

Федеральное государственное автономное  
образовательное учреждение высшего профессионального образования  
“Московский физико-технический институт  
(государственный университет)”  
факультет инноваций и высоких технологий

На правах рукописи

Самосват Егор Александрович

## МОДЕЛИРОВАНИЕ ИНТЕРНЕТА С ПОМОЩЬЮ СЛУЧАЙНЫХ ГРАФОВ

05.13.18 — математическое моделирование, численные методы  
и комплексы программ

Диссертация на соискание ученой степени  
кандидата физико-математических наук

Научный руководитель —  
д.ф.-м.н. А.М. Райгородский

Москва, 2014

# Оглавление

|   |           |
|---|-----------|
| <b>Введение</b>   | <b>4</b>  |
| Актуальность темы . . . . .   | 4         |
| Модели и основные характеристики сложных сетей . . . . .                              | 5         |
| Предпочтительное присоединение . . . . .  | 6         |
| Свойства медиа-веба и модели с устареванием . . . . .                                 | 7         |
| Приложение моделей к задаче обхода эфемерных страниц поиско-<br>вым роботом . . . . . | 8         |
| <b>1 Модели предпочтительного присоединения</b>                                       | <b>11</b> |
| 1.1 Модель Барабаши–Альберт . . . . .   | 11        |
| 1.1.1 Предпочтительное присоединение . . . . .  | 11        |
| 1.1.2 LCD-модель $G_m^{(n)}$ . . . . .  | 13        |
| 1.1.3 Модификация LCD-модели: модель $G_m^n$ . . . . .                                | 15        |
| 1.2 О числе подграфов случайного графа в модели $G_m^n$ . . . . .                     | 16        |
| 1.2.1 Подсчет количества треугольников . . . . .                                      | 16        |
| 1.2.2 Обобщение на случай произвольного подграфа . . . . .                            | 25        |
| 1.2.3 Доказательство теоремы о коротком спуске . . . . .                              | 28        |
| 1.2.4 Доказательство теоремы о длинном спуске . . . . .                               | 30        |
| 1.2.5 Доказательство теоремы о произвольном подграфе . . . . .                        | 35        |
| 1.3 Обобщенное предпочтительное присоединение . . . . .                               | 38        |
| 1.3.1 Определение $PA$ -класса . . . . .  | 38        |
| 1.3.2 Степенной закон распределения степеней вершин . . . . .                         | 39        |
| 1.3.3 Кластерный коэффициент . . . . .  | 40        |
| 1.3.4 Полиномиальная модель . . . . .   | 43        |
| 1.3.5 Описание модели, изученной эмпирически . . . . .                                | 44        |
| 1.3.6 Эмпирические результаты . . . . .   | 45        |
| 1.3.7 Обсуждение . . . . .  | 48        |
| 1.3.8 Доказательства теорем . . . . .   | 49        |

|          |   |           |
|----------|---|-----------|
| <b>2</b> | <b>Свойства медиа-веба и модели с устареванием</b>                            | <b>56</b> |
| 2.1      | Базовые модели . . . . .  | 56        |
| 2.2      | Свойство устаревания медиа-веба . . . . .                                     | 57        |
| 2.2.1    | Данные . . . . .  | 57        |
| 2.2.2    | Свойство устаревания . . . . .  | 58        |
| 2.3      | Модель медиа-веба . . . . .   | 58        |
| 2.4      | Теоретический анализ предложенной модели . . . . .                            | 59        |
| 2.4.1    | Распределение входящей степени . . . . .                                      | 59        |
| 2.4.2    | Свойство устаревания . . . . .  | 62        |
| 2.5      | Эмпирический анализ предложенной модели . . . . .                             | 64        |
| 2.5.1    | Оценивание параметров . . . . .   | 65        |
| 2.5.2    | Правдоподобие . . . . .   | 65        |
| <b>3</b> | <b>Приложение моделей к задаче обхода эфемерных страниц поисковым роботом</b> | <b>70</b> |
| 3.1      | Формализация проблемы . . . . .   | 70        |
| 3.2      | Источники контента . . . . .  | 73        |
| 3.3      | Оптимальный обход источников . . . . .  | 76        |
| 3.3.1    | Теоретический анализ . . . . .  | 76        |
| 3.3.2    | Реализация . . . . .  | 80        |
| 3.4      | Эксперименты . . . . .  | 82        |
| 3.4.1    | Данные . . . . .  | 83        |
| 3.4.2    | Упрощения предложенного алгоритма . . . . .                                   | 84        |
| 3.4.3    | Результаты . . . . .  | 85        |
| 3.5      | Обсуждение . . . . .  | 89        |
|          | <b>Заключение</b>   | <b>92</b> |
|          | <b>Список литературы</b>  | <b>94</b> |

# Введение

## Актуальность темы

Современный этап изучения графовой структуры сложных сетей начался сравнительно недавно, в конце 1990-х годов. По всей видимости, главным толчком к активному развитию данной области послужило появление и рост сети Интернет. Под сложными сетями обычно понимают совершенно разные графы (сети), которые встречаются в природе и обладают нетривиальными топологическими свойствами, от компьютерных и социальных сетей до биологических и экономических. Удивительно, но несмотря на столь разные области происхождения, все эти сети обладают многими общими свойствами: малый диаметр (теория шести рукопожатий), степенной закон распределения степеней вершин, выраженная кластерная структура и др., что, с одной стороны, отличает их от сильно регулярных графов вроде решеток, а с другой стороны, от случайных графов в стиле Эрдеша–Реньи. А это значит, что можно пытаться построить общую теорию подобных сетей. В эту работу включились и физики, и математики, и исследователи в области информационных технологий (computer scientists).

