Федеральное государственное автономное образовательное учреждение
высшего образования «Московский физико-технический институт
(национальный исследовательский университет)»

*На правах рукописи*

**Прохоренкова Людмила Александровна**

# МОДЕЛИ СЛОЖНЫХ СЕТЕЙ И АЛГОРИТМЫ НА ГРАФАХ

РЕЗЮМЕ

диссертации на соискание ученой степени
доктора компьютерных наук

Москва - 2021

## Тема диссертации

Многие наблюдаемые системы можно представить в виде графа, вершины которого — элементы системы, а ребра — отношения между этими элементами. Многочисленные экспериментальные исследования показывают, что наблюдаемые сложные сети как правило обладают некоторыми типичными свойствами: степенное распределение степеней вершин, малый диаметр, наличие кластеров и так далее. В литературе предложен ряд моделей случайных графов, которые обадают некоторыми количественными и топологическими свойствами наблюдаемых сетей. Такие модели используются в экспериментальной физике, биоинформатике, информационном поиске и многих других приложениях [2, 9]. Изучение свойств сложных сетей и их моделей имеет важнейшее значение для понимания принципов формирования таких структур, прогнозирования их динамики и разработки эффективных алгоритмов.

Наиболее известным свойством сложных сетей является их распределение степеней вершин. Для большинства проанализированных сетей показано, что распределение степеней вершин подчиняется (приблизительно) степенному закону [9]. Это явление часто объясняется принципом *предпочтительного присоединения* [6], который лежит в основе многих моделей сложных сетей [10, 17, 22].

Другой ключевой характеристикой реальных сетей является наличие в них *сообществ* (кластеров), то есть групп вершин, сильно взаимосвязанных между собой и относительно хорошо отделенных от остальной части сети. Например, в социальных сетях сообщества формируются пользователями со схожими интересами; в сетях цитирования они объединяют статьи в определенных областях; в Интернете сообщества могут соответствовать страницам по смежным темам и так далее. Наличие сообществ сильно влияет, например, на рекламное продвижение продуктов с помощью вирусного маркетинга, на распространение инфекционных заболеваний, компьютерных вирусов и информации [47]. Поэтому выделение сообществ является важной и активно изучаемой исследовательской задачей [14, 23, 30].

Помимо выделения сообществ, существуют и другие важные приложения анализа графов, которые будут более подробно рассмотрены ни-

же: обнаружение влиятельных элементов сети [45], поиск ближайших соседей на основе графов близости [4] и другие.

**Цели и задачи исследования** Цель диссертации состоит из двух частей. Во-первых, проанализировать свойства существующих моделей сложных сетей и разработать новые реалистичные модели с желаемыми количественными и топологическими свойствами. Во-вторых, применить методы, основанные на графах, к различным практическим задачам: выделение сообществ, оценка дат публикации веб-страниц, обнаружение влиятельных элементов в сложных сетях и поиск ближайших соседей на основе графов близости.

## Основные результаты

**Модели сложных сетей и их анализ**

- Предложен широкий класс моделей *обобщенного предпочтительного присоединения*, который включает в себя многие существующие модели. Для всего класса теоретически проанализировано распределение степеней вершин, локальный и глобальный коэффициенты кластеризации, а также корреляции степеней вершин [20, 21, 35, 37].
- Доказано, что глобальный коэффициент кластеризации стремится к нулю для всех графов со степенным распределением степеней вершин с бесконечной дисперсией [36, 37].
- Проанализировано асимптотическое поведение модулярности (меры, характеризующей наличие сообществ в графе) для многих моделей случайных графов, включая $d$-регулярные графы, модель предпочтительного присоединения и модель пространственного предпочтительного присоединения [38].
- Предложен новый принцип *предпочтительного размещения*, который позволяет генерировать структуры со степенным распределением размеров кластеров [13].

- Предложена новая модель *предпочтительного присоединения с устареванием*. Показано, что эта модель лучше других моделирует часть Интернета, связанную с медиаконтентом. Теоретически проанализированы основные свойства этой модели [24, 40].

**Выделение сообществ**

- Проведено теоретическое и экспериментальное сравнение алгоритмов выделения сообществ, основанных на максимизации правдоподобия для ряда вероятностных моделей. Предложена более теоретически обоснованная модель для этой задачи [42].

- Проведен систематический анализ следующей проблемы: имея только времена "заражения" вершин графа, требуется выделить сообщества сильно взаимосвязанных вершин. Сравнение существующих и новых подходов на нескольких больших датасетах показало, что наиболее хорошее и стабильное качество достигается предложенными простыми эвристическими подходами [43].

- Проанализирована важная проблема выбора подходящего индекса для оценки алгоритмов выделения сообщества. Предложен теоретический подход: сформулирован список желаемых свойств для индексов сравнения кластеризаций и формально проверено каждое свойство для всех популярных индексов [15].

**Другие приложения анализа графов**

- На основе предложенной модели предпочтительного присоединения с устареванием разработан новый алгоритм датирования вебстраниц [39].

- Предложен новый алгоритм для быстрого выделения вершин большой степени в сложных сетях [5].

- Получены новые теоретические гарантии для алгоритмов поиска ближайших соседей, основанных на графах близости [41].

# Публикации и апробация работы

**Публикации повышенного уровня**

1. L. Ostroumova Prokhorenkova. General results on preferential attachment and clustering coefficient (Общие результаты о предпочтительном присоединении и коэффициенте кластеризации). Optimization Letters, 11(2):279–298, 2017. Web of Science Q2, Scopus Q2.

2. A. Krot and L. Ostroumova Prokhorenkova. Local clustering coefficient in generalized preferential attachment models (Локальный коэффициент кластеризации в моделях обобщенного предпочтительного присоединения). Internet Mathematics, 2017. Scopus Q2. (Главный соавтор; автор диссертации сформулировала проблему, предложила технику доказательства и курировала исследование.)

3. A. Krot and L. Ostroumova Prokhorenkova. Assortativity in generalized preferential attachment models (Ассортативность в моделях обобщенного предпочтительного присоединения). Internet Mathematics, 2017. Scopus Q2. (Главный соавтор; автор диссертации сформулировала проблему, предложила технику доказательства и курировала исследование.)

4. L. Ostroumova Prokhorenkova and E. Samosvat. Recency-based preferential attachment models (Модели предпочтительного присоединения с устареванием). Journal of Complex Networks, 4(4):475-499, 2016. Scopus Q1, Web of Science. (Главный соавтор; автор диссертации сформулировала и доказала все теоретические результаты.)

5. L. Prokhorenkova and A. Tikhonov. Community detection through likelihood optimization: in search of a sound model (Выделение сообществ на основе оптимизации правдоподобия: в поисках надежной модели). In The World Wide Web Conference, pages 1498–1508, 2019. CORE A*. (Главный соавтор; автор диссертации предложила модель ILFR и провела теоретический анализ.)

6. L. Prokhorenkova, A. Tikhonov, and N. Litvak. Learning clusters through information diffusion (Выделение сообществ на основе распространения информации). In The World Wide Web Conference, pages 3151–3157, 2019. CORE A*. (Главный соавтор; автор диссертации разработала предложенные алгоритмы и дизайн экспериментов.)

7. M. Gösgens, A. Tikhonov, and L. Prokhorenkova. Systematic analysis of cluster similarity indices: How to validate validation measures (Систематический анализ индексов похожести кластеризаций: как валидировать метрики валидации). In International Conference on Machine Learning, pages 3799–3808. PMLR, 2021. CORE A*. (Главный соавтор; автор диссертации сформулировала проблему и курировала исследование.)

8. L. Ostroumova Prokhorenkova, P. Prokhorenkov, E. Samosvat, and P. Serdyukov. Publication date prediction through reverse engineering of the web (Предсказание дат публикации страниц на основе оптимизации правдоподобия ссылочной структуры). In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, pages 123–132, 2016. CORE A*. (Главный соавтор; основная идея была предложена в неразрывном сотрудничестве с Е. Самосватом; автор диссертации предложила алгоритмические детали, частично провела эксперименты и курировала исследование.)

9. K. Avrachenkov, N. Litvak, L. Ostroumova Prokhorenkova, and E. Suyargulova. Quick detection of high-degree entities in large directed networks (Быстрое обнаружение вершин большой степени в больших направленных сетях). In 2014 IEEE International Conference on Data Mining, pages 20–29. IEEE, 2014. CORE A*. (Главный соавтор; автор диссертации предложила алгоритм, провела теоретический анализ совместно с Н. Литвак и провела эксперименты.)

10. L. Prokhorenkova and A. Shekhovtsov. Graph-based nearest neighbor search: from practice to theory (Поиск ближайших соседей на основе графов близости: от практики к теории). In International Conference

on Machine Learning, pages 7803–7813. PMLR, 2020. CORE A*. (Главный соавтор; автор диссертации сформулировала проблему, курировала исследование и доказала результаты Разделов 4.1 и 4.2 — анализ жадного поиска и все вспомогательные результаты.)

**Публикации стандартного уровня**

1. L. Ostroumova, A. Ryabchenko, and E. Samosvat. Generalized preferential attachment: tunable power-law degree distribution and clustering coefficient (Обобщенное предпочтительное присоединение: настраиваемые распределение степеней и коэффициент кластеризации). In International Workshop on Algorithms and Models for the Web-Graph, pages 185–202. Springer, 2013. Scopus. (Автор диссертации доказала Теорему 3.)

2. L. Ostroumova Prokhorenkova. Global clustering coefficient in scale-free weighted and unweighted networks (Глобальный коэффициент кластеризации в графах со степенным распределением степеней вершин). Internet Mathematics, 12(1-2):54–67, 2016. Scopus.

3. L. Ostroumova Prokhorenkova, P. Prałat, and A. Raigorodskii. Modularity of complex networks models (Модулярность в моделях сложных сетей). Internet Mathematics, 2017. Web of Science. (Автор диссертации доказала результаты Раздела 4.)

4. A. Dorodnykh, L. Ostroumova Prokhorenkova, and E. Samosvat. Preferential placement for community structure formation (Принцип предпочтительного размещения для формирования реалистичной кластерной структуры). In International Workshop on Algorithms and Models for the Web-Graph, pages 75–89. Springer, 2017. Scopus, Web of Science. (Главный соавтор; автор диссертации формализовала модель и доказала теоретические результаты.)

5. D. Lefortier, L. Ostroumova, and E. Samosvat. Evolution of the Media Web (Эволюция медиа-веба). In International Workshop on Algorithms and Models for the Web-Graph, pages 80–92. Springer, 2013. Scopus. (Автор диссертации, в неразрывном сотрудничестве с Е. Самосватом, теоретически проанализировала предложенную модель.)

**Приглашенные доклады на международных конференциях**

1. Август 2014 – "Crawling of new web pages", 8th Russian Summer School in Information Retrieval, Нижний Новгород, Россия.
2. Июнь 2017 – "Some general results on preferential attachment and clustering coefficient", 7th International Conference on Network Analysis, Нижний Новгород, Россия.
3. Сентябрь 2020 – "Some applications of graphs and probability theory to machine learning", 17th Workshop on Algorithms and Models for the Web Graph, Онлайн.

**Доклады на международных конференциях**

1. Ноябрь 2012 – "Generalized preferential attachment", Workshop on Internet Topology and Economics, Атланта, США.
2. Июнь 2013 – "Preferential attachment models and their generalizations", Franco-Russian workshop on Algorithms, complexity and applications, Москва, Россия.
3. Август 2013 – "Recency-based preferential attachment models", International Conference on Random Structures and Algorithms, Познань, Польша.
4. Сентябрь 2013 – "Recency-based preferential attachment models", Palanga Conference in Combinatorics and Number Theory, Паланга, Литва.
5. Октябрь 2013 – "Timely crawling of high-quality ephemeral new content", International conference on Machine learning and Very Large Data Sets, Москва, Россия.
6. Октябрь 2013 – "Generalized preferential attachment", Workshop on Random Graphs and their Applications, Москва, Россия.
7. Октябрь 2013 – "Timely crawling of high-quality ephemeral new content", ACM International Conference on Information and Knowledge Management, Сан-Франциско, США.

