IGGM лимит равен 1 млн.

Для генерации достаточно разнообразного начального набора $\hat{S}$ жадного алгоритма используется смесь семи моделей случайных графов с разнообразными структурными свойствами: Эрдёша–Рени, присоединение по предпочтению (Preferential Attachment), Holme–Kim, случайные графы со степенно-законным ожидаемым распределением степеней, случайные геометрические графы, случайные регулярные графы и стохастическая блочная модель [9,21–25].

4.2 Численное сравнение алгоритмов разнообразия

В табл. 2 приведены результаты оптимизации Energy со стандартными отклонениями по пяти независимым запускам для графов с $N = 16$ вершинами.

Все предложенные алгоритмы существенно превосходят базовые методы ER-mix и Random Graph Generators. Разнообразие GraphWorld также далеко от оптимального. Среди нейронных алгоритмов наилучший результат достигается комбинацией Greedy→Genetic→LocalOpt. Такая комбинация естественна: Greedy из смеси генераторов обеспечивает начальный разнообразный набор; Genetic вносит достаточную случайность для создания новых структурных паттернов; LocalOpt выполняет финальную тонкую подстройку малыми модификациями. Нейросетевой метод IGGM даёт значительный прирост разнообразия для GCD и Portrait-div и конкурентоспособные результаты для NetLSD-heat, используя меньше сгенерированных графов за счёт необходимости многократного обучения генеративной модели.

Эксперименты на графах с $N = 64$ вершинами показаны в табл. 3 и согласуются с результатами на меньших графах: предложенные алгоритмы существенно превосходят базовые методы, а комбинация Greedy→Genetic→LocalOpt продолжает давать наилучшие среди алгоритмических (не-нейронных) методов.

Табл. 2. Результаты оптимизации Energy для графов с $N = 16$ вершинами ($|S| = 100$), стандартные отклонения по пяти независимым запускам. Значения представляют собой отрицание меры Energy; значения, более близкие к нулю, означают большее разнообразие.
Table 2. Energy optimization results for graphs with $N = 16$ vertices ($|S| = 100$), standard deviations across five independent runs. Values represent the negative of the Energy measure; values closer to zero indicate greater diversity.

Метод	GCD	Portrait-div	NetLSD-heat	NetLSD-wave
ER-mix	0.281 ± 0.0350	43.057 ± 8.812	72.387 ± 26.192	0.583 ± 0.0892
GraphWorld	0.466 ± 0.0151	3.917 ± 0.0896	5.108 ± 0.2856	0.621 ± 0.0117
Random Graph Generators	0.553 ± 0.0167	6.009 ± 0.2368	116.685 ± 8.009	1.334 ± 0.0831
Greedy[1M]	0.160 ± 0.0004	1.287 ± 0.0029	0.682 ± 0.0005	0.124 ± 0.0003
Greedy[2M]	0.157 ± 0.0010	1.278 ± 0.0033	0.681 ± 0.0008	0.124 ± 0.0004
Greedy[3M]	0.156 ± 0.0004	1.274 ± 0.0018	0.681 ± 0.0001	0.123 ± 0.0003
ER-mix→Genetic[3M]	0.139 ± 0.0025	1.264 ± 0.0031	0.677 ± 0.0013	0.117 ± 0.0003
Greedy[1M]→Genetic[2M]	0.139 ± 0.0018	1.263 ± 0.0020	0.674 ± 0.0003	0.118 ± 0.0005
Greedy[2M]→Genetic[1M]	0.141 ± 0.0012	1.263 ± 0.0027	0.674 ± 0.0004	0.118 ± 0.0005
ER-mix→Genetic[1M]→LocalOpt[2M]	0.138 ± 0.0002	1.259 ± 0.0007	0.675 ± 0.0000	0.117 ± 0.0001
Greedy[1M]→LocalOpt[2M]	0.139 ± 0.0012	1.255 ± 0.0006	0.679 ± 0.0001	0.118 ± 0.0001
Greedy[1M]→Genetic[1M]→LocalOpt[1M]	0.135 ± 0.0000	1.245 ± 0.0004	0.673 ± 0.0001	0.117 ± 0.0001
IGGM[1M]	0.120 ± 0.0003	1.213 ± 0.0001	0.675 ± 0.0002	0.148 ± 0.0001

Табл. 3. Результаты оптимизации Energy для графов с $N = 64$ вершинами. Значения, более близкие к нулю, означают большее разнообразие. Результаты одного запуска; стандартные отклонения не приводятся из-за высокой вычислительной стоимости.
Table 3. Energy optimization results for graphs with $N = 64$ vertices. Values closer to zero indicate greater diversity. Results from a single run; standard deviations are not provided due to computational cost.