Физики находят аналогии между сетями и термодинамическими системами, ищут фазовые переходы в сетях, применяют методы статистической физики. Математики подходят к вопросу более формально, строго доказывая гипотезы физиков: изучение сложных сетей оказалось хорошим полигоном для приложения теории вероятностей и случайных процессов, дискретного анализа и теории графов. Исследователи в области информационных технологий пытаются извлечь практическую пользу из изучения сетевых структур: разрабатывают алгоритмы поиска сообществ в сетях и их оптимального обхода, считают PageRank и подобные ему характеристики. В этой работе автору удалось попробовать себя в каждом из этих трех амплуа.

## Модели и основные характеристики сложных сетей

В последние 10-15 лет было предложено множество моделей сетей на основе случайных графов. Идея состоит в том, чтобы с их помощью объяснять и предсказывать важные количественные и топологические характеристики растущих реальных сетей, от Интернета и социальных сетей до биологических и экономических сетей [3, 8, 28, 31, 32, 39, 48].

Простейшая характеристика вершины в сети — это ее степень, т.е. количество примыкающих к ней ребер. Поэтому не удивительно, что наиболее глубоко изученное свойство сетей — это распределение степеней вершин в них. Для большинства изученных реальных сетей оказалось, что доля вершин степени  $d$  убывает как  $d^{-\gamma}$ , причем обычно  $2 < \gamma < 3$  [5, 8, 14, 26].

Другой важной характеристикой сети является ее кластерный коэффициент — величина, отражающая склонность сети формировать кластеры, т.е. густо соединенные множества вершин. В литературе встречаются разные подходы к определению кластерного коэффициента [9]. Мы будем использовать глобальный и средний локальный кластерные коэффициенты, точные определения которых будут даны в параграфе 1.3.3. Для большинства изученных реальных сетей средний локальный кластерный коэффициент варьируется в диапазоне от 0.01 до 0.08 и не меняется значительно с ростом сети [8]. Создание моделей сетей, которые реалистично отражают не только степенной закон распределения степеней вершин, но и поведение кластерного коэффициента, является непростой задачей.

Для того, чтобы совместить в одной модели и настраиваемое распределение степеней вершин, и кластеризацию, некоторые авторы [4, 45, 49] предложили стартовать с заданного априорного распределения степеней вершин и заданной кластеризации, а затем генерировать случайный граф, подчиненный этим ограничениям. Другими словами, среди всего множества графов с заданным количеством вершин они рассматривали подмножество графов с желаемыми свойствами, а затем равновероятно выбирали случайный граф из этого подмножества. Поступая таким образом, авторы автоматически получали желаемые характеристики случайного графа. Однако, такой подход не дает ответа на вопрос о происхождении и природе этих характеристик, более того, он кажется не достаточно общим и может быть заподозрен в “переобучении”.

Более естественный способ — рассмотреть граф как результат некоторого случайного процесса, определенного в терминах простых естественных правил, которые гарантируют желаемые свойства, наблюдаемые в реальных се-

тях. Видимо, наиболее широко изученной реализацией этого подхода является предпочтительное присоединение.

## Предпочтительное присоединение

Механизм предпочтительного присоединения (preferential attachment) был положен в основу модели развития Интернета в 1999 году Барабаши и Альберт [5]. Их гипотеза состояла в том, что в Интернете новые страницы “предпочитают” цитировать более популярные страницы, т.е. с большей вероятностью ссылаются на те страницы, которые до этого уже много цитировались. С помощью идеи предпочтительного присоединения удалось объяснить малый диаметр Интернета, степенной закон распределения степеней вершин в нем, а также фазовый переход в размерах компонент связности.

Здесь надо отметить, что Барабаши и Альберт предложили именно идею предпочтительного присоединения, а не конкретную модель. Эта идея нашла выражение в целом множестве моделей с различными свойствами. Одной из популярных моделей предпочтительного присоединения является LCD-модель, созданная Боллобашем и Риорданом.

Для определения глобального кластерного коэффициента необходимо знать распределение числа треугольников и пар примыкающих ребер в графе, такой анализ для LCD-модели был проведен в [9]. Мы в свою очередь провели этот анализ для модификации LCD-модели, которая также удобна для анализа распределения более сложных подграфов. Этому вопросу посвящен раздел 1.2. Интересный результат состоит в том, что асимптотическое поведение с ростом графа математического ожидания числа копий фиксированного подграфа определяется лишь количеством вершин степени ноль, один и два в этом подграфе. Так, если в фиксированном подграфе все степени вершин больше трех, то математическое ожидание числа его копий ограничено постоянной.

Как уже говорилось, существует множество моделей, использующих идею Барабаши–Альберт, а также ее модификации. При этом доказательства теорем во всех этих моделях довольно похожи между собой, но каждый раз занимают несколько страниц текста. Нами придуман новый подход к моделям предпочтительного присоединения, который позволяет получать аналогичные результаты сразу для множества моделей и не повторять доказательства заново. Этому подходу посвящен раздел 1.3.

На основе этого подхода предложена модель с качественно более реалистичным поведением глобального кластерного коэффициента, чем в преды-

дущих моделях предпочтительного присоединения. Мы также показываем, что в моделях предпочтительного присоединения поведение глобального кластерного коэффициента и среднего локального кластерного коэффициента разительно отличаются. Все наши теоретические результаты проиллюстрированы экспериментально.

Первая глава этой работы соответствует математическому подходу к анализу сложных сетей и основана на статьях автора [51, 52, 53].

## Свойства медиа-веба и модели с устареванием

Вторая глава этой работы соответствует физическому подходу к анализу сложных сетей и основана на статье автора [54].