8. Декабрь 2013 – “Generalized preferential attachment”, Workshop on Algorithms and Models for the Web Graph, Кембридж, США.

9. Декабрь 2013 – “Evolution of the Media Web”, Workshop on Algorithms and Models for the Web Graph, Кембридж, США.

10. Апрель 2014 – “Crawling Policies Based on Web Page Popularity Prediction”, 36th European Conference on Information Retrieval, Амстердам, Нидерланды.

11. Июль 2014 – “Model of the Media Web and its application to crawling of ephemeral web pages”, Talk at the 11th International Vilnius Conference on Probability Theory and Mathematical Statistics, Вильнюс, Литва.

12. Июль 2014 – “Recency-based preferential attachment models”, International Conference Sum(m)it240, Будапешт, Венгрия.

13. Декабрь 2014 – “Quick Detection of High-degree Entities in Large Directed Networks”, IEEE International Conference on Data Mining, Шэньчжэнь, Китай.

14. Декабрь 2014 – “Global clustering coefficient in scale-free networks”, 11th Workshop on Algorithms and Models for the Web Graph, Пекин, Китай.

15. Май 2015 – “Global clustering coefficient in scale-free networks”, 5th International Conference on Network Analysis, Нижний Новгород, Россия.

16. Апрель 2016 – “Global clustering coefficient in scale-free networks”, Workshop Critical and collective effects in graphs and networks, Москва, Россия.

17. Май 2017 – “Modularity of complex networks models”, Second workshop Critical and collective effects in graphs and networks, Москва, Россия.

18. Июнь 2017 – “Preferential placement for community structure formation”, 14th Workshop on Algorithms and Models for the Web Graph, Торонто, Канада.

19. Июль 2017 – “Modularity of random graph models”, 39th Conference on Stochastic Processes and their Applications, Москва, Россия.

20. Август 2017 – “Modularity of random graph models”, 18th International Conference on Random Structures and Algorithms, Гнезно, Польша.

21. Май 2018 – “Community detection through likelihood optimization: in search of a sound model”, Workshop on graphs, networks, and their applications, Москва, Россия.

22. Май 2019 – “Learning clusters through information diffusion”, Conference on graphs, networks, and their applications, Москва, Россия.

23. Май 2019 – “Community detection through likelihood optimization: in search of a sound model”, The Web Conference, Сан-Франциско, США.

24. Май 2019 – “Learning clusters through information diffusion”, The Web Conference, Сан-Франциско, США.

25. Июль 2019 – “Using synthetic networks for parameter tuning in community detection”, 16th Workshop on Algorithms and Models for the Web Graph, Брисбен, Австралия.

26. Июль 2020 – “Graph-based nearest neighbor search: from practice to theory”, 37th International Conference on Machine Learning (ICML), Онлайн.

27. Июль 2021 – “Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures”, 38th International Conference on Machine Learning (ICML), Онлайн.

## Анализ моделей сложных сетей

В настоящее время анализ сложных сетей является активной областью исследований. В частности, предложено большое количество моделей случайных графов, которые обладают некоторыми свойствами сетей реального мира. Изучение таких моделей и их свойств чрезвычайно важно, поскольку это помогает понять фундаментальные принципы,

лежащие в основе формирования сложных сетей, предсказать будущее поведение сетей и построить эффективные алгоритмы для их анализа.

## Обобщенное предпочтительное присоединение

Результаты этого раздела основаны на статьях [20, 21, 35, 37].

Наиболее широко изученным свойством сложных сетей является их распределение степеней вершин. Для большинства проанализированных сетей реального мира доля вершин степени $d$ убывает (приблизительно) как $d^{-\gamma-1}$, обычно с $1 < \gamma < 2$ [9]. Таким образом, доля вершин степени больше $d$ (*кумулятивное* распределение) как правило убывает как $d^{-\gamma}$. Наиболее известным подходом к моделированию сложных сетей со степенным распределением степеней вершин является принцип *предпочтительного присоединения* [6]. Основная идея этого подхода состоит в том, что вершины добавляются по очереди, и каждая новая вершина соединяется с несколькими существующими вершинами с вероятностями, пропорциональными их степеням. Многие существующие модели сложных сетей основаны на идее предпочтительного присоединения: модель Боллобаша–Риордана [10], Бакли–Остхуса [11], Холма–Кима [17], RAN [48] и другие. Мы предлагаем общий подход, который позволяет получать строгие теоретические результаты одновременно для всех таких моделей.

### PA-класс моделей

Пусть $G_m^n$ $(n \geqslant n_0)$ — граф с $n$ вершинами $\{1, \ldots, n\}$ и $mn$ ребрами, полученный следующим образом. Стартуем в момент времени $n_0$ с произвольного графа $G_m^{n_0}$ с $n_0$ вершинами и $mn_0$ ребрами. На шаге $n+1$ $(n \geq n_0)$ получаем граф $G_m^{n+1}$ из $G_m^n$ добавлением вершины $n+1$ и $m$ ребер, соединяющих эту вершину с $m$ вершинами из $\{1, \ldots, n, n+1\}$. Обозначим за $d_v^n$ степень вершины $v$ в $G_m^n$. Если для некоторых констант $A$ и $B$ выполнены следующие условия

$$\mathrm{P}\left(d_v^{n+1} = d_v^n \mid G_m^n\right) = 1 - A\frac{d_v^n}{n} - B\frac{1}{n} + O\left(\frac{(d_v^n)^2}{n^2}\right),\ 1 \leqslant v \leqslant n\ , \qquad (1)$$

$$\mathrm{P}\left(d_v^{n+1} = d_v^n + 1 \mid G_m^n\right) = A\frac{d_v^n}{n} + B\frac{1}{n} + O\left(\frac{\left(d_v^n\right)^2}{n^2}\right),\ 1 \leqslant v \leqslant n\ , \qquad (2)$$

$$\mathrm{P}\left(d_v^{n+1} = d_v^n + j \mid G_m^n\right) = O\left(\frac{\left(d_v^n\right)^2}{n^2}\right),\ 2 \leqslant j \leqslant m,\ 1 \leqslant v \leqslant n\ , \qquad (3)$$

$$\mathrm{P}(d_{n+1}^{n+1} = m + j) = O\left(\frac{1}{n}\right),\ 1 \leqslant j \leqslant m\ , \qquad (4)$$

то мы говорим, что полученная модель $G_m^n$ принадлежит PA-классу (от "preferential attachment"). При этом должно быть выполнено $2mA + B = m$ и $0 \leqslant A \leqslant 1$.

Заметим, что даже для фиксированных параметров $A$ и $m$ процедура построения графа не определена полностью, поскольку мы не задали совместное распределение концов добавленных ребер. Поэтому существует целый ряд моделей, обладающих очень разными свойствами и удовлетворяющих условиям (1)–(4). Например, модели Боллобаша–Риордана [10], Холма–Кима [17] и RAN [48] принадлежат PA-классу с $A = 1/2$ и $B = 0$. Модель Бакли–Остухса [11] тоже принадлежит PA-классу с $A = \frac{1}{2+\beta}$ и $B = \frac{m\beta}{2+\beta}$.

Оказывается, некоторые строгие результаты могут быть доказаны для всего PA-класса, не указывая конкретную модель. Такие утверждения обобщают ряд предыдущих теоретических результатов, полученных для каждой модели независимо.

#### Распределение степеней в PA-классе

Обозначим за $N_n(d)$ количество вершин степени $d$ в $G_m^n$. Доказан следующий результат про математическое ожидание случайной величины $N_n(d)$.

**Теорема 1.** *Для $d \geqslant m$ имеем $\mathbb{E}N_n(d) = c(m,d)\left(n + O\left(d^{2+\frac{1}{A}}\right)\right)$, где*

$$c(m,d) = \frac{\Gamma\left(d + \frac{B}{A}\right)\Gamma\left(m + \frac{B+1}{A}\right)}{A\Gamma\left(d + \frac{B+A+1}{A}\right)\Gamma\left(m + \frac{B}{A}\right)} \overset{d\to\infty}{\sim} \frac{\Gamma\left(m + \frac{B+1}{A}\right) d^{-1-\frac{1}{A}}}{A\Gamma\left(m + \frac{B}{A}\right)}$$

*и $\Gamma(x)$ — гамма-функция.*

Мы также показываем, что количество вершин степени $d$ сконцентрировано около своего математического ожидания.

**Теорема 2.** *Для любой модели из PA-класса и для любого $\delta > 0$ существует функция $\varphi(n) = o(1)$ такая, что*

$$\lim_{n\to\infty} \mathrm{P}\left(\exists\, d \leqslant n^{\frac{A-\delta}{4A+2}} : |N_n(d) - \mathbb{E}N_n(d)| \geqslant \varphi(n)\,\mathbb{E}N_n(d)\right) = 0\ .$$

Из Теорем 1 и 2 следует, что распределение степеней в $G_m^n$ подчиняется (асимптотически) степенному закону с параметром $1+\frac{1}{A}$. Таким образом, *кумулятивное* распределение степеней подчиняется степенному закону с параметром $\gamma = \frac{1}{A}$.

#### Коэффициент кластеризации в PA-классе

Другой важной характеристикой сложных сетей является коэффициент кластеризации — мера, отражающая наличие кластеров (плотно взаимосвязанных групп вершин) в графе. В литературе можно найти несколько определений коэффициента кластеризации, наиболее распространенные — *глобальный* и *средний локальный* коэффициенты. Глобальный коэффициент кластеризации $C_1(G)$ — это отношение трехкратного числа треугольников к числу пар смежных ребер в $G$. Средний локальный коэффициент кластеризации определяется следующим образом: $C_2(G) = \frac{1}{n}\sum_{i=1}^{n} C(i)$, где $C(i) = \frac{T^i}{P_2^i}$ — локальный коэффициент кластеризации для вершины $i$, $T^i$ — количество ребер между соседями вершины $i$, а $P_2^i$ — количество пар соседей. Считается, что для многих сетей реального мира и средний локальный и глобальный коэффициенты кластеризации стремятся к ненулевому пределу по мере роста сети. Таким образом, возникает естественный вопрос: можем ли мы что-то сказать о коэффициенте кластеризации во всем PA-классе?

**T-подкласс** Модели из PA-класса могут иметь очень разные коэффициенты кластеризации даже для фиксированных параметров $A$ и $m$. Поэтому, чтобы иметь возможность анализировать поведение коэффициентов кластеризации, нужно ввести дополнительные ограничения. Для этого определим *T-подкласс* моделей.

Чтобы принадлежать к T-подклассу, модель должна удовлетворять следующему свойству, в дополнение к (1)–(4):

$$\mathrm{P}\left(d_i^{n+1} = d_i^n + 1, d_j^{n+1} = d_j^n + 1 \mid G_m^n\right) = e_{ij}\frac{D}{mn} + O\left(\frac{d_i^n d_j^n}{n^2}\right) . \quad (5)$$

Здесь $e_{ij}$ — количество ребер между вершинами $i$ и $j$ в $G_m^n$, а $D$ — неотрицательная константа. Стоит заметить, что это условие все еще не полностью определяет зависимости между добавленными ребрами. Нетрудно показать, что модели Боллобаша–Риордана и Бакли-Остхуса принадлежат T-подклассу с $D = 0$, модель Холма-Кима с $D = P_t \cdot (m-1)$, а модель RAN с $D = 3$.

**Глобальный коэффициент кластеризации в T-подклассе** Сначала проанализируем поведение глобального коэффициента кластеризации $C_1(G_m^n)$. Справедлива следующая теорема.