Метод	GCD	Portrait-div	NetLSD-heat	NetLSD-wave
ER-mix	0.400	2.236	452.885	0.454
GraphWorld	0.442	3.746	3.509	0.510
Random Graph Generators	0.540	5.868	112.685	1.298
Greedy[1M]	0.177	1.155	0.812	0.172
Greedy[3M]	0.175	1.148	0.796	0.169
ER-mix→Genetic [3M]	0.167	1.128	0.567	0.128
Greedy[1M]→Genetic[2M]	0.158	1.126	0.673	0.117
ER-mix→Genetic[1M]→LocalOpt[2M]	0.133	1.086	0.551	0.118
Greedy[1M]→LocalOpt[2M]	0.132	1.082	0.603	0.117
Greedy[1M]→Genetic[1M]→LocalOpt[1M]	0.128	1.060	0.673	0.116

4.3 Примеры сгенерированных графов

Поскольку в основных экспериментах сгенерировано 100 графов, каждый из которых имеет всего 16 вершин, возможна визуальная инспекция результатов. Рис. 4 показывает выборку из результата генетического алгоритма с разнообразием на основе Portrait-div. Очевидно, что графы отличаются по плотности, внутренней структуре, числу циклов и планарности, и существенно отличаются от входного распределения ER-mix.

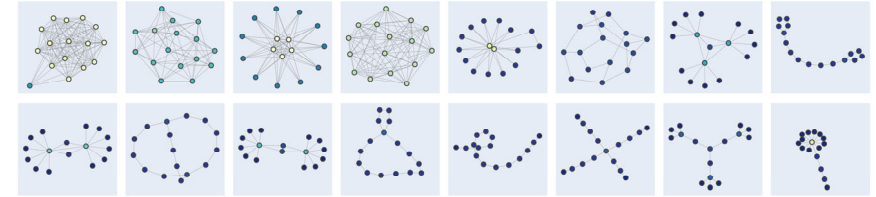


Рис. 4. Выборка графов, сгенерированных генетическим алгоритмом с мерой разнообразия Portrait-div. Графы демонстрируют разнообразные структуры: различные плотности, числа циклов, планарность и локальные паттерны.

Fig. 4. A sample of graphs generated by a genetic algorithm with the Portrait-div diversity measure. The graphs exhibit diverse structures: different densities, cycle numbers, planarity, and local patterns.

4.4 Структурные характеристики и чувствительность расстояний

Анализируются структурные характеристики сгенерированных графов. Рис. 5 визуализирует различные характеристики для базового метода ER-mix, IGGM и комбинации Greedy, Genetic и LocalOpt. Наглядно видно, что по сравнению с ER-mix наши алгоритмы приводят к значительно более разнообразным парам характеристик.

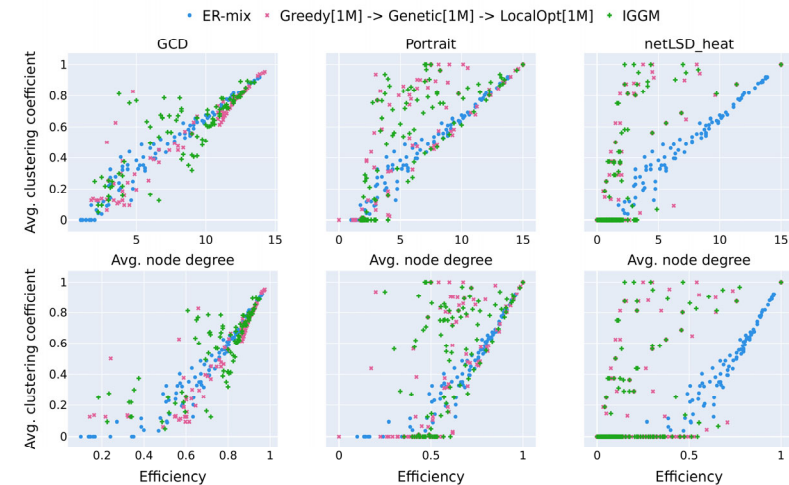


Рис. 5. Совместное распределение характеристик графов для расстояний GCD, Portrait-div и NetLSD-heat. По сравнению с ER-mix, предложенные алгоритмы (IGGM и Greedy→Genetic→LocalOpt) покрывают более широкие области пространства характеристик.