Один из главных недостатков моделей предпочтительного присоединения состоит в том, что они уделяют слишком много внимания старым страницам и не реалистично объясняют, как появляются ссылки на вновь созданные страницы. В главе 2 мы изучаем медиа-веб, т.е. высокодинамическую часть веба, где ежедневно появляется множество новых страниц, связанных с медиа-контентом: новостями, постами в блогах и форумах. Отметим, что некоторые отдельные части веба уже изучались ранее, например, в [35] была предложена модель для социального веба.

Ясно, что предпочтительное присоединение плохо подходит для описания эволюции этой части веба. Действительно, в новостях и блогах редко цитируют сюжеты, потерявшие свою актуальность, какими бы популярными они ни были до этого. В разделе 2.2 мы подробно анализируем этот факт и вводим *свойство устаревания* медиа-страниц. Это наблюдение натолкнуло нас на мысль модифицировать предпочтительное присоединение, понижая вероятность ссылки на неактуальные страницы. Модели, обладающие таким свойством, мы называем моделями *с устареванием*. Мы предположили, что “качество” статьи тоже влияет на ее цитируемость. Это предположение близко к идее модели приспособления (fitness model) [7], в ней каждая страница имеет свою приспособленность, которая повышает вероятность того, что она будет процитирована.

Таким образом, в разделе 2.3 мы предлагаем новый класс моделей эволюции сетей, в котором разные функции *привлекательности* страниц возможны, включая взятые из моделей предпочтительного присоединения и модели приспособления, но также и новые — с устареванием, отвечающие за особенности медиа-веба.

В разделе 2.4 мы анализируем модели с устареванием теоретически и по-

казываем, какие из них реалистично предсказывают одновременно и распределение степеней вершин, и свойство устаревания медиа-веба. Наконец, в 2.5 мы сравниваем эти модели путем оценивания для каждой модели правдоподобия реальных данных при условии, что данные появились в соответствии с этой моделью. Такой подход позволяет нам сравнить модели количественно, при этом предложенные модели оказываются значительно лучше предыдущих. Таким образом, следуя духу физики, мы начинаем с экспериментов, чтобы предложить модель, объясняющую реальные данные, а затем с помощью новых экспериментов проверяем ее достоверность.

Один из самых удивительных выводов состоит в том, что в медиа-вебе вероятность процитировать страницу определяется скорее качеством страницы, чем ее текущей популярностью.

## Приложение моделей к задаче обхода эфемерных страниц поисковым роботом

В третьей главе мы рассматриваем приложение моделей с устареванием для одной из задач информационного поиска, эта глава основана на статье автора [55].

Лучшее понимание эволюции медиа-веба помогло нам придумать алгоритм для эффективного обхода этой части Интернета поисковым роботом. Поясним, в чем состояла особенность решаемой задачи.

Поисковый робот традиционно выполняет две задачи: обнаружение неизвестных качественных страниц и обновление обнаруженных ранее. Обе эти проблемы активно исследовались в течение последнего десятилетия [40]. Однако, в последнее время роль веба как средства массовой информации стала особенно важной. Благодаря этой тенденции на первый план выходит вопрос о скорости реакции поискового робота, т.е. задача уменьшения задержки между моментом создания новой страницы и моментом ее обнаружения поисковым роботом.

Как уже обсуждалось, медиа-страницы интересны пользователям лишь несколько дней после своего появления. Чтобы подчеркнуть эту особенность, мы называем страницы, к которым быстро пропадает интерес пользователей, *эфемерными*. Такие страницы появляются в медиа-вебе, но это не единственный их источник, например, объявления о продажах, анонсы мероприятий также являются эфемерными.

Цена задержки между моментом создания эфемерной страницы и момен-



том ее скачивания поисковым роботом очень велика с точки зрения удовлетворения пользовательского интереса. Более того, если поисковый робот не сможет найти эфемерную страницу на пике пользовательского интереса, то, видимо, вообще нет необходимости ее скачивать. Таким образом, проблема быстрого обнаружения и скачивания новых эфемерных страниц является важной, но, насколько известно автору, слабо изученной в литературе. Действительно, множество метрик было предложено для измерения покрытия и свежести скачанного корпуса [17, 19, 40], но все они не берут в расчет особенности эфемерных страниц, т.е. деградацию полезности скачанных страниц для поисковой системы со временем. В разделе 3.1 мы формализуем задачу обхода эфемерных страниц путем введения подходящей метрики. Затем мы описываем предлагаемый алгоритм решения поставленной задачи.

Наш опыт использования Интернета подсказывает, что большинство публикуемого свежего контента (новости, посты в блогах, объявления) может быть найдено на сравнительно небольшом количестве “хабов” или *источников контента*. Собственно с таких хабов мы обычно и начинаем поиск нового — это главные страницы хостов, новостные рубрики или специальные страницы с новостями. Довольно естественная идея состоит в том, чтобы искать свежий контент путем мониторинга появления новых ссылок на хабах. В разделе 3.2 мы проверяем, что большинство интересных ссылок действительно могут быть найдены на небольшом множестве источников контента, а также предлагаем простую процедуру для поиска источников контента.

Сложность состоит в том, что надо не только найти новые ссылки, но и скачать страницы, на которые они ведут. В итоге возникает вопрос, как оптимальным образом распределить ресурсы между мониторингом хабов и обходом найденных страниц. В разделе 3.3 мы предлагаем алгоритм, который динамически оценивает для каждого источника контента скорость появления новых ссылок (она зависит от времени суток и дня недели), а также их “качество”, затем в результате решения оптимизационной задачи алгоритм подбирает оптимальное расписание “переобхода” (т.е. повторного обхода) источников контента и скачивания новых страниц.