**Теорема 3.** *Пусть $G_m^n$ принадлежит T-подклассу с $D > 0$. Для любого $\varepsilon > 0$ с вероятностью $1 - o(1)$ верно следующее:*

*(1) Если $2A < 1$, то $\frac{6(1-2A)D-\varepsilon}{m(4(A+B)+m-1)} \leqslant C_1(G_m^n) \leqslant \frac{6(1-2A)D+\varepsilon}{m(4(A+B)+m-1)}$ ;*

*(2) Если $2A = 1$, то $\frac{6D-\varepsilon}{m(4(A+B)+m-1)\log n} \leqslant C_1(G_m^n) \leqslant \frac{6D+\varepsilon}{m(4(A+B)+m-1)\log n}$ ;*

*(3) Если $2A > 1$, то $n^{1-2A-\varepsilon} \leqslant C_1(G_m^n) \leqslant n^{1-2A+\varepsilon}$ .*

Заметим, что в некоторых случаях ($2A \geqslant 1$, то есть $\gamma \leqslant 2$) глобальный коэффициент кластеризации $C_1(G_m^n)$ стремится к нулю (для любого $D$) с ростом числа вершин. Обобщение этого результата на произвольные графы со степенным распределением степеней вершин будет обсуждаться ниже.

**Локальный коэффициент кластеризации в T-подклассе** Далее проанализируем поведение локального коэффициента кластеризации. Во-первых, несложно показать, что $C_2(G_m^n)$ не стремится к нулю, если выполняется условие (5) с $D > 0$.

Однако, чтобы проанализировать локальный коэффициент кластеризации более детально, рассмотрим функцию $C_2(d)$ — локальный коэффициент кластеризации для вершин степени $d$. Известно, что в реальных

сетях $C_2(d)$ обычно убывает как $d^{-\psi}$ с некоторым параметром $\psi > 0$. Для некоторых сетей $C_2(d)$ ведет себя как $d^{-1}$ [26].

Оказывается, во *всех* моделях T-подкласса локальный коэффициент кластеризации $C_2(d)$ асимптотически ведет себя как $\frac{2D}{Am} \cdot d^{-1}$. Выполнен следующий результат.

**Теорема 4.** *Пусть $G_m^n$ принадлежит T-подклассу PA-класса. Тогда для любого $\delta > 0$ существует функция $\varphi(n) = o(1)$ такая, что*

*(1) если $2A \leqslant 1$:* $\lim_{n\to\infty} \mathrm{P}\left(\exists\, d \leqslant n^{\frac{A-\delta}{4A+2}} : |C_2(d) - F(d)| \geqslant \frac{\varphi(n)}{d}\right) = 0,$

*(2) если $2A > 1$:* $\lim_{n\to\infty} \mathrm{P}\left(\exists\, d \leqslant n^{\frac{A(3-4A)-\delta}{4A+2}} : |C_2(d) - F(d)| \geqslant \frac{\varphi(n)}{d}\right) = 0,$

*где* $F(d) = \frac{2D}{d\,(d-1)\,m}\left(m + \sum_{i=m}^{d-1} \frac{i}{Ai+B}\right) \overset{d\to\infty}{\sim} \frac{2D}{mA} \cdot d^{-1}$.

Таким образом, несмотря на то, что T-подкласс обобщает множество различных моделей, можно проанализировать локальный коэффициент кластеризации для всего класса. Оказывается, $C_2(d)$ асимптотически убывает как $\frac{2D}{Am} \cdot d^{-1}$. В частности, этот результат означает, что нельзя изменить показатель $-1$, изменяя параметры $A, D$ и $m$. Другими словами, модели предпочтительного присоединения не позволяют моделировать $C(d) \propto d^{-\psi}$ с $\psi \neq 1$.

#### Ассортативность

Рассмотрим еще одну важную характеристику сложных сетей: *коэффициент ассортативности*, который был впервые введен Ньюманом [32] как коэффициент корреляции Пирсона для степеней смежных вершин. Однако известно, что этот коэффициент имеет определенные недостатки при применении к сетям со степенным распределением степеней вершин [28]. Таким образом, более информативной характеристикой является функция $d_{nn}(d)$ — средняя степень соседа вершин степени $d$. Граф называется ассортативным, если $d_{nn}(d)$ является возрастающей функцией $d$, и дисассортативным, когда $d_{nn}(d)$ является убывающей функцией $d$.

Было эмпирически показано, что в ряде сетей реального мира $d_{nn}(d)$ ведет себя как $d^{\nu}$ для некоторого значения $\nu$, которое может быть поло-

жительным (ассортативные сети) или отрицательными (дисассортативные сети) [9]. Ассортативность имеет большое количество применений; например, ее можно использовать в эпидемиологии. Социальные сети обычно ассортативны, поэтому болезни, затрагивающие людей с большим количеством социальных связей, скорее всего распространятся на другие узлы с высокой степенью. С другой стороны, биологические сети обычно дисассортативны; поэтому стратегии вакцинации, нацеленные на элементы с большой степенью, могут быстро уничтожить эпидемию.

Мы проанализировали $d_{nn}(d)$ во всем T-подклассе моделей для $\gamma > 3$ (случай конечной дисперсии). Мы доказали, что математическое ожидание $d_{nn}(d)$ асимптотически ведет себя как $\log(d)$ (с точностью до константного множителя). Однако это приближение работает достаточно хорошо только для больших значений $d$, а для $d < 10^4$ мы эмпирически наблюдаем другое поведение, которое может выглядеть как $d^\nu$ для некоторого $\nu > 0$. Сформулируем основной результат, а детали можно найти в работе [20].

**Теорема 5.** *Пусть $G_m^n$ принадлежит T-подклассу PA-класса с $A < \frac{1}{2}$. Тогда для любого $\varepsilon > 0$ и для любого $d = d(n) \geqslant m$ имеем*

$$\mathbb{E}d_{nn}(d) = F(d)\left(1 + O\left(\frac{n^{2A+\varepsilon}d^{2+\frac{1}{A}}}{n} + \frac{d^{2+\frac{1}{A}}\log n}{\sqrt{n}}\right)\right),$$

*где $F(d) \overset{d\to\infty}{\sim} \frac{Am+B}{A} \cdot \log(d)$ (точное выражение для $F(d)$ можно найти в работе [20]).*

Согласно Теореме 5, все графы T-подкласса с $A < \frac{1}{2}$ являются ассортативными. Однако $\mathbb{E}d_{nn}(d)$ увеличивается медленно (как $\log(d)$), в отличие от $d^\nu$ в реальных сетях.

## Глобальный коэффициент кластеризации в графах со степенным распределением степеней вершин

Результаты этого раздела основаны на статьях [36, 37].

Хотя распределение степеней вершин в моделях предпочтительного присоединения соответствует реальным сетям, для коэффициента кластеризации это не совсем так. Действительно, для большинства сетей

реального мира параметр $\gamma$ их кумулятивного распределения степеней вершин принадлежит интервалу $(1, 2)$. Однако, как обсуждалось выше (Теорема 3), если $\gamma < 2$ в любой модели PA-класса, то глобальный коэффициент кластеризации убывает к нулю с ростом графа, что не соответствует наблюдениям на реальных сетях. Основная причина такого поведения заключается в том, что количество ребер, добавляемых на каждом шаге, является постоянным. Следовательно, количество треугольников может расти не быстрее чем линейно с числом вершин $n$, в то время как число пар смежных ребер растет как $n^{2/\gamma}$.

Мы доказываем, что модель со степенным распределением степеней вершин с $\gamma < 2$ и с асимптотически постоянным глобальным коэффициентом кластеризации не может существовать. Для этого рассмотрим последовательность графов $G_n$ ($n$ — количество вершин) со степенями, выбранными независимо из *правильно меняющегося распределения* $F$ с параметром $\gamma$ кумулятивного распределения. Правильно меняющиеся распределения — это широкий класс распределений с тяжелым хвостом, обобщающий степенные распределения. Мы предполагаем, что для заданных степеней вершин граф может быть построен произвольным образом. Мы доказываем, что если графы (без петель и кратных ребер) имеют правильно меняющееся распределение степеней вершин с бесконечной дисперсией ($1 < \gamma < 2$), то глобальный коэффициент кластеризации для любой такой последовательности графов стремится к нулю с высокой вероятностью. Важно отметить, что здесь не предполагается никакой модели случайного графа.

**Теорема 6.** *Для любого $\varepsilon > 0$ и любого $\alpha$ такого, что $0 < \alpha < \frac{1}{\gamma+1}$, с вероятностью $1 - O(n^{-\alpha})$ глобальный коэффициент кластеризации графа $G_n$ удовлетворяет следующему неравенству:*

$$C_1(G_n) \leqslant n^{-\frac{(2-\gamma)}{\gamma(\gamma+1)}+\varepsilon}.$$

Таким образом, при заданных ограничениях на распределение степеней вершин, $C_1(G_n)$ убывает к нулю с ростом $n$. С другой стороны, во многих наблюдаемых сетях глобальный коэффициент кластеризации имеет достаточно большие значения [33]. Еще более интересным является тот факт, что существуют модели сложных сетей с асимптотически

степенным распределением степеней вершин с бесконечной дисперсией и с не убывающим к нулю глобальным коэффициентом кластеризации. Например, такие результаты были получены в работе [8]. Это кажущееся противоречие можно объяснить нашей формализацией степенного распределения. Действительно, мы предполагаем, что степени выбираются независимо из правильно меняющегося распределения. Предположение о независимости является довольно сильным и может не выполняться для некоторых моделей случайных графов. В нашем анализе мы используем это предположение, чтобы показать, что наибольшие степени в графе достаточно велики. Без предположения о независимости наибольшие степени могут быть меньше, в таком случае результат Теоремы 6 неприменим.

Формально наблюдения из [33] не противоречат Теореме 6. Существует несколько возможных объяснений. Во-первых, большие значения глобального коэффициента кластеризации обычно экспериментально наблюдаются в небольших сетях. Во-вторых, для сетей со степенным распределением степеней вершин наблюдаемый глобальный коэффициент кластеризации обычно меньше, чем средний локальный коэффициент кластеризации, что согласуется с теорией. Наконец, наши результаты могут быть применены только к сетям с правильно меняющимися распределениями степеней. Если сеть имеет, например, степенное распределение степеней вершин с экспоненциальным отсечением, то наши результаты уже не могут быть применены.

Помимо верхней оценки на глобальный коэффициент кластеризации, мы также предлагаем алгоритм, позволяющий построить граф с почти максимальным (с точностью до множителя $n^{o(1)}$) коэффициентом кластеризации [36].

### Анализ модулярности

Результаты этого раздела основаны на статье [38].

Коэффициент кластеризации является базовой характеристикой для анализа наличия в сети кластерной структуры. Более продвинутым показателем является *модулярность* [34]. Модулярность используется одновременно как численная характеристика наличия структуры сооб-

ществ в графе, как способ определить что такое сообщества и как функция качества алгоритмов выделения сообществ. Многие алгоритмы обнаружения сообществ основаны на поиске разбиений с высокой модулярностью.

Основная идея модулярности состоит в том, чтобы сравнить фактическую плотность ребер внутри сообществ с плотностью, которую можно было бы ожидать, если бы вершины были соединены случайным образом, независимо от структуры сообществ. Формально, для заданного разбиения $\mathcal{A} = \{A_1, \ldots, A_k\}$ множества вершин $V(G)$, пусть

$$q_{\mathcal{A}} = \sum_{A \in \mathcal{A}} \left( \frac{e(A)}{|E(G)|} - \frac{(\sum_{v \in A} \deg(v))^2}{4|E(G)|^2} \right), \tag{6}$$

где $E(G)$ — количество ребер в $G$, $e(A) = |\{uv \in E(G) : u, v \in A\}|$ — количество ребер в подграфе, индуцированном множеством $A$, а $\deg(v)$ — степень вершины $v$.

*Модулярность* графа $G$ определяется как

$$q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G).$$

Если $q^*(G)$ близка к 1 (максимум), то мы имеем явную структуру сообществ. И наоборот, если $q^*(G)$ близка к нулю, мы имеем граф без структуры сообществ.

Мы теоретически анализируем модулярность в случайных $d$-регулярных графах, в графах с ограниченной средней степенью, в модели предпочтительного присоединения [10] и в модели пространственного предпочтительного присоединения [1].

**Случайные $d$-регулярные графы** Рассмотрим вероятностное пространство *случайных $d$-регулярных графов* с равномерным распределением. Это пространство обозначим $\mathcal{G}_{n,d}$, а асимптотики будут приведены для $n \to \infty$ при фиксированном $d \geqslant 2$ и четном $n$ если $d$ нечетно. Сначала мы получаем численную верхнюю оценку на модулярность $\mathcal{G}_{n,d}$. Мы также доказываем следующую более слабую, но явную оценку.