Fig. 5. Joint distribution of graph features for GCD, Portrait-div, and NetLSD-heat distances. Compared to ER-mix, the proposed algorithms (IGGM and Greedy→Genetic→LocalOpt) cover significantly wider regions of the feature space.

Рис. 6 расширяет структурный анализ на дополнительные пары характеристик и включает базовый метод GraphWorld. Ряд закономерностей воспроизводится по всем столбцам.

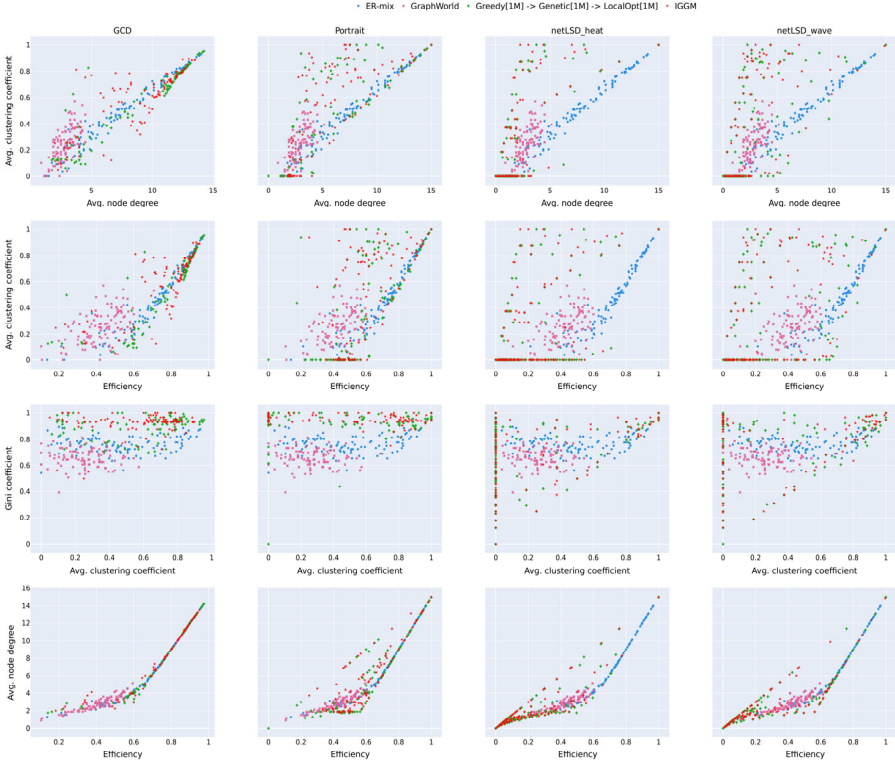


Рис. 6. Расширенные совместные распределения структурных характеристик по всем четырём мерам расстояния (столбцы: GCD, Portrait, netLSD-heat, netLSD-wave) и четырём парам характеристик (строки: ср. коэффициент кластеризации против ср. степени вершины; ср. коэффициент кластеризации против эффективности; коэффициент Джини против ср. коэффициента кластеризации; ср. степень вершины против эффективности). Отображены четыре метода: ER-mix, GraphWorld, Greedy→Genetic→LocalOpt, IGGM. Предложенные алгоритмы покрывают значительно более широкие области каждого пространства характеристик.

Fig. 6. Extended joint distributions of structural features across all four distance measures (columns: GCD, Portrait, netLSD-heat, netLSD-wave) and four pairs of features (rows: avg. clustering coefficient vs. avg. vertex degree; avg. clustering coefficient vs. efficiency; Gini coefficient vs. avg. clustering coefficient; avg. vertex degree vs. efficiency). Four methods are shown: ER-mix, GraphWorld, Greedy→Genetic→LocalOpt, IGGM. The proposed algorithms cover significantly wider regions of each feature space.

Во-первых, ER-mix и GraphWorld образуют узкие кластеры или почти диагональные полосы в большинстве панелей, что подтверждает: стандартные генераторы обеспечивают ограниченное структурное покрытие. Во-вторых, предложенные алгоритмы, – особенно Greedy→Genetic→LocalOpt и IGGM (рис. 7) – рассеиваются широко по каждой панели, заполняя области, недостижимые базовыми методами. В-третьих, столбцы netLSD демонстрируют выраженное смещение в сторону разреженности: большинство сгенерированных графов кластеризуются при нулевом среднем значении степени и эффективности, тогда как GCD и Portrait порождают более структурно разнообразные популяции.