“Качество” страницы можно измерять разными способами, и оно, например, может быть основано на ссылочной структуре веба (входящая степень [33] или PageRank [2, 20] ) или на каких-то внешних сигналах (например, логге запросов [27, 41, 42] или количестве раз, которое страница была показана в поисковой выдаче [42]). В этой работе для оценки качества мы предлагаем использовать количество кликов на страницу в поисковой выдаче, которое наиболее достоверно с точки зрения поисковой системы отражает интерес

пользователей к содержанию страницы. Более того, выбор такой меры качества нам в итоге позволяет использовать обратную связь от пользователей для того, чтобы наш алгоритм лучше удовлетворял их текущим интересам.

**Благодарности.** Автор признателен профессору Андрею Михайловичу Райгородскому за неоценимую помощь в работе. Автор также благодарен своим научным коллегам Людмиле Остроумовой, Дамьену Лефортье и Александру Рябченко за интересные дискуссии и плодотворную работу.

# Глава 1

## Модели предпочтительного присоединения

### 1.1 Модель Барабаши–Альберт

#### 1.1.1 Предпочтительное присоединение

В 1999 году Барабаши и Альберт заметили [5], что распределение степеней вершин в ссылочном графе Интернета подчиняется степенному закону. В качестве возможного объяснения этого феномена они предложили случайный процесс, строящий граф шаг за шагом. На каждом шаге процесса к графу добавляется новая вершина и она соединяется с  $m$  разными вершинами, уже существующими в графе, причем вершины выбираются с вероятностью, пропорциональной их текущим степеням. Указанное условие задает класс моделей, которые Барабаши и Альберт назвали моделями предпочтительного присоединения.

Поясним, почему идея Барабаши и Альберт приводит не к одной модели, а к целому классу моделей. Обозначим через  $d_v^n$  степень вершины  $v$  в растущем графе в момент времени  $n$ . На каждом шаге добавляется  $m$  ребер, таким образом мы имеем  $\sum_v d_v^n = 2mn$ . Это наблюдение и правило предпочтительного присоединения означают, что

$$\mathbf{P}(d_v^{n+1} = d + 1 \mid d_v^n = d) = \frac{d}{2n}, \quad (1.1)$$

где  $\mathbf{P}$  — вероятность события. Однако, условие (1.1) на вероятность присоединения не задает полностью распределение  $m$  вершин, выбираемых на каждом шаге. Так,  $m$  ребер на очередном шаге могут проводиться независимо, а могут существовать и сложные зависимости. Таким образом, действительно, Барабаши и Альберт предложили не одну модель, а класс моделей. Как было

показано позже, существуют модели с принципиально разными свойствами, удовлетворяющие описанию Барабаши и Альберт.

**Теорема 1** (Боллобаш, Риордан [9]). Пусть  $f(n), n \geq 2$ , — произвольная целочисленная функция, такая, что  $f(2) = 0$ ,  $f(n) \leq f(n+1) \leq f(n) + 1$  для всех  $n \geq 2$  и  $f(n) \rightarrow \infty$  при  $n \rightarrow \infty$ . Тогда существует такая модель из класса Барабаши–Альберт, что в ней с вероятностью, стремящейся к единице при  $n \rightarrow \infty$ , случайный граф содержит в точности  $f(n)$  треугольников.

В [10] Боллобаш и Риордан предложили конкретную модель из класса Барабаши–Альберт, известную как LCD-модель, и доказали, что для  $d < n^{\frac{1}{15}}$  доля вершин степени  $d$  асимптотически почти наверное подчиняется степенному закону с параметром 3. Недавно Гречников заметно улучшил этот результат [29] и избавился от ограничения на  $d$ . Также было показано, что математическое ожидание глобального кластерного коэффициента асимптотически пропорционально  $\frac{(\ln n)^2}{n}$  и таким образом стремится к нулю с ростом размера графа [9].

Можно получить естественное обобщение LCD-модели, потребовав, чтобы вероятность присоединения вершины  $n+1$  к вершине  $v$  была пропорциональна величине  $d_v^n + t\beta$ , где  $\beta$  является константой, отражающей начальную привлекательность вершины. Бакли и Остус [15] предложили точно определенную модель для неотрицательных целых  $\beta$ . Мори [38] обобщил модель для действительных  $\beta > -1$ . Для обеих моделей было показано, что распределение степеней вершин подчиняется степенному закону с показателем  $3 + \beta$  в диапазоне малых степеней. В соответствии с недавним результатом Эгмана и Нобла [25] математическое ожидание глобального кластерного коэффициента в модели Мори с  $\beta > 0$  асимптотически пропорционально величине  $\frac{\ln n}{n}$ . Для  $\beta = 0$  модель Мори практически идентична LCD-модели. Поэтому авторы подчеркивают [25] неожиданную разницу в поведении глобального кластерного коэффициента при  $\beta = 0$  и  $\beta > 0$  ( $\frac{(\ln n)^2}{n}$  против  $\frac{\ln n}{n}$ ).

Главный недостаток описанных моделей — нереалистичное поведение кластерного коэффициента. Фактически, для всех описанных моделей кластерный коэффициент стремится к нулю с ростом размера графа, тогда как для реальных сетей он считается постоянным [8].

Модель с асимптотически постоянным кластерным коэффициентом (средним локальным) была предложена Холмом и Кимом [30]. Идея состоит в том, чтобы чередовать шаги предпочтительного присоединения с шагами создания треугольников. Эта модель позволяет настраивать кластерный коэффициент

путем изменения вероятности шага, на котором образуется треугольник. Однако, эксперименты и эмпирический анализ показали, что распределение степеней вершин в этой модели подчиняется степенному закону с фиксированным показателем экспоненты, близким к 3, тогда как большинство реальных сетей имеют показатель меньше 3. RAN (random Apollonian network) — еще один интересный пример [50] модели типа Барабаши–Альберт с асимптотически постоянным средним локальным кластерным коэффициентом.