**Теорема 7.** *Пусть $d \in \mathbb{N} \setminus \{1, 2\}$ и $\varepsilon > 0$ — произвольно малая константа. Тогда, асимптотически почти наверное (а.п.н.), $q^*(\mathcal{G}_{n,d}) \leqslant \frac{2}{\sqrt{d}}$.*

**Графы с ограниченной средней степенью** Для графов с ограниченной средней степенью справедлива следующая теорема.

**Теорема 8.** *Пусть $\{G_n\}$ — последовательность графов, в которой $G_n$ — связный граф на $n$ вершинах со средней степенью $\frac{2|E(G_n)|}{n} \leqslant D$ для некоторого фиксированного $D$ и максимальной степенью $\Delta = \Delta(G_n) = o(n)$. Тогда $q^*(G_n) \geqslant \frac{2}{D} - O\left(\sqrt{\frac{\Delta}{n}}\right) = \frac{2}{D} - o(1)$.*

**Модель предпочтительного присоединения** Далее мы анализируем модулярность в модели предпочтительного присоединения [10]. Следующая теорема следует из Теоремы 8 и того факта, что $\Delta(G_m^n) = O\left(n^{\frac{1}{2}+2\varepsilon}\right)$ для любого $\varepsilon > 0$ а.п.н.

**Теорема 9.** *Для любого $\varepsilon > 0$ а.п.н. $q^*(G_m^n) \geqslant \frac{1}{m} - O\left(n^{-1/4+\varepsilon}\right) = \frac{1}{m} - o(1)$.*

Эту теорему можно улучшить и получить асимптотически более точную нижнюю оценку. Здесь мы приведем асимптотический результат для $m \to \infty$, а полное утверждение и численные значения для малых $m$ можно найти в работе [38].

**Теорема 10.** *А.п.н. $q^*(G_m^n) = \Omega\left(1/\sqrt{m}\right)$.*

Кроме того, верна следующая верхняя оценка.

**Теорема 11.** *Для любого $\varepsilon > 0$ а.п.н. $q^*(G_2^n) \leqslant \frac{15+\varepsilon}{16}$, а для любого $m \geqslant 3$ а.п.н. $q^*(G_m^n) \leqslant \frac{15}{16}$.*

**Модель пространственного предпочтительного присоединения** Мы также анализируем модулярность в модели *пространственного предпочтительного присоединения* [1]. Эта модель комбинирует предпочтительное присоединение с пространственной структурой через понятие “сфер влияния”, объем которых увеличивается с ростом степеней вершин. Параметрами модели являются число вершин $n$, размерность пространства $m$, вероятность ребер $p \in [0, 1]$ и константы $A_1$, $A_2$, для которых выполнено $0 < A_1 < \frac{1}{p}$ и $A_2 > 0$.

Известно, что такая модель позволяет генерировать сети со степенным распределением степеней вершин и рядом других характеристик реальных сетей. Следующая теорема показывает, что в моделях пространственного предпочтительного присоединения модулярность стремится к

единице с ростом графа, в отличие от $d$-регулярных графов и классической модели предпочтительного присоединения.

**Теорема 12.** *Пусть* $p \in (0,1]$, $A_1, A_2 > 0$ *и* $pA_1 < 1$. *Тогда а.п.н. модулярность графа в модели пространственного предпочтительного присоединения равна* $1 - O\left(n^{\max\{-1/m,-1+pA_1\}/2}\log^{9/2} n\right) = 1 - o(1)$.

### Предпочтительное размещение

Результаты этого раздела основаны на статье [13].

Важным аспектом моделирования сложных сетей является возможность генерировать реалистичные сообщества (кластеры). Несколько эмпирических исследований показали, что структура сообществ различных реальных сетей обладает некоторыми типичными свойствами: например, кумулятивное распределение размеров сообществ подчиняется степенному закону с некоторым параметром [3, 12, 16]. К сожалению, модель предпочтительного присоединения, как и многие другие модели, не позволяет получить кластера с нужными свойствами.

Мы предлагаем процесс *предпочтительного размещения*, который естественным образом генерирует реалистичные кластерные структуры. Мы предполагаем, что вершины вложены в пространство $\mathbb{R}^d$ для некоторого $d \geqslant 1$. Вершины добавляются по очереди, и их положения определяются следующим образом: каждая новая вершина выбирает 'родителя' среди существующих вершин случайно и равномерно, а итоговое положение выбирается случайным образом на некотором расстоянии от родителя. Расстояние до родителя имеет распределение $\Xi$. Мы показываем, что для получения реалистичной кластерной структуры следует в качестве $\Xi$ взять распределение с тяжелым хвостом. В этом случае, согласно описанной выше процедуре, новые вершины будут часто появляться в плотных областях, близких к некоторым ранее добавленным вершинам; однако из-за тяжелого хвоста $\Xi$ время от времени будут происходить выбросы, которые порождают новые кластеры.

Наши эмпирические исследования подтверждают, что если $\Xi$ имеет степенное распределение с подходящим параметром, мы получаем реалистичную кластерную структуру. Мы показываем, что распределение

размеров кластеров (полученных с помощью алгоритма DBSCAN) подчиняется степенному закону. Более того, полученная структура является иерархической, что согласуется с многочисленными наблюдениями на реальных сетях.

Далее мы проводим теоретический анализ, который показывает, почему в предложенной модели естественно ожидать степенное распределение размеров кластеров. Основная трудность состоит в том, что нет формального определения понятия *кластер*. Таким образом, мы можем дать только интуитивное объяснение того, почему предлагаемый алгоритм приводит к степенному распределению размеров кластеров. Для этого мы делаем некоторые сильные предположения, которые позволяют строго доказать нужный результат.

Формально, пусть $F_t(s)$ — количество кластеров размера $s$ на шаге $t$. Мы предполагаем, что кластеры могут только расти, они не могут объединяться или делиться. Пусть на шаге $t+1$ новый кластер появляется с вероятностью $p(t) = \frac{c}{t^\alpha}$, $c > 0$, $0 \leqslant \alpha \leqslant 1$. Наконец, если вершина $t+1$ не создает новый кластер, вероятность присоединения к кластеру $C$ с $|C| = s$ равна $\frac{s}{t}$. Последнее предположение мотивировано тем, что вероятность выбрать родителя из некоторого кластера $C$ с $|C| = s$ равна $\frac{s}{t}$ по определению модели. Эти предположения являются довольно сильными, но они позволяют формально проанализировать поведение $F_t(s)$. А именно, верна следующая теорема (полное утверждение можно найти в работе [13]).

**Теорема 13.** *В предположениях, описанных выше, верно следующее.*

1. *Если* $\alpha = 0$ *и* $0 < c < 1$, *то:* $\mathbb{E}F_n(s) \sim \frac{c\,\Gamma\left(2+\frac{1}{1-c}\right)}{(2-c)} \cdot \frac{n}{s^{1+\frac{1}{1-c}}}$.
2. *Если* $0 < \alpha \leqslant 1$, *то для любого* $\epsilon > 0$*:* $\mathbb{E}F_n(s) \sim \frac{c\,\Gamma(3-\alpha)}{2-\alpha} \cdot \frac{n^{1-\alpha}}{s^{2-\alpha}}$.

Таким образом, если вероятность $p(n)$ образования нового кластера имеет порядок $\frac{1}{n^\alpha}$ с $\alpha > 0$, то распределение размеров кластеров подчиняется степенному закону с параметром $2-\alpha$, растущим с $p(n)$ от 1 до 2; если $p(n) = c$, $0 < c < 1$, то параметр степенного распределения размеров кластеров растет с ростом $c$ от 2 до бесконечности.

Заметим, что идея предпочтительного размещения позволяет генерировать положения вершин в некотором пространстве. Затем, чтобы по-

лучить граф, можно использовать любую из существующих пространственных моделей. В работе [13] мы анализируем несколько возможных вариантов и показываем, что полученные графы действительно обладают желаемыми свойствами, включая реалистичное распределение степеней вершин и степенное распределение размеров кластеров.

### Предпочтительное присоединение с устареванием

Результаты этого раздела основаны на статьях [24, 40].

В работе [24] был предложен новый принцип *предпочтительного присоединения с устареванием*. Этот принцип мотивирован эмпирическим исследованием части Интернета, связанной с медиа-контентом. Мы проанализировали динамику входящих и исходящих ссылок для новостных страниц. В частности, было определено и проанализировано *свойство устаревания*. А именно, обозначим через $e(T)$ долю ребер, соединяющих вершины, разница в возрасте которых превышает $T$. Мы показали, что медиа-страницы, как правило, ссылаются на близкие по дате создания страницы, и $e(T)$ экспоненциально убывает с ростом $T$.

Предложенный принцип предпочтительного присоединения с устареванием позволяет моделировать это свойство. А именно, для каждой новой страницы вероятность добавить ссылку на существующую страницу $p$ пропорциональна *привлекательности* $p$, которая является некоторой функцией от $d(p)$ (текущая степень $p$), $q(p)$ (качество $p$) и $a(p)$ (текущий возраст $p$). Рассматриваются различные функции привлекательности:

$$\mathrm{attr}(p) = q(p)^{\alpha_1} \cdot d(p)^{\alpha_2} \cdot e^{-\frac{a(p)}{\tau} \cdot \alpha_3},$$

где $(\alpha_1, \alpha_2, \alpha_3) \in \{0, 1\}^3$, а $\tau$ отвечает за скорость убывания привлекательности со временем. Заметим, что $\mathrm{attr}(p) = d(p)$ соответствует классическому предпочтительному присоединению, а $\mathrm{attr}(p) = q(p) \cdot d(p)$ соответствует другой известной модели — fitness model.

Чтобы получить свойство устаревания, необходимо включить в функцию привлекательности множитель $e^{-\frac{a(p)}{\tau}}$. Проведенные нами теоретический анализ и компьютерное моделирование показывают, что для получения степенного распределения степеней вершин с реалистичным параметром нужно выбрать функцию привлекательности $\mathrm{attr}(p) = q(p) \cdot$

$e^{-\frac{a(p)}{\tau}}$. Более того, качество $q$ должно иметь степенное распределение. Анализ реальных ссылочных структур показал, что функция привлекательности $\mathrm{attr}(p) = q(p) \cdot e^{-\frac{a(p)}{\tau}}$ действительно дает большее правдоподобие по сравнению с другими моделями.

Предложенный принцип формализован и теоретически проанализирован в работе [40]. Для этого мы фокусируемся на функции привлекательности $q(p) \cdot e^{-\frac{a(p)}{\tau}}$ и формально определяем последовательность случайных графов $\{G_n\}$ следующим образом. Последовательность параметризуется натуральным числом $m$ (исходящая степень новой вершины) и целочисленной функцией $N(n)$. Пусть, кроме того, имеется последовательность независимых одинаково распределенных случайных величин $\zeta_1, \zeta_2, \ldots$, принимающих положительные значения.

Каждый граф $G_n$ строится независимо от других следующим образом. В начале процесса есть две вершины, соединенные ребром (граф $\tilde{G}_2^n$). Первые две вершины имеют качество $q(1) := \zeta_1$ и $q(2) := \zeta_2$. На шаге $t+1$ $(2 \leqslant t \leqslant n-1)$ к графу $\tilde{G}_t^n$ добавляется одна вершина и $m$ ребер. Новая вершина $t+1$ имеет качество $q(t+1) := \zeta_{t+1}$. Новые ребра проводятся независимо и соединяют добавленную вершину с предыдущими. Для каждого ребра вероятность того, что оно будет проведено в вершину $i$ $(1 \leqslant i \leqslant t)$, равна

$$\frac{\mathrm{attr}_t(i)}{\sum_{j=1}^{t} \mathrm{attr}_t(j)}, \text{ где } \mathrm{attr}_t(i) = q(i)\, e^{-\frac{t-i}{N(n)}}.$$

Далее мы предполагаем, что $N = N(n) \to \infty$ при $n \to \infty$.