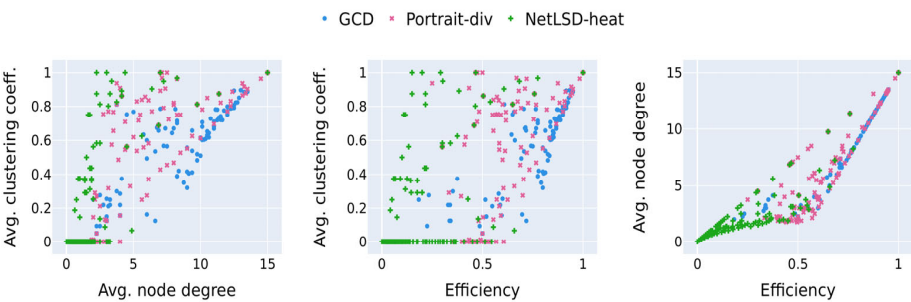


Рис. 7. Совместное распределение характеристик графов из IGGM в зависимости от используемой меры расстояния. Каждое расстояние порождает структурно различные типы разнообразных графов.

Fig. 7. Joint distribution of graph characteristics from IGGM depending on the distance measure used. Each distance generates structurally different types of diverse graphs.

4.5 Оценка на задачах нейронного алгоритмического рассуждения

Как обсуждалось в разд. 1, ключевой мотивацией для структурно разнообразных графов является оценка моделей NAR на тестовом наборе данных CLRS [4]. Мы показываем это, обучая стандартную модель NAR на обучающих графах данного тестового набора и оценивая её на двух тестовых наборах: стандартном генераторе CLRS и наборе, порождённом генетическим алгоритмом с расстоянием Portrait-Div. В табл. 4 приведена средняя точность по пяти запускам. Точность падает во всех трёх случаях, подтверждая, что структурно разнообразные графы являются более сложным тестом, чем стандартное распределение ER-0.5.

Табл. 4. Точность NAR (%) на тестовой выборке CLRS против набора, порождённого комбинацией Genetic + Portrait-Divergence.

Table 4. NAR accuracy (%) on the CLRS test set versus the set generated by the Genetic + Portrait-Divergence combination.

Алгоритм	CLRS	Genetic
Bellman-Ford	92	87
Dijkstra	92	85
MST-Prim	84	83

5. Заключение

В настоящей работе сформулирована и исследована задача генерации структурно разнообразных графов, способных служить репрезентативными экземплярами для различных задач, связанных с графами. Показано, что задача является нетривиальной: не просто определить, что значит «разнообразный» набор графов, а стандартные модели случайных графов не обеспечивают достаточного разнообразия. Предложены желательные свойства (монотонность и уникальность) и идентифицирована мера Energy как наилучший из доступных вариантов. Предложены несколько алгоритмов – жадный, генетический, локальная оптимизация и нейронный (IGGM), – которые эмпирически существенно превосходят наивную генерацию случайных графов как по Energy, так и по среднему попарному расстоянию. Наилучшие ненейронные результаты достигает комбинация Greedy→Genetic→LocalOpt, тогда как IGGM достигает наибольшего разнообразия для мер GCD и Portrait-div.

Дополнительным результатом настоящей работы является более глубокое эмпирическое понимание расстояний между графами. Поскольку разные расстояния порождают разные типы разнообразных графов, визуализация структурных свойств сгенерированных наборов раскрывает, к чему чувствительно каждое расстояние. NetLSD сильно смещён в сторону разреженных графов; GCD предпочитает более плотные структуры; Portrait-div захватывает и локальные, и глобальные свойства и порождает наиболее визуально разнообразные графы. Несколько перспективных направлений остаются для будущих исследований. Хотя текущие эксперименты фиксируют число вершин N , сами предложенные алгоритмы поддерживают наборы графов переменного размера; эксперименты со смешанными по размеру коллекциями – естественное расширение. Масштабируемость зависит не только от алгоритма разнообразия, но и от скорости базовой генеративной модели и вычисления расстояний между графами. Расширение фреймворка на графы с признаками вершин и рёбер, а также на ориентированные графы – ещё одно перспективное направление. Тем не менее, многие задачи остаются нерешёнными, и мы надеемся, что настоящая работа будет способствовать дальнейшим теоретическим и практическим исследованиям в этой области.