Существует множество моделей, не упомянутых здесь, которые также основаны на идее предпочтительного присоединения. Анализ свойств таких моделей очень часто схож. В разделе 1.3 мы предлагаем общий подход к моделям предпочтительного присоединения и доказываем теоремы, нацеленные на упрощение этого анализа. Хотя предложенные ранее модели и имели постоянный средний локальный кластерный коэффициент, их глобальный кластерный коэффициент стремился к нулю с ростом размера графа. Мы предлагаем модель, которая позволяет одновременно настраивать и распределение степеней вершин, и глобальный кластерный коэффициент.

### 1.1.2 LCD-модель $G_m^{(n)}$

Как было сказано выше, одной из популярных моделей из класса Барабаши и Альберт является LCD-модель (LCD от Linearized Chord Diagram), предложенная Боллобашем и Риорданом (см. [9, 10]).

Ниже мы описываем эту модель.

Мы будем строить случайный граф с  $t$  вершинами и  $mt$  ребрами,  $m \in \mathbb{N}$ . Разберем сначала случай  $m = 1$ . Рассмотрим последовательность вершин  $\{v_1, v_2, \dots\}$ . Пусть  $d_G(v)$  — степень вершины  $v$  в графе  $G$ . Мы индуктивно определим процесс  $(G_1^{(t)})_{t \geq 1}$  так, что  $G_1^{(t)}$  является графом на множестве вершин  $\{v_i : 1 \leq i \leq t\}$ . Начнем с  $G_1^{(1)}$  — графа с одной вершиной и одной петлей. Имея  $G_1^{(t-1)}$ , мы построим  $G_1^{(t)}$ , добавляя вершину  $v_t$  вместе с ребром между вершиной  $v_t$  и вершиной  $v_i$ , где  $i$  выбирается случайно с вероятностью

$$P(i = s) = \begin{cases} d_{G_1^{(t-1)}}(v_s)/(2t - 1) & 1 \leq s \leq t - 1 \\ 1/(2t - 1) & s = t \end{cases}$$

В случае, когда  $m > 1$ , добавим  $m$  ребер из  $v_t$  по одному за раз (независимо от всех предыдущих ребер), каждый раз пересчитывая степени вершин. Это определение эквивалентно следующему:

сначала построим процесс  $(G_1^{(t)})_{t \geq 1}$  на последовательности вершин  $\{v'_1, v'_2, \dots\}$ , а потом сформируем графы  $(G_m^{(t)})_{t \geq 1}$  из графов  $(G_1^{(mt)})_{t \geq 1}$  посредством склеивания вершин  $\{v'_1, v'_2, \dots, v'_m\}$  в вершину  $v_1$ , склеивания вершин  $\{v'_{m+1}, v'_{m+2}, \dots, v'_{2m}\}$  в вершину  $v_2$  и т.д.

Для графов в LCD-модели было получено много интересных результатов, среди них наиболее близок к нашему исследованию следующий результат, который можно найти в [9].

Пусть  $S$  – граф на вершинах  $V(S) = \{1, 2, \dots, n\}$ , возможно, с петлями. Пусть каждое ребро  $ij$  в  $S$  при  $i \leq j$  ориентировано из  $j$  в  $i$ . Обозначим через  $V^-(S)$  множество вершин, в которые приходят ребра. Для  $i \in V(S)$  обозначим через  $d_S^{\text{in}}(i)$  степень по входящим ребрам вершины  $i$  в  $S$ . Обозначим через  $C_S(t)$  число ребер в  $S$ , “пересекающих”  $t$ , т.е. число ребер  $ij$  в  $S$ , для которых  $i \leq t \leq j$ . Будем говорить, что  $S$  – это подграф графа  $G_1^{(n)}$ , и обозначать такое событие записью  $S \subseteq G_1^{(n)}$ , если в точности ребра  $S$  встречаются в  $G_1^{(n)}$ , а не  $G_1^{(n)}$  содержит подграф, изоморфный  $S$ . Будем говорить также, что  $S$  является *допустимым*, если степени по исходящим ребрам всех вершин  $S$  не превосходят 1. Таким образом,  $S$  допустимый, если  $\mathbf{P}(S \subseteq G_1^{(n)}) \neq 0$ . Справедлива следующая теорема (см. [9]).

**Теорема 2.** Пусть  $S$  – допустимый подграф. Тогда для вероятности  $p_S$  того, что  $S \subseteq G_1^{(n)}$ , верно равенство

$$p_S = \prod_{i \in V^-(S)} d_S^{\text{in}}(i)! \prod_{ij \in E(S)} \frac{1}{2\sqrt{ij}} \exp \left( O \left( \sum_{i \in V(S)} C_S(i)^2 / i \right) \right),$$

где  $\sum_{i \in V(S)} C_S(i)^2 / i \rightarrow 0$  при  $n \rightarrow \infty$ .

Непосредственным следствием этой теоремы является следующий более общий результат.

**Теорема 3** (о произвольном подграфе). Пусть задан граф  $G_0$ , степени вершин которого равны  $d_1, \dots, d_s$ . Обозначим через  $\#(d_i = k)$  число вершин в  $G_0$ , степень каждой из которых равна  $k$ . Обозначим через  $\#(G_0, G_m^{(n)})$  число подграфов, изоморфных графу  $G_0$ , в графе  $G_m^{(n)}$ . Тогда

$$\mathbf{E} \left( \#(G_0, G_m^{(n)}) \right) = \Theta \left( n^{\#(d_i=0)} \cdot (\sqrt{n})^{\#(d_i=1)} \cdot (\ln n)^{\#(d_i=2)} \cdot m^{\frac{\sum d_i}{2}} \right).$$

Любопытно, что такого следствия мы в литературе не встречали, и оно, в частности, не приводится в [9]. Однако в [9] приводятся следующие результаты, которые прекрасно согласуются с теоремой 3.