Предположим, что случайные величины $\zeta_1, \zeta_2, \ldots$ имеют распределение Парето с функцией плотности $f(x) = \frac{\gamma a^\gamma I[x>a]}{x^{\gamma+1}}$, где $\gamma > 1$, $a > 0$. В этом случае математическое ожидание количества вершин степени $d$ в графе $G_n$ $(N_n(d))$ убывает как $d^{-\gamma-1}$.

**Теорема 14.** *Определим константу $\alpha$ следующим образом: если $\gamma > 2$, то $\alpha = 2$; если $1 < \gamma \leqslant 2$, то $\alpha$ удовлетворяет $1 < \alpha < \gamma$. Пусть $d = d(n)$ растет с ростом $n$ и $d = o\left(\min\left\{\left(\frac{n}{N \log N}\right)^{\frac{1}{\gamma+1}},\ N^{\frac{\alpha-1}{\alpha+(\gamma+1)(\alpha+1)}}\right\}\right)$, тогда*

$$\frac{\mathbb{E}[N_n(d)]}{n} = \frac{\gamma}{d^{\gamma+1}} \left(\frac{(\gamma-1)m}{\gamma}\right)^{\gamma} (1 + o(1)) \ .$$

Следующая теорема показывает, что количество вершин степени $d$ сконцентрировано около своего математического ожидания.

**Теорема 15.** *Верно следующее:*

$$\mathrm{P}\left(|N_n(d) - \mathbb{E}[N_n(d)]| > \sqrt{Nn\log n}\right) = O\left(\frac{1}{\log n}\right).$$

Заметим, что для $d = o\left(\left(\frac{n}{N\log n}\right)^{\frac{1}{2(\gamma+1)}}\right)$ мы имеем $\sqrt{Nn\log n} = o\left(\mathbb{E}[N_n(d)]\right)$ и Теорема 15 действительно показывает концентрацию.

Помимо этого, мы доказываем, что функция $e(T)$ ведет себя желаемым образом, то есть экспоненциально убывает с ростом $T$.

**Теорема 16.** *Для любого целого $T$,*

$$\mathbb{E}[e(T)] = e^{-\frac{T}{N}} + O\left(\frac{N}{n}\right),$$

$$\mathrm{P}\left(|e(T) - \mathbb{E}[e(T)]| \geqslant \sqrt{\frac{N\log n}{n}}\right) = O\left(\frac{1}{\log n}\right).$$

## Выделение сообществ

Как обсуждалось выше, кластерная структура является одним из наиболее важных свойств реальных графов. Она характеризуется наличием сильно взаимосвязанных групп вершин, относительно хорошо отделенных от остальной части графа. В социальных сетях сообщества (кластеры) формируются пользователями со схожими интересами; в сетях цитирования они представляют статьи в определенных областях; в Интернете сообщества могут соответствовать страницам по смежным темам и так далее. Наличие сообществ сильно влияет, например, на продвижение продуктов с помощью вирусного маркетинга, распространение инфекционных заболеваний, компьютерных вирусов, информации и так далее. Умение находить сообщества важно для ряда приложений: кластеры в графах цитирования полезны для поиска похожих научных работ, обнаружение пользователей со схожими интересами важно для целевой рекламы, кластеризация также может использоваться для сжатия и визуализации сети. В этом разделе мы обсуждаем различные аспекты задачи *выделения сообществ* [14].

**Выделение сообществ на основе максимизации правдоподобия**

Результаты этого раздела основаны на статье [42].

Среди многих существующих алгоритмов выделения сообществ следует отметить методы, основанные на статистическом выводе. В таких методах предполагается некоторая базовая (нулевая) модель случайного графа, наблюдение — это данная структура графа (матрица смежности), а скрытые переменные — параметры модели, которые включают в себя разбиение вершин на сообщества. Такие методы играют важную роль, поскольку они теоретически обоснованы и согласованы: например, было доказано, что если метод максимального правдоподобия применяется к сетям, созданным на основе стохастической блочной модели (stochastic block model), то он возвращает правильное разбиение на сообщества, при условии что степени вершин достаточно большие [7]. Кроме того, правдоподобие может быть использовано, чтобы определить само понятие *сообщества*.

Выбор правильной нулевой модели имеет важное значение для алгоритмов статистического вывода, поскольку он сильно влияет на качество. Наиболее часто используемая модель называется *planted partition model* (PPM). В этой модели вершины разделены на $k$ кластеров, и для каждой пары вершин $i, j$ мы проводим ребро между ними с вероятностью $p_{in}$, если они принадлежат одному кластеру, и $p_{out}$ в противном случае, $p_{in} > p_{out}$.

Однако в модели PPM нельзя смоделировать реалистичное распределение степеней вершин. Для решения этой проблемы были предложены *стохастическая блочная модель с поправкой на степени* (degree-corrected SBM) [18] и ее упрощенный вариант *degree-corrected planted partition model* (DCPPM) [31]. В DCPPM вершины разбиваются на $k$ кластеров, а ребра проводятся независимо случайным образом. Ожидаемое количество ребер между вершинами $i$ и $j$ равно $\frac{d(i)d(j)}{2m}p_{in}$, если они принадлежат одному кластеру, или $\frac{d(i)d(j)}{2m}p_{out}$ в противном случае. Здесь $d(i)$ — желаемая степень вершины $i$, а $m$ - общее количество ребер. Однако можно показать, что в DCPPM при любых (осмысленных) $p_{in}$ и $p_{out}$ ожидаемая степень вершины $i$ не равна $d(i)$. Руководствуясь этим наблюдением, а также моделью случайного графа LFR [23], мы предлагаем

однопараметрическую модель, у которой нет вышеуказанной проблемы.

Неспособность DCPPM сохранить заданную последовательность степеней вершин вызвана тем фактом, что вероятность внутрикластерного ребра не зависит от размера кластера и, следовательно, ожидаемая доля внутренних ребер *зависит от размера кластера*, к которому принадлежит вершина. Вместо этого, как и в модели LFR [23], мы используем параметр $\mu$, который определяет эту долю и делает ее равной для всех вершин графа. Мы предполагаем, что все ребра независимы и ожидаемое число ребер между двумя вершинами $i$ и $j$ равно $\frac{\mu d(i)d(j)}{2m}$, если $c(i) \neq c(j)$ или $\frac{(1-\mu)d(i)d(j)}{D(C_q)} + \frac{\mu d(i)d(j)}{2m}$, если $c(i) = c(j) = q$. Здесь $\mu$ — параметр, регулирующий долю внутрикластерных ребер ($0 < \mu < 1$), $c(i)$ — метка кластера для вершины $i$, а $D(C)$ — сумма степеней вершин, принадлежащих кластеру $C$. При таком определении модели ожидаемая степень вершины $i$ равна $d(i)$. Более того, модель имеет только один параметр $\mu$ вместо $p_{in}$ и $p_{out}$. Мы называем предложенную модель *Independent LFR* или ILFR.

Мы получаем точную формулу правдоподобия для предлагаемой модели ILFR. Для простоты приведем здесь ее приближение:

$$\log L_{ILFR'}(\mathcal{C}, G, \mu) = m_{in} \log(1-\mu) + m_{out} \log \mu \\ - m_{out} \log 2m - \sum_C \frac{D_{in}(C)}{2} \log D(C) + \sum_i d(i) \log d(i) - m\,, \quad (7)$$

где $m_{in}$ и $m_{out}$ — это количество внутрикластерных и межкластерных ребер, а $D_{in}(C)$ — удвоенное число ребер, индуцированных $C$. Оптимальным значением $\mu$ в соответствии с (7) является $\mu = \frac{m_{out}}{m}$.

Полученная формула (7) позволяет нам применять методы, основанные на максимизации правдоподобия, к модели ILFR. В ряде экспериментов мы сравниваем алгоритмы оптимизации правдоподобия, основанные на трех базовых моделях, рассмотренных выше — PPM, DCPPM и ILFR. Мы также показываем, что предлагаемая модель ILFR наилучшим образом (с точки зрения правдоподобия) подходит для различных сетей реального мира.

### Выделение сообществ на основе каскадов

Результаты этого раздела основаны на статье [43].

Рассмотрим более сложною проблему выделения сообществ в сети, основываясь исключительно на *каскадах*, распространяющихся в этой сети. Каскады могут соответствовать распространению, например, информации или эпидемии. По сравнению с традиционным выделением сообществ, эта задача существенно отличается, поскольку у нас нет информации про структуру графа, а есть только данные наблюдаемых каскадов. Для каждого каскада мы знаем только зараженные элементы и время их заражения.

Формально, пусть мы наблюдаем набор каскадов $\mathcal{C} = \{C_1, \ldots, C_r\}$, которые распространяются по неизвестному неориентированному графу $G = (V, E)$ с $|V| = n$ вершинами и $|E| = m$ ребрами. Каждый каскад $C \in \mathcal{C}$ представляет собой времена активации (заражения) вершин, то есть $C = \{(v_i, t^C_{v_i})\}_{i=1}^{n_C}$, где $v_i$ — вершина, $t^C_{v_i}$ — время ее активации в $C$, а $|C| = n_C$ — размер каскада. При этом информация про источник заражения данной вершины не доступна.

Мы предполагаем, что вершины графа $G$ разбиты на сообщества: $\mathcal{A} = \{A_1, \ldots, A_k\}$, $\cup_{i=1}^k A_i = V$, $A_i \cap A_j = \emptyset$ для $i \neq j$. Мы ожидаем, что плотность ребер внутри сообществ высокая по сравнению с плотностью ребер между сообществами. Наблюдая только набор каскадов $\mathcal{C}$, мы хотим найти разбиение $\mathcal{A}'$, близкое к $\mathcal{A}$.

Мы предлагаем и анализируем два типа подходов к этой проблеме: основанные на максимизации правдоподобия при определенных модельных предположениях и основанные на кластеризации вспомогательных графов.

Подходы, основанные на максимизации правдоподобия, называются ClustOpt и GraphOpt. В ClustOpt мы предполагаем модель каскада, в которой каждая активная вершина может заразить любую незараженную вершину независимо от других, время до заражения распределено экспоненциально. При этом, если незараженная вершина принадлежит к тому же сообществу, то интенсивность заражения равна $\alpha_{in}$, а в противном случае она равна $\alpha_{out}$, $\alpha_{out} < \alpha_{in}$. Эпидемия прекращается в момент времени $T_{max}$. Эта модель эпидемии является упрощенной, по-

скольку распространение зависит только от структуры сообществ, а не от графа $G$. Мы выводим формулу для вероятности каскадов $L(\mathcal{C}, \mathcal{A})$ в этой модели. Эта вероятность зависит от параметров $\alpha_{in}$ и $\alpha_{out}$. Для оптимизации правдоподобия мы предлагаем следующий алгоритм (детали можно найти в [43]).

**АЛГОРИТМ 1:** ClustOpt

1. Строим начальное разбиение $\mathcal{A}_{init}$;
2. Находим $\hat{\alpha}_{in}, \hat{\alpha}_{out} = \arg\max_{\alpha_{in},\alpha_{out}} \log L(\mathcal{C}, \mathcal{A}_{init})$;
3. Для фиксированных $\hat{\alpha}_{in}, \hat{\alpha}_{out}$ находим $\hat{\mathcal{A}} = \arg\max_{\mathcal{A}} \log L(\mathcal{C}, \mathcal{A})$.

GraphOpt основан на более сложной модели эпидемий, которая была предложена в [44]. В этой модели активная вершина заражает своих соседей через экспоненциально распределенное время с интенсивностью $\alpha$. Эпидемия заканчивается в момент времени $T_{max}$. GraphOpt основан на максимизации правдоподобия, при этом граф $G$ является скрытой переменной. Мы предполагаем следующий генеративный вероятностный процесс. Для данного разбиения вершин $\mathcal{A}$ мы строим граф $G$ в соответствии с моделью ILFR, рассмотренной в предыдущем разделе. Затем, используя граф $G$, мы генерируем набор каскадов $\mathcal{C}$ в соответствии с моделью эпидемии $\mathrm{P}(\mathcal{C}|G)$. Мы наблюдаем $\mathcal{C}$, а наша цель — восстановить разбиение $\bar{\mathcal{A}}$, которое максимизирует правдоподобие, то есть $\bar{\mathcal{A}} = \arg\max_{\mathcal{A}} \mathrm{P}(\mathcal{C}|\mathcal{A})$ . Мы предлагаем следующий алгоритм (вывод этого алгоритма и его детали можно найти в [43]).