Список литературы / References

- [1]. Boccaletti S. et al. Complex networks: Structure and dynamics. Physics Reports. Elsevier, 2006, vol. 424, no. 4-5, pp. 175-308.
- [2]. Georgiev D. G. et al. Beyond Erdős-Rényi: Generalization in algorithmic reasoning on graphs. The Second Learning on Graphs Conference, 2023.
- [3]. Veličković P., Blundell C. Neural algorithmic reasoning. Patterns. Elsevier, 2021, vol. 2, no. 7, p. 100273.
- [4]. Veličković P. et al. The CLRS Algorithmic Reasoning Benchmark. Proceedings of the 39th International Conference on Machine Learning, 2022.
- [5]. Rodionov G., Prokhorenkova L. Neural Algorithmic Reasoning Without Intermediate Supervision. Advances in Neural Information Processing Systems, 2023, vol. 36.
- [6]. Rodionov G., Prokhorenkova L. Discrete Neural Algorithmic Reasoning. Proceedings of the 42nd International Conference on Machine Learning, 2025.
- [7]. Palowitch J. et al. GraphWorld: Fake graphs bring real insights for GNNs. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 3691-3701.
- [8]. Karrer B., Newman M. E. Stochastic blockmodels and community structure in networks. Physical review E. APS, 2011, vol. 83, no. 1, p. 016107.
- [9]. Erdős P., Rényi A. et al. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 1960, vol. 5, no. 1, pp. 17-60.
- [10]. Tantardini M. et al. Comparing methods for comparing networks. Scientific Reports. Springer, 2019, vol. 9, no. 1, pp. 1-19.
- [11]. Tsitsulin A. et al. NetLSD: hearing the shape of a graph. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2347-2356.
- [12]. Yaveroglu Ö. N. et al. Revealing the hidden language of complex networks. Scientific Reports. Nature Publishing Group UK London, 2014, vol. 4, p. 4547.
- [13]. Bagrow J. P., Bollt E. M. An information-theoretic, all-scales approach to comparing networks. Applied Network Science. Springer, 2019, vol. 4, no. 1.
- [14]. Friedman D., Dieng A. B. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. Transactions on Machine Learning Research, 2023.
- [15]. Xie Y. et al. How Much Space Has Been Explored? Measuring the Chemical Space Covered by Databases and Machine-Generated Molecules. The Eleventh International Conference on Learning Representations, 2023.
- [16]. Velikonitvsev F., Mironov M., Prokhorenkova L. Challenges of Generating Structurally Diverse Graphs. Advances in Neural Information Processing Systems (NeurIPS), 2024.
- [17]. Mironov M., Prokhorenkova L. Measuring Diversity: Axioms and Challenges. arXiv preprint, 2024. DOI: 10.48550/arXiv.2410.14556. 2024.
- [18]. You J. et al. GraphRNN: Generating realistic graphs with deep auto-regressive models. International conference on machine learning. PMLR, 2018, pp. 5708-5717.
- [19]. Martinkus K. et al. SPECTRE: Spectral Conditioning Helps to Overcome the Expressivity Limits of One-shot Graph Generators, 2022.

- [20]. Vignac C. et al. DiGress: Discrete Denoising diffusion for graph generation. The Eleventh International Conference on Learning Representations, 2023.
- [21]. Barabási A.-L., Albert R. Emergence of scaling in random networks. Science. American Association for the Advancement of Science, 1999, vol. 286, no. 5439, pp. 509-512.
- [22]. Holme P., Kim B. J. Growing scale-free networks with tunable clustering. Physical review E. APS, 2002, vol. 65, no. 2, p. 026107.
- [23]. Chung F., Lu L. Connected components in random graphs with given expected degree sequences. Annals of Combinatorics. Springer, 2002, vol. 6, no. 2, pp. 125-145.
- [24]. Penrose M. Random geometric graphs. Oxford University Press, 2003.
- [25]. Holland P. W., Laskey K. B., Leinhardt S. Stochastic blockmodels: First steps. Social networks. Elsevier, 1983, vol. 5, no. 2, pp. 109-137.

Информация об авторах / Information about authors

Фёдор Сергеевич ВЕЛИКОНИВЦЕВ – аспирант, младший научный сотрудник научно-учебной лаборатории компании Яндекс факультета компьютерных наук НИУ ВШЭ. Сфера научных интересов: машинное обучение на графах, методы эффективного глубокого обучения, программирование на графических ускорителях.

Fedor Sergeevich VELIKONITVSEV – postgraduate student, Junior Research Fellow at the Yandex Research and Education Laboratory, Faculty of Computer Science, HSE University. His research interests include graph machine learning, efficient deep learning methods, and GPU programming.