**Теорема 4.** Пусть  $m \geq 2$  фиксировано. Пусть также  $K_3$  – полный граф на трех вершинах. Тогда

$$\mathbb{E} \left( \# \left( K_3, G_m^{(n)} \right) \right) = (1 + o(1)) \cdot \frac{m \cdot (m - 1) \cdot (m + 1)}{48} \cdot (\ln n)^3$$

при  $n \rightarrow \infty$ .

**Теорема 5.** Пусть фиксированы  $m \geq 2$  и  $l \geq 3$ . Пусть также  $C_l$  – цикл на  $l$  вершинах. Тогда

$$\mathbb{E} \left( \# \left( C_l, G_m^{(n)} \right) \right) = (1 + o(1)) \cdot c_{m,l} \cdot (\ln n)^l$$

при  $n \rightarrow \infty$ , где  $c_{m,l}$  – это положительная константа. Более того, при  $m \rightarrow \infty$  имеем  $c_{m,l} = \Theta(m^l)$ .

В настоящей работе мы рассмотрим еще одну модель из класса Барабаши–Альберт. С одной стороны, эта модель очень близка к только что описанной. С другой стороны, она чуть более естественна. В этой модели нам удастся получить два весьма общих результата, которые мы называем *теоремами о длинном и коротком спуске*. Из этих теорем следует теорема, аналогичная теореме 3 (о произвольном подграфе), а также некоторые другие результаты.

### 1.1.3 Модификация LCD-модели: модель $G_m^n$

Опишем модель, во многом схожую с LCD, но более удобную для ряда вычислений. Мы будем различать их, используя обозначения  $G_m^n$  и  $G_m^{(n)}$ . Рассмотрим последовательность вершин  $\{1, 2, \dots\}$ . Мы индуктивно определим процесс  $(G_m^t)_{t \geq 1}$  так, что  $G_m^t$  является графом на множестве вершин  $\{1, \dots, t\}$  с  $mt$  ребрами. Начнем с  $G_m^1$  – графа с одной вершиной и  $m$  петлями. Имея  $G_m^{t-1}$ , мы построим  $G_m^t$ , добавляя вершину  $t$  вместе с  $m$  ребрами, выходящими из нее. Ребра взаимно независимы, и каждое ребро соединяет вершину  $t$  с вершиной  $i$ , где  $i \in \{1, \dots, t\}$  выбирается случайно, причем

$$\mathbb{P}(i = s) = \begin{cases} d_s^{t-1} / (m \cdot (2t - 1)) & 1 \leq s \leq t - 1 \\ 1 / (2t - 1) & s = t \end{cases}$$

Здесь, как и прежде, под  $d_s^{t-1} = d_{G_m^{t-1}}(s)$  мы подразумеваем степень вершины  $s$  в графе  $G_m^{t-1}$ .

В текущих обозначениях, в отличие от обозначений для LCD, нет скобок в верхнем индексе у  $G_m^t$ . Будем использовать это отличие для указания того, о какой модели идет речь.

Главное отличие нашей модели от LCD-модели состоит в том, что мы проводим сразу все  $m$  ребер из очередной вершины, а в LCD-модели ребра проводятся последовательно и после проведения каждого ребра степени вершин пересчитываются.

## 1.2 О числе подграфов случайного графа в модели $G_m^n$

### 1.2.1 Подсчет количества треугольников

Обозначим через  $\#(K_3, G_m^n)$  количество подграфов, изоморфных  $K_3$ , в графе  $G_m^n$ .

**Теорема 6** (о числе треугольников). *Имеет место асимптотика*

$$\mathbb{E}(\#(K_3, G_m^n)) = \left(1 + O\left(\frac{1}{\ln n}\right)\right) \frac{(m-1)m(m+1)}{48} \ln^3 n.$$

Для доказательства этой теоремы нам потребуются ввести вспомогательные обозначения и доказать ряд вспомогательных утверждений. Через  $e_{ij}$  обозначим количество ребер между  $i$ -ой и  $j$ -ой вершинами. Для числа треугольников тогда имеем:

$$\#(K_3, G_m^n) = \sum_{1 \leq i < j < k \leq n} e_{ij} \cdot e_{jk} \cdot e_{ik},$$

где  $e_{ij} \cdot e_{jk} \cdot e_{ik}$  — количество треугольников с вершинами  $i$ ,  $j$  и  $k$ . В свою очередь, для искомого математического ожидания получаем формулу

$$\mathbb{E}(\#(K_3, G_m^n)) = \sum_{1 \leq i < j < k \leq n} \mathbb{E}(e_{ij} \cdot e_{jk} \cdot e_{ik}). \quad (1.2)$$

Обозначим через  $D_n$  суммарную степень вершин на  $n$ -ом шаге:

$$D_n = \sum_{i=1}^{n+1} d_i^n = (2n+1) \cdot m,$$

где мы формально полагаем  $d_{n+1}^n = m$ , это будет удобным в дальнейшем при применении рекуррентных соотношений.

Вначале докажем две технические леммы.



**Лемма 1.** *Имеет место асимптотика*

$$\prod_{k=a}^b \left(1 + \frac{m}{D_k}\right) = \left(1 + O\left(\frac{1}{a}\right)\right) \cdot \sqrt{\frac{b}{a}},$$

когда скоро  $a < b$ .