**АЛГОРИТМ 2:** GraphOpt

1. Выбираем некоторое начальное разбиение $\hat{\mathcal{A}}$ и граф $\hat{G}$;
2. Обновляем $\hat{G}$: $\hat{G} = \arg\max_G \left( \log \mathrm{P}(\mathcal{C}|G) + \log \mathrm{P}(G|\hat{\mathcal{A}}) \right)$;
3. Обновляем $\hat{\mathcal{A}}$: $\hat{\mathcal{A}} = \arg\max_{\mathcal{A}} \log \mathrm{P}(\hat{G}|\mathcal{A})$;
4. Повторяем (2)-(3) до сходимости.

Вторая группа алгоритмов основана на построении вспомогательного графа $\hat{G}$ и последующей кластеризации этого графа (с использованием, например, алгоритма Louvain). Стоит отметить, что граф $\hat{G}$ не должен быть похож на $G$. Но важно, чтобы $\hat{G}$ отражал информацию про разбиение вершин на сообщества. Например, в алгоритме Path мы соединяем все последовательные вершины, которые участвовали в одном каскаде.

А в алгоритме CLIQUE мы соединяем все пары вершин одного каскада взвешенными ребрами (веса зависят от разницы во времени между временами заражения).

Чтобы сравнить предложенные подходы между собой и с ранее существующими, мы провели ряд экспериментов на различных реальных и синтетических датасетах. Оказывается, наиболее стабильное качество достигается с помощью эвристических методов, основанных на кластеризации вспомогательных графов, которые не используют сильных модельных предположений. Эти методы одинаково хорошо работают в разных сетях и для эпидемий разных типов.

### Анализ функций качества в задаче кластеризации

Результаты этого раздела основаны на статье [15].

При разработке и анализе алгоритмов выделения сообществ крайне важно иметь возможность оценивать результаты, то есть измерять качество различных алгоритмов и сравнивать их между собой. Мы показываем, что эта проблема очень важна: существуют десятки мер похожести кластеризаций (*индексов*), они часто не согласованы друг с другом, эти разногласия влияют на то, какие алгоритмы будут выбраны, и это может привести к снижению качества в реальных системах.

Мы предлагаем теоретический подход к этой проблеме: для ряда желаемых свойств мы теоретически проверяем, какие индексы им удовлетворяют. Это позволяет делать осознанный выбор: для конкретного приложения можно сначала выбрать свойства, которые являются желательными для данной задачи, а затем выбрать индексы, удовлетворяющие этим требованиям. Наши результаты позволяют дать практические рекомендации, которые значительно отличаются от того, какие индексы обычно выбираются на практике. Показано, что у наиболее популярных индексов есть ряд недостатков, в то время как некоторые менее известные меры похожести кластеризаций обладают хорошими свойствами.

Далее мы кратко обсудим свойства, которые подробно описаны в работе [15]. За $V(A, B)$ мы обозначим индекс похожести двух разбиений $A$ и $B$ данного набора элементов.

1. Во-первых, значение индекса должно быть интерпретируемым. В

частности, должно быть легко понять, что кластеризация-кандидат максимально похожа на эталонную кластеризацию (то есть совпадает с ней). Формально, мы требуем $V(A, A) = c_{\max}$ и $V(A, B) < c_{\max}$ для всех $A \neq B$. Это свойство называется *максимальная согласованность*.

2. Во многих случаях желательно, чтобы хороший индекс похожести кластеризаций был *симметричным*, то есть $V(A, B) = V(B, A)$ для всех разбиений $A, B$.

3. Вычислительная сложность имеет важнейшее значение для задач кластеризации больших наборов данных: алгоритмы и индексы со сверхлинейным временем могут быть неприемлемыми. Мы говорим, что индекс имеет *линейную сложность*, если сложность его вычисления линейна по количеству элементов.

4. Для некоторых приложений может быть желательна метрическая интерпретация похожести кластеризаций. В частности, если разбиение $A$ похоже на $B$, а $B$ похоже на $C$, тогда $A$ должно быть достаточно близко к $C$. Мы говорим, что индекс удовлетворяет свойству *расстояния*, если его можно линейно преобразовать в метрику.

5. Свойство *монотонности* является наиболее естественным для индексов похожести кластеризаций. Мы требуем, чтобы индекс был монотонным относительно изменений, которые увеличивают сходство кластеризаций. Это свойство формализовано в работе [15].

6. Наконец, определим свойство *constant baseline*, которое является наиболее важным: это свойство менее интуитивно, чем предыдущие, и поэтому может привести к неожиданным последствиям на практике. Неформально, хороший индекс похожести не должен отдавать предпочтение кластеризации $B$ по сравнению с другой кластеризацией $C$ только потому, что $B$ содержит много или мало кластеров. Эту интуицию можно формализовать с помощью случайных разбиений: предположим, что у нас есть эталонная кластеризация $A$ и два случайных разбиения $B$ и $C$. Интуитивно, оба случайных разбиения являются одинаково плохими приближениями $A$. Поэтому

Таблица 1: Свойства индексов похожести кластеризаций

| | Макс. соглас. | Симметричность | Расстояние | Лин. сложность | Монотонность | Const. baseline |
|---|---|---|---|---|---|---|
| **NMI** | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ |
| **$\text{NMI}_{\max}$** | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| **FNMI** | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ |
| **VI** | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ |
| **SMI** | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| **FMeasure** | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ |
| **BCubed** | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ |
| **AMI** | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ |

Таблица 2: Свойства pair-counting индексов похожести кластеризаций

| | Макс. соглас. | Мин. соглас. | Симметричность | Расстояние | Лин. сложность | Монотонность | Сильная монот. | Const. baseline | As. const. baseline | Тип смещения |
|---|---|---|---|---|---|---|---|---|---|---|
| **R** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ⤭ |
| **AR** | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | |
| **J** | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ↘ |
| **W** | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ↘ |
| **D** | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ↘ |
| **CC** | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| **S&S** | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| **CD** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | |

мы требуем, чтобы значение похожести случайного кандидата $B$ на эталонную кластеризацию $A$ имело фиксированное ожидаемое значение $c_{\text{base}}$, которое не зависит от $A$ или размеров кластеров в $B$. Осталось формализовать понятие случайного разбиения, что сделано в работе [15].

Основные полученные результаты приведены в Таблицах 1 и 2, формальные определения индексов и свойств можно найти в работе [15].

## Другие приложения

В этом разделе обсуждаются другие приложения анализа графов. Сначала будет описан алгоритм определения дат публикации вебстраниц, основанный на модели с устареванием, рассмотренной выше. Затем мы обсудим задачу выделения вершин большой степени в сложных сетях. Наконец, будет рассмотрена проблема поиска ближайших соседей и алгоритмы для решения этой задачи, основанные на графах близости.

### Определение дат публикации веб-страниц

Начнем с примера того, как представленная выше *модель предпочтительного присоединения с устареванием* может быть использована для повышения качества определения дат публикации веб-страниц. Результаты этого раздела основаны на статье [39].

Знание дат публикации веб-страниц имеет важное значение, например, для ранжирования документов с учетом их свежести. К сожалению, для значительной части веб-страниц их даты публикации не могут быть достоверно определены. Наиболее распространенным способом определения даты публикации является поиск этой даты в тексте страницы. Однако веб-страницы могут не содержать или содержать несколько дат-кандидатов; эти даты могут быть записаны в разных форматах и для разных часовых поясов. В некоторых случаях датой публикации веб-страницы может считаться дата первого обхода этой страницы поисковым роботом. Однако из-за нехватки ресурсов не все веб-сайты просматриваются достаточно часто, чтобы можно было обнаружить новые страницы сразу после их публикации.

Предлагаемый алгоритм сочетает методы извлечения дат из текста с методами, основанными на анализе *ссылочной структуры* сайтов. На первом этапе извлекаются даты-кандидаты из URL-адреса и HTML-текста веб-страницы и выбирается наиболее вероятная дата публикации из полученных кандидатов. Для некоторых страниц можно выбрать надежные даты, которые будут зафиксированы на протяжении всего дальнейшего процесса; такие даты называются *фиксированными*. Для некоторых других страниц можно извлечь даты-кандидаты, которые менее надежны, и их оценки могут быть улучшены на третьем этапе алгоритма; такие даты называются *предварительными*. Для остальных страниц извлечение дат на основе текста невозможно.

На втором этапе алгоритма выбираются *предварительные* даты для всех оставшихся страниц, то есть страниц без извлеченных дат. Для этого мы используем и сравниваем несколько методов *распространения* дат, в которых мы итеративно распространяем известные даты с датированных страниц на недатированные, используя, например, усреднение по соседям.

Полученные *предварительные* даты могут быть дополнительно улучшены на третьем этапе с помощью предложенного метода *вероятностной оптимизации*, основанного на модели предпочтительного присоединения с устареванием [24]. В работе [24] даты публикации веб-страниц используются для прогнозирования динамики структуры веб-ссылок. Здесь мы выполняем обратную операцию, то есть используем ссылочную структуру для оценки дат публикации веб-страниц. А именно, мы находим такие даты публикации, которые максимизируют вероятность того, что веб-граф, наблюдаемый в реальности, получен из модели предпочтительного присоединения с устареванием.

Конкретная модель, которую мы предполагаем, имеет несколько параметров. Для каждой страницы $p$ у нас есть количество исходящих ссылок $m_p$, качество страницы $q_p$ и время публикации $t_p$. Помимо параметров страниц, у нас также есть скорость угасания привлекательности $\lambda$, вспомогательная константа $c$ ($c > \lambda$) и количество страниц $n$. В начале процесса есть $n$ страниц, каждая страница $p$ имеет время публикации $t_p$ и качество $q_p$. Затем для каждой страницы $p$ мы генерируем $m_p$ исходящих ссылок. Все ссылки генерируются независимо друг от друга. Вероятность того, что страница $r$ будет выбрана в качестве целевой страницы для ссылки со страницы $p$, пропорциональна привлекательности $r$ относительно $p$. Привлекательность является функцией от $q_r$ (качество $r$) и разницы в возрасте $a_{p,r}$ для страниц $p$ и $r$, то есть $a_{p,r} = t_p - t_r$. Стоит заметить, что разница $a_{p,r}$ может быть отрицательной, то есть вероятность наличия ребра между $p$ и $r$ с $t_p < t_r$ не равна нулю. В реальном веб-графе такая ссылка может быть добавлена в момент $t > t_r$, если страница $p$ была обновлена в момент $t$. Функция привлекательности определяется следующим образом:

$$\operatorname{attr}(q_r, a_{p,r}) = \begin{cases} q_r \cdot e^{-\lambda a_{p,r}} \cdot \left(1 - \frac{e^{-c a_{p,r}}}{2}\right), & \text{если } a_{p,r} \geqslant 0\,, \\ q_r \cdot e^{-\lambda a_{p,r}} \cdot \frac{e^{c a_{p,r}}}{2}, & \text{если } a_{p,r} < 0\,. \end{cases} \tag{8}$$

Во-первых, привлекательность $r$ пропорциональна ее качеству. Во-вторых, привлекательность снижается с возрастом $r$, то есть старые страницы менее популярны. Эти два множителя предложены в [24]. Третий множитель — это сигмоида, которая заменяет индикатор $\mathbb{1}_{a_{p,r} \geqslant 0}$

из [24]. Это делается для того, чтобы вероятности ребер стали дифференцируемыми. Сигмоида также позволяет нам избежать вырожденной вероятности, поскольку все вероятности ребер становятся больше нуля.

Мы используем эту модель для оценки дат публикации веб-страниц. А именно, пусть нам дан некоторый ориентированный граф (вершины — это веб-страницы, а ребра — связи между ними), и мы предполагаем, что этот граф построен в соответствии с моделью, описанной выше. Нам даны наблюдаемые значения некоторых параметров (количество исходящих ссылок $m_p$ и некоторые даты публикации (*фиксированные* даты) $t_p$). Мы хотим найти остальные неизвестные значения, максимизируя вероятность того, что наблюдаемый граф построен в соответствии с описанной моделью. Параметры с неизвестными значениями — это скорость угасания привлекательности $\lambda$, константа $c$, качество всех страниц $q_p$ и даты публикации $t_p$ для страниц с *предварительными* датами.

Мы оптимизируем неизвестные параметры с помощью градиентного спуска. Формулы для вероятности и ее производных можно найти в [39].

В работе мы оцениваем предложенный алгоритм на двух наборах данных: датасет, полученный поисковым роботом Яндекса (4 миллиона страниц с 70 хостов), и общедоступный датасет MemeTracker, который состоит из сообщений в блогах и новостных статей (12 миллионов страниц с 250 тысяч хостов).

### Поиск популярных вершин в больших сетях

Результаты этого раздела основаны на статье [5].

Мы решаем проблему быстрого обнаружения объектов (вершин) большой степени в крупных социальных сетях. Объектами могут быть пользователи, группы по интересам, географические местоположения и так далее. Например, задача может состоять в том, чтобы найти список пользователей с большим количеством подписчиков в Twitter или группы с большим количеством участников в Facebook.

Полный перебор позволяет найти топ-$k$ объектов с наибольшей входящей степенью в ориентированном графе $G$ на $N$ вершинах со сложностью $O(N)$. Для очень больших сетей даже такая линейная сложность является неприемлемой. Кроме того, данные социальных сетей, как правило,

доступны только владельцам сетей и могут быть получены другими лицами через API-запросы, а количество API-запросов в единицу времени обычно существенно ограничено.

Формально, пусть $V$ — набор из $N$ объектов, например пользователей, к которым можно получить доступ с помощью API-запросов. Пусть $W$ — набор из $M$ других объектов (возможно, совпадающий с $V$). Рассмотрим двудольный граф $(V, W, E)$, в котором направленное ребро $(v, w) \in E$, $v \in V$ и $w \in W$, соответствует некоторому отношению между $v$ и $w$. Например, в сети Twitter мы можем считать, что $V$ — это набор пользователей, $W = V$, и $(v, w) \in E$ означает, что $v$ подписан $w$ или что $v$ ретвитнул твит $w$. Заметим, что любой ориентированный граф $G = (V, E)$ может быть эквивалентно представлен двудольным графом $(V, V, E)$. Можно также предположить, что $V$ — это набор пользователей, $W$ — набор групп по интересам, а ребро $(v, w)$ означает, что пользователь $v$ состоит в группе $w$. Задача — быстро найти группы с большим числом участников в $W$.

Пусть $n$ — заданный лимит на количество API-запросов. Предлагаемый алгоритм состоит из двух этапов, первый использует $n_1$ запросов, второй — $n_2$, $n_1 + n_2 = n$.

**Первый этап** Выберем случайный набор $A$ из $n_1$ элементов $v_1, \ldots, v_{n_1} \in V$, элементы выбираются независимо. Для каждого элемента в $A$ мы запоминаем его (исходящих) соседей в $W$. На этот шаг тратится $n_1$ API-запросов — каждый запрос выдает нам список соседей заданной вершины. Для каждого $w \in W$ мы вычисляем величину $S[w]$, равную числу элементов в $A$, которые ссылаются на $w$.

**Второй этап** Мы используем $n_2$ API-запросов для получения фактических значений (входящих) степеней $n_2$ элементов из $W$ с наибольшими значениями $S[w]$. Мы ожидаем, что элементы с наибольшими степенями в $W$ с большой вероятностью будут среди $n_2$ элементов с наибольшими значениями $S[w]$, если $n_2$ достаточно большое.

В работе [5] мы экспериментально показываем, что предложенный алгоритм существенно превосходит другие известные методы. Например, требуется всего тысяча запросов к API, чтобы найти топ-100 самых популярных пользователей с точностью более 90% в социальной сети Twitter

с примерно миллиардом зарегистрированных пользователей (на момент публикации). Далее, мы анализируем предложенный алгоритм, используя теорию экстремальных значений, и получаем достаточно точный прогноз качества. Мы показываем, что количество запросов к API для поиска $k$ наиболее популярных объектов сублинейно по количеству объектов. Более того, мы формально показываем, что тяжелый хвост распределения степеней вершин является важной составляющей эффективности алгоритма.

### Анализ алгоритмов поиска ближайших соседей, основанных на графах близости

В этом разделе обсуждается еще одно важное практическое применение графового анализа — в задаче поиска ближайших соседей. Результаты основаны на статье [41].

Многие методы машинного обучения, распознавания образов, теории кодирования и других областей исследований основаны на поиске ближайших соседей. В частности, метод $k$-ближайших соседей входит в топ-10 алгоритмов анализа данных [46]. Поскольку современные датасеты как правило имеют очень большие размеры (как по количеству элементов $n$, так и по размерности $d$), важно снизить вычислительную сложность алгоритмов поиска ближайших соседей. Задача заключается в том, чтобы предварительно обработать датасет $\mathcal{D}$ таким образом, чтобы для произвольного пришедшего вектора запроса $q$ мы могли быстро (за время $o(n)$) найти его ближайших соседей в $\mathcal{D}$.

Существует большое количество эффективных методов поиска ближайших соседей. В последние годы было показано, что подходы, основанные на графах, превосходят алгоритмы других типов во многих практических приложениях [4]. Большинство графовых методов используют графы ближайших соседей (или их аппроксимацию), где вершины соответствуют элементам $\mathcal{D}$, и каждая вершина соединена со своими ближайшими соседями направленными ребрами. Для заданного запроса $q$ сначала берется некоторый элемент из $\mathcal{D}$ (например случайный или заранее фиксированный) и делаются жадные шаги по направлению к $q$ на графе: на каждом шаге рассматриваются все соседи текущего элемента

и выбирается тот, который ближе всего к $q$. Существует большое количество дополнительных эвристик, позволяющих ускорить методы этого типа [29].

В то время как эмпирически известно, что алгоритмы на основе графов показывают хорошее качество в практических задачах, существует очень мало теоретических исследований, объясняющих это. Мы делаем шаг в этом направлении. Наш анализ предполагает равномерное распределение элементов датасета на сфере, и мы в основном фокусируемся на режиме $d \ll \log n$.

Предположим, что нам дан набор векторов $\mathcal{D} = \{\mathbf{x}_1, \dots \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^{d+1}$ и пусть все элементы $\mathcal{D}$ принадлежат единичной сфере, $\mathcal{D} \subset \mathcal{S}^d$. Такая постановка является важной для практических приложений, поскольку векторы признаков часто нормируются на сферу. Для заданного запроса $q \in \mathcal{S}^d$ пусть $\bar{\mathbf{x}} \in \mathcal{D}$ является его ближайшим соседом. Задача поиска ближайшего соседа — найти $\bar{\mathbf{x}}$. Еще выделяют задачу поиска приближенного близкого соседа $c, R$-ANN (approximate near neighbor): для заданных $R > 0$, $c > 1$ нам нужно найти такой элемент $\mathbf{x}'$, что $\rho(q, \mathbf{x}') \leqslant cR$, если $\rho(q, \bar{\mathbf{x}}) \leqslant R$. Здесь и далее $\rho(\cdot, \cdot)$ обозначает сферическое расстояние.

Мы предполагаем, что элементы $\mathbf{x}_i \in \mathcal{D}$ являются случайными, независимыми и равномерно распределенными на $\mathcal{S}^d$. Хотя такое предположение является весьма сильным, получение теоретических гарантий для равномерных датасетов является важным шагом на пути к пониманию свойств алгоритмов поиска ближайших соседей на основе графов близости.[1] Далее, мы предполагаем, что вектор запроса $q \in \mathcal{S}^d$ равномерно распределен на расстоянии не более $R$ от ближайшего соседа $\bar{\mathbf{x}}$ (поскольку проблема $c, R$-ANN формулируется с условием $\rho(q, \bar{\mathbf{x}}) \leqslant R$).

Мы предполагаем, что размерность $d = d(n)$ растет с ростом $n$. Есть три принципиально разных режима: плотный с $d \ll \log(n)$; разреженный с $d \gg \log(n)$; умеренный с $d = \Theta(\log(n))$. Далее будут рассматриваться первые два, из них плотный — более подробно.

Пусть мы построили некоторый граф $G$ на элементах $\mathcal{D}$. Для посту-

[1]На практике датасеты обычно распределены далеко не равномерно. Однако в наших экспериментах мы показываем, что датасеты общего вида можно отобразить в меньшую размерность, одновременно сделав их распределение более равномерным, и эта процедура позволяет улучшить алгоритмы поиска ближайших соседей на основе графов близости [41].

пившего запроса $q$ мы выбираем случайный элемент $\mathbf{x} \in \mathcal{D}$ и выполняем жадный спуск по графу: на каждом шаге мы измеряем расстояния между соседями текущего элемента и $q$ и переходим к ближайшему соседу, пока можем делать такие шаги.

**Граф ближайших соседей**

Сначала проанализируем, как жадный поиск по графу ближайших соседей работает в плотном и разреженном режимах. В плотном режиме, когда $d \ll \log(n)$, возьмем произвольное $M > 1$ и построим граф $G(M)$, соединяя пары $\mathbf{x}_i$ и $\mathbf{x}_j$ с $\rho(\mathbf{x}_i, \mathbf{x}_j) \leqslant \arcsin\left(M\, n^{-1/d}\right)$. Верна следующая теорема.

**Теорема 17.** *Пусть* $\log\log n \ll d \ll \log n$ *и* $c \geqslant 1$. *Пусть, кроме того, имеется фиксированное* $M$ *такое, что* $M > \sqrt{\frac{4c^2}{3c^2-1}}$. *Тогда, с вероятностью* $1-o(1)$, *жадный поиск ближайших соседей на основе графа* $G(M)$ *решает задачу* $c, R$*-ANN для любого* $R$ *(при* $c = 1$ *— задачу точного поиска); сложность поиска для одного запроса —* $\Theta\left(d^{1/2} \cdot n^{1/d} \cdot M^d\right) = n^{o(1)}$; *а требуемая память —* $\Theta\left(n \cdot d^{-1/2} \cdot M^d \cdot \log n\right) = n^{1+o(1)}$.

Другими словами, для плотного режима основная составляющая сложности — $n^{1/d} \cdot M^d$ для некоторой константы $M$. Здесь $M^d$ соответствует сложности одного шага, а $n^{1/d}$ — количеству шагов.

В разреженном режиме, когда $d \gg \log(n)$, возьмем произвольное $M$, $0 < M < 1$, и построим граф $G(M)$, соединяя пары $\mathbf{x}_i$ и $\mathbf{x}_j$ с $\rho(\mathbf{x}_i, \mathbf{x}_j) \leqslant \arccos\left(\sqrt{\frac{2M \ln n}{d}}\right)$. Справедлива следующая теорема.

**Теорема 18.** *Для любого* $c > 1$ *пусть* $\alpha_c = \cos\left(\frac{\pi}{2c}\right)$ *и пусть для фиксированного* $M$ *выполнено* $M < \frac{\alpha_c^2}{\alpha_c^2+1}$. *Тогда, с вероятностью* $1 - o(1)$, *жадный поиск ближайших соседей на основе графа* $G(M)$ *решает задачу* $c, R$*-ANN (для любого* $R$ *и для сферического расстояния); сложность поиска для одного запроса —* $\Theta\left(n^{1-M+o(1)}\right)$; *а требуемая память —* $\Theta\left(n^{2-M+o(1)}\right)$.

Из доказательства этой теоремы следует, что в разреженном режиме жадный алгоритм сходится не более чем за два шага с вероятностью $1 - o(1)$ (для равномерно распределенных данных).