*Доказательство.* Преобразуем произведение:

$$\prod_{k=a}^b \left(1 + \frac{m}{D_k}\right) = \exp\left(\ln\left(\prod_{k=a}^b \left(1 + \frac{m}{D_k}\right)\right)\right) = \exp\left(\sum_{k=a}^b \ln\left(1 + \frac{1}{2k+1}\right)\right).$$

Воспользуемся оценкой  $\ln(1+x) = x + O(x^2)$ :

$$\begin{aligned} \exp\left(\sum_{k=a}^b \left(\frac{1}{2k+1} + O\left(\frac{1}{k^2}\right)\right)\right) &= \exp\left(\frac{1}{2} \ln\left(\frac{b}{a}\right) + O\left(\frac{1}{a}\right)\right) \\ &= \sqrt{\frac{b}{a}} \exp\left(O\left(\frac{1}{a}\right)\right) = \left(1 + O\left(\frac{1}{a}\right)\right) \cdot \sqrt{\frac{b}{a}}. \end{aligned}$$

□

**Лемма 2.** *Имеет место асимптотика*

$$\prod_{k=a}^b \left(1 + \frac{2m}{D_{k-1}} + \frac{m(m-1)}{(D_{k-1})^2}\right) = \left(1 + O\left(\frac{1}{a}\right)\right) \cdot \frac{b}{a},$$

когда скоро  $a < b$ .

*Доказательство.* Преобразуем произведение:

$$\begin{aligned} \prod_{k=a}^b \left(1 + \frac{2m}{D_{k-1}} + \frac{m(m-1)}{(D_{k-1})^2}\right) &= \exp\left(\ln\left(\prod_{k=a}^b \left(1 + \frac{2m}{D_{k-1}} + \frac{m(m-1)}{(D_{k-1})^2}\right)\right)\right) = \\ &= \exp\left(\sum_{k=a}^b \ln\left(1 + \frac{2m}{D_{k-1}} + \frac{m(m-1)}{(D_{k-1})^2}\right)\right). \end{aligned}$$

Воспользуемся оценкой  $\ln(1+x) = x + O(x^2)$ :

$$\begin{aligned} \exp\left(\sum_{k=a}^b \left(\frac{2}{2k-1} + O\left(\frac{1}{k^2}\right)\right)\right) &= \exp\left(\ln\left(\frac{b}{a}\right) + O\left(\frac{1}{a}\right)\right) = \\ &= \frac{b}{a} \exp\left(O\left(\frac{1}{a}\right)\right) = \left(1 + O\left(\frac{1}{a}\right)\right) \cdot \frac{b}{a}. \end{aligned}$$

□

Теперь перейдем к оценке величины  $\mathbf{E}(e_{ij} \cdot e_{jk} \cdot e_{ik})$ . Введем случайную величину  $I_\alpha^{ij}$ , которая представляет собой индикатор события, состоящего в том, что при добавлении  $j$ -той вершины  $\alpha$ -ое ребро из нее ( $\alpha \in \{1, \dots, m\}$ ) было проведено в  $i$ -тую вершину.

Заметим, что

$$d_i^k = d_i^{k-1} + \sum_{\alpha=1}^m I_\alpha^{ik}. \quad (1.3)$$

Это соотношение верно для  $k > i$  по определению случайного процесса; для  $k = i$  оно тоже верно, поскольку мы положили  $d_i^{i-1} = m$ .

Так как ребро между вершинами  $i$  и  $j$  ( $i < j$ ) появляется только при добавлении вершины  $j$ , имеем

$$e_{ij} = \sum_{\alpha=1}^m I_\alpha^{ij}, \quad i < j. \quad (1.4)$$

В дальнейшем нам часто придется вычислять условные математические ожидания произведений индикаторов вида  $I_\alpha^{ij}$ , при этом индикаторы могут быть либо независимыми, либо несовместными, либо совпадающими. Поэтому

$$\begin{aligned} & \mathbf{E} (I_\mu^{in} \cdot I_\nu^{jn} | G_m^{n-1}) = \\ & = \begin{cases} \mathbf{E} (I_\mu^{in} | G_m^{n-1}) \cdot \mathbf{E} (I_\nu^{jn} | G_m^{n-1}), & \text{если } \mu \neq \nu; \\ 0, & \text{если } \mu = \nu, i \neq j; \\ \mathbf{E} (I_\mu^{in} | G_m^{n-1}), & \text{если } \mu = \nu, i = j. \end{cases} \quad (*) \end{aligned}$$

Здесь условные математические ожидания берутся относительно сигма-алгебры графа  $G_m^{n-1}$ , под которой подразумевается естественная фильтрация процесса в момент времени  $n - 1$ . Заметим, что последние два равенства в (\*) выполнены и безусловно тоже.

Техника доказательства последующих лемм состоит в применении соотношений (1.3) и (1.4) и в аккуратном раскрытии произведений индикаторов с учетом описанных трех случаев.

**Лемма 3.** *Имеет место соотношение*

$$\mathbf{E} (e_{ij} \cdot e_{jk} \cdot e_{ik}) = \left( 1 + O \left( \frac{1}{k} \right) \right) \cdot \frac{m-1}{4m} \cdot \frac{1}{k^2} \cdot \mathbf{E} (e_{ij} \cdot d_i^{k-1} \cdot d_j^{k-1}).$$

*Доказательство.* Воспользуемся выражением (1.4):

$$\begin{aligned} \mathbb{E}(e_{ij} \cdot e_{jk} \cdot e_{ik}) &= \mathbb{E} \left( e_{ij} \cdot \sum_{\alpha} I_{\alpha}^{jk} \cdot \sum_{\beta} I_{\beta}^{ik} \right) = \\ &= \mathbb{E} \left( e_{ij} \cdot \sum_{\alpha=\beta} I_{\alpha}^{jk} \cdot I_{\beta}^{ik} \right) + \mathbb{E} \left( e_{ij} \cdot \sum_{\alpha \neq \beta} I_{\alpha}^{jk} \cdot I_{\beta}^{ik} \right). \end{aligned}$$

Заметим, что первое слагаемое в этом выражении обращается в ноль, так как одно и то же ребро не может идти в разные вершины. Во втором слагаемом выражения перейдем к условному математическому ожиданию (воспользуемся формулой полной вероятности):

$$\begin{aligned} \mathbb{E} \left( e_{ij} \cdot \sum_{\alpha \neq \beta} I_{\alpha}^{jk} \cdot I_{\beta}^{ik} \right) &= \mathbb{E} \left( \mathbb{E} \left( e_{ij} \cdot \sum_{\alpha \neq \beta} I_{\alpha}^{jk} \cdot I_{\beta}^{ik} \mid G_m^{k-1} \right) \right) = \\ &= \mathbb{E} \left( e_{ij} \cdot \sum_{\alpha \neq \beta} \mathbb{E} (I_{\alpha}^{jk} \cdot I_{\beta}^{ik} \mid G_m^{k-1}) \right) = \end{aligned}$$

(пользуемся (\*) при  $\alpha \neq \beta$ )

$$\begin{aligned} &= \mathbb{E} \left( e_{ij} \cdot \sum_{\alpha \neq \beta} \mathbb{E} (I_{\alpha}^{jk} \mid G_m^{k-1}) \cdot \mathbb{E} (I_{\beta}^{ik} \mid G_m^{k-1}) \right) = \\ &= \mathbb{E} \left( e_{ij} \cdot m(m-1) \cdot \frac{d_i^{k-1} d_j^{k-1}}{(D_{k-1})^2} \right) = \frac{m-1}{m} \frac{1}{(2k-1)^2} \cdot \mathbb{E} (e_{ij} \cdot d_i^{k-1} \cdot d_j^{k-1}) = \\ &= \left( 1 + O \left( \frac{1}{k} \right) \right) \cdot \frac{m-1}{4m} \cdot \frac{1}{k^2} \cdot \mathbb{E} (e_{ij} \cdot d_i^{k-1} \cdot d_j^{k-1}). \end{aligned}$$

□

**Лемма 4.** *Имеет место соотношение*

$$\mathbb{E} (e_{ij} \cdot d_i^{k-1} \cdot d_j^{k-1}) = \left( 1 + O \left( \frac{1}{j} \right) \right) \cdot \frac{k}{j} \cdot \mathbb{E} (e_{ij} \cdot d_i^j \cdot d_j^j).$$

*Доказательство.* Воспользуемся соотношением (1.3):

$$\begin{aligned}
\mathbb{E} \left( e_{ij} \cdot d_i^{k-1} \cdot d_j^{k-1} \right) &= \mathbb{E} \left( e_{ij} \cdot \left( d_i^{k-2} + \sum_{\alpha} I_{\alpha}^{i(k-1)} \right) \cdot \left( d_j^{k-2} + \sum_{\beta} I_{\beta}^{j(k-1)} \right) \right) = \\
&= \mathbb{E} \left( e_{ij} \cdot \left( d_i^{k-2} d_j^{k-2} + d_i^{k-2} \cdot \sum_{\beta} I_{\beta}^{j(k-1)} + \right. \right. \\
&\quad \left. \left. + d_j^{k-2} \cdot \sum_{\alpha} I_{\alpha}^{i(k-1)} + \sum_{\alpha} I_{\alpha}^{i(k-1)} \cdot \sum_{\beta} I_{\beta}^{j(k-1)} \right) \right) =
\end{aligned}$$

(как в лемме 3, сделаем переход к условному математическому ожиданию при условии сигма-алгебры графа на предыдущем шаге процесса)

$$\begin{aligned}
&= \mathbb{E} \left( e_{ij} \cdot \left( d_i^{k-2} d_j^{k-2} + d_i^{k-2} \frac{m d_j^{k-2}}{D_{k-2}} + d_j^{k-2} \frac{m d_i^{k-2}}{D_{k-2}} + \frac{m(m-1) d_i^{k-2} d_j^{k-2}}{(D_{k-2})^2} \right) \right) = \\
&= \mathbb{E} \left( e_{ij} \cdot d_i^{k-2} \cdot d_j^{k-2} \right) \left( 1 + \frac{2m}{D_{k-2}} + \frac{m(m-1)}{(D_{k-2})^2} \right) =
\end{aligned}$$

(мы получили рекуррентное соотношение, применим его несколько раз)

$$= \mathbb{E} \left( e_{ij} \cdot d_i^j \cdot d_j^j \right) \prod_{l=j}^{k-2} \left( 1 + \frac{2m}{D_l} + \frac{m(m-1)}{(D_l)^2} \right) =$$

(по лемме 2)

$$= \left( 1 + O \left( \frac{1}{j} \right) \right) \cdot \frac{k}{j} \cdot \mathbb{E} \left( e_{ij} \cdot d_i^j \cdot d_j^j \right).$$

□

**Лемма 5.** *Имеет место соотношение*

$$\mathbb{E} \left( e_{ij} \cdot d_i^j \cdot d_j^j \right) = \left( 1 + O \left( \frac{1}{j} \right) \right) \frac{m}{2j} \left( \mathbb{E} \left( (d_i^{j-1})^2 \right) + \mathbb{E} \left( d_i^{j-1} \right) \right).$$

*Доказательство.* Воспользуемся выражениями (1.3) и (1.4):

$$\begin{aligned}
\mathbb{E} \left( e_{ij} \cdot d_i^j \cdot d_j^j \right) &= \mathbb{E} \left( \left( \sum_{\alpha} I_{\alpha}^{ij} \right) \cdot \left( d_i^{j-1} + \sum_{\beta} I_{\beta}^{ij} \right) \cdot \left( d_j^{j-1} + \sum_{\gamma} I_