**Влияние длинных ребер**

Согласно рассуждениям выше, если $d \ll \sqrt{\log n}$ ("очень плотный" режим), количество шагов становится доминирующим фактором сложности. В этом случае имеет смысл сократить количество шагов, добавив так называемые *длинные ребра* — ребра, соединяющие элементы, расположенные далеко друг от друга. Такие ребра могут ускорить поиск на ранних стадиях алгоритма.

Наш подход к добавлению таких длинных ребер основан на статье Клейнберга [19], в которой рассматривается двумерная сетка с некоторыми добавленными ребрами. Предполагается, что в дополнение к локальным ребрам сетки каждая вершина создает одно случайное исходящее длинное ребро, и вероятность ребра из $u$ в $v$ пропорциональна $\rho(u,v)^{-r}$. В работе доказано, что для $r = 2$ жадный поиск на основе полученного графа находит целевой элемент решетки за $O\left(\log^2 n\right)$ шагов, в то время как любое другое значение $r$ дает по крайней мере $n^{\varphi}$ шагов с $\varphi > 0$. Этот результат может быть расширен до *фиксированного* $d > 2$: в этом случае следует взять $r = d$ для достижения полилогарифмического числа шагов.

Как и в работе [19], мы проводим длинные ребра со следующими вероятностями:

$$\mathrm{P}(\text{ребра из } u \text{ в } v) = \frac{\rho(u,v)^{-d}}{\sum_{w \neq u} \rho(u,w)^{-d}}\,. \tag{9}$$

**Теорема 19.** *При условиях Теоремы 17 добавление для каждой вершины* $\Theta(\log n)$ *независимых длинных ребер с распределением* (9) *снижает количество шагов до* $O(\log n)$ *(с вероятностью* $1 - o(1)$*).*

Важно отметить, что в отличие от [19], мы предполагаем $d \to \infty$. Теорема 19 показывает, что длинные ребра позволяют гарантировать $O\left(\log n\right)$ шагов, в то время как простые графы ближайших соседей дают $\Theta\left(n^{1/d}\right)$. Следовательно, сокращение числа шагов имеет смысл, если $\log n < n^{1/d}$, то есть $d < \frac{\log n}{\log\log n}$.

Однако Теорему 19 сложно применять на практике, поскольку вероятности в (9) зависят от $d$, а реальные датасеты обычно имеют внутреннюю размерность, отличную от исходной [27]. Кроме того, внутренняя размер-

ность может меняться в зависимости от региона внутри одного датасета. Таким образом, не ясно как выбрать правильное значение $d$ в (9).

Однако, как мы более подробно обсуждаем в [41], можно сделать распределение в (9) не зависящим от размерности. Для этого мы переформулируем вероятности в терминах *рангов* вместо *расстояний*. А именно, отсортируем все элементы по их близости к некоторому элементу $u$. Затем определим вероятность добавления ребра из $u$ в другой элемент следующим образом:

$$\mathrm{P}(\text{ребра в } k\text{-го соседа}) = \frac{1/k}{\sum_{i=1}^{n} 1/i} \sim \frac{1}{k \ln n}\,. \tag{10}$$

Это распределение *не зависит от размерности*, а для равномерных $d$-мерных датасетов обладает теми же гарантиями, что и (9).

#### Влияние лучевого поиска

*Лучевой поиск* (beam search) — метод обхода графа, в котором на каждом шаге рассматривается наилучший элемент из некоторого фиксированного множества. Эта техника широко используется в графовых алгоритмах поиска ближайших соседей, поскольку она позволяет значительно улучшить точность по сравнению с жадным обходом графа [29]. Следующая теорема показывает, что лучевой поиск действительно снижает сложность алгоритма в рассматриваемых условиях.

**Теорема 20.** *Пусть $M > 1$, $L > 1$ фиксированы и $M^2\left(1 - \frac{M^2}{4L^2}\right) > 1$. Пусть, кроме того, $\log\log n \ll d \ll \log n$. Предположим, что мы используем лучевой поиск с $\frac{C\,L^d}{\sqrt{d}}$ кандидатами (для достаточно большого $C$) и мы добавляем $\Theta(\log n)$ длинных ребер по схеме, описанной выше. Тогда поиск ближайших соседей на основе графа $G(M)$ находит ближайшего соседа с вероятностью $1-o(1)$. Сложность поиска для одного запроса — $O\left(L^d \cdot M^d\right)$.*

Из этой теоремы следует, что лучевой поиск позволяет заметно понизить степени графа, что в итоге приводит к снижению времени обработки одного запроса. Чтобы это показать, возьмем $M = \sqrt{\frac{3}{2}}$ и произвольное $L > \sqrt{\frac{9}{8}}$. Видим, что в таком случае сложность поиска снижается до $\left(\frac{27}{16}\right)^{d/2}$, что меньше чем $2^{d/2}$ в Теореме 17.

# Заключение

В процессе подготовки данной диссертации были опубликованы статьи [5, 13, 15, 20, 21, 24, 35, 36, 37, 38, 39, 40, 41, 42, 43].

В статьях [13, 20, 21, 24, 35, 36, 37, 38, 40] проанализированы свойства существующих моделей сложных сетей и разработаны новые реалистичные модели с желаемыми количественными и топологическими свойствами.

В статьях [15, 42, 43] исследуется задача выделения сообществ в сложных сетях: выбор правильной модели в методах вероятностной оптимизации, выделение сообществ на основе распространения информации в графе, а также проблема выбора функции качества при сравнении алгоритмов выделения сообществ.

В статьях [5, 25, 41] изучаются приложения анализа графов к задачам датирования веб-страниц, поиска влиятельных вершин в графах, а также эффективного поиска ближайших соседей.

Основные результаты работы, выносимые на защиту:

- Новый класс моделей *обобщенного предпочтительного присоединения* и теоретические результаты, полученные для всего класса (распределение степеней вершин, локальный и глобальный коэффициенты кластеризации, корреляции степеней вершин).
- Утверждение о том, что глобальный коэффициент кластеризации стремится к нулю с ростом числа вершин для всех графов со степенным распределением степеней вершин с бесконечной дисперсией.
- Теоретический анализ модулярности для $d$-регулярных графов, модели предпочтительного присоединения и модели пространственного предпочтительного присоединения.
- Принцип *предпочтительного размещения*, который позволяет генерировать структуры со степенным распределением размеров кластеров, и анализ полученных структур.
- Модель *предпочтительного присоединения с устареванием* и анализ ее свойств.

- Анализ алгоритмов выделения сообществ, основанных на максимизации правдоподобия, и новая модель ILFR для этой задачи.
- Системный анализ задачи выделения сообществ на основе распространения информации в графе и новые эффективные подходы для этой задачи.
- Теоретический анализ проблемы выбора подходящей функции качества для алгоритмов выделения сообществ.
- Новый алгоритм датирования веб-страниц, основанный на вероятностной оптимизации.
- Новый алгоритм для быстрого выделения вершин большой степени в сложных сетях.
- Теоретические гарантии для алгоритмов поиска ближайших соседей, основанных на графах близости.

**Благодарности**

Автор диссертации благодарна Андрею Михайловичу Райгородскому за многочисленные научные консультации в процессе исследования.

# Литература


[1] W. Aiello, A. Bonato, C. Cooper, J. Janssen, and P. Prałat. A spatial web graph model with local influence regions. *Internet Mathematics*, 5(1-2):175–196, 2008.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[3] A. Arenas, L. Danon, A. Diaz-Guilera, P. M. Gleiser, and R. Guimera. Community analysis in social networks. *The European Physical Journal B*, 38(2):373–380, 2004.

[4] M. Aumüller, E. Bernhardsson, and A. Faithfull. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87:101374, 2020.

[5] K. Avrachenkov, N. Litvak, L. Ostroumova Prokhorenkova, and E. Suyargulova. Quick detection of high-degree entities in large directed networks. In *2014 IEEE International Conference on Data Mining*, pages 20–29. IEEE, 2014.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[7] P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[8] M. Bloznelis and V. Kurauskas. Clustering coefficient of random intersection graphs with infinite degree variance. *Internet Mathematics*, 2016.

[9] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[10] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. In *The Structure and Dynamics of Networks*, pages 384–395. Princeton University Press, 2011.

[11] P. G. Buckley and D. Osthus. Popularity based random graph models leading to a scale-free degree sequence. *Discrete Mathematics*, 282(1-3):53–68, 2004.

[12] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[13] A. Dorodnykh, L. Ostroumova Prokhorenkova, and E. Samosvat. Preferential placement for community structure formation. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 75–89. Springer, 2017.

[14] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

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

[16] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical review E*, 68(6):065103, 2003.

[17] P. Holme and B. J. Kim. Growing scale-free networks with tunable clustering. *Physical review E*, 65(2):026107, 2002.

[18] B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[19] J. Kleinberg. The small-world phenomenon: an algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, 2000.

[20] A. Krot and L. Ostroumova Prokhorenkova. Assortativity in generalized preferential attachment models. *Internet Mathematics*, 2017.

[21] A. Krot and L. Ostroumova Prokhorenkova. Local clustering coefficient in generalized preferential attachment models. *Internet Mathematics*, 2017.

[22] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 57–65. IEEE, 2000.

[23] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[24] D. Lefortier, L. Ostroumova, and E. Samosvat. Evolution of the media web. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 80–92. Springer, 2013.

[25] D. Lefortier, L. Ostroumova, E. Samosvat, and P. Serdyukov. Timely crawling of high-quality ephemeral new content. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 745–750, 2013.

[26] J. Leskovec. *Dynamics of large networks*. PhD thesis, Carnegie Mellon University, School of Computer Science, 2008.

[27] P.-C. Lin and W.-L. Zhao. Graph based nearest neighbor search: Promises and failures. *arXiv preprint arXiv:1904.02077*, 2019.

[28] N. Litvak and R. Van Der Hofstad. Uncovering disassortativity in large scale-free networks. *Physical Review E*, 87(2):022801, 2013.

[29] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[30] F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics reports*, 533(4):95–142, 2013.

[31] M. Newman. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. *arXiv preprint arXiv:1606.02319*, 2016.

[32] M. E. Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[33] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[34] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[35] L. Ostroumova, A. Ryabchenko, and E. Samosvat. Generalized preferential attachment: tunable power-law degree distribution and clustering coefficient. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 185–202. Springer, 2013.

[36] L. Ostroumova Prokhorenkova. Global clustering coefficient in scale-free weighted and unweighted networks. *Internet Mathematics*, 12(1-2):54–67, 2016.

[37] L. Ostroumova Prokhorenkova. General results on preferential attachment and clustering coefficient. *Optimization Letters*, 11(2):279–298, 2017.

[38] L. Ostroumova Prokhorenkova, P. Prałat, and A. Raigorodskii. Modularity of complex networks models. *Internet Mathematics*, 2017.

[39] L. Ostroumova Prokhorenkova, P. Prokhorenkov, E. Samosvat, and P. Serdyukov. Publication date prediction through reverse engineering

of the web. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 123–132, 2016.

[40] L. Ostroumova Prokhorenkova and E. Samosvat. Recency-based preferential attachment models. *Journal of Complex Networks*, 4(4):475–499, 2016.

[41] L. Prokhorenkova and A. Shekhovtsov. Graph-based nearest neighbor search: From practice to theory. In *International Conference on Machine Learning*, pages 7803–7813. PMLR, 2020.

[42] L. Prokhorenkova and A. Tikhonov. Community detection through likelihood optimization: in search of a sound model. In *The World Wide Web Conference*, pages 1498–1508, 2019.

[43] L. Prokhorenkova, A. Tikhonov, and N. Litvak. Learning clusters through information diffusion. In *The World Wide Web Conference*, pages 3151–3157, 2019.

[44] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. *arXiv preprint arXiv:1105.0697*, 2011.

[45] A. Sheikhahmadi, M. A. Nematbakhsh, and A. Shokrollahi. Improving detection of influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications*, 436:833–845, 2015.

[46] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[47] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang. Dynamics of information diffusion and its applications on complex networks. *Physics Reports*, 651:1–34, 2016.

[48] T. Zhou, G. Yan, and B.-H. Wang. Maximal planar networks with large clustering coefficient and power-law degree distribution. *Physical Review E*, 71(4):046141, 2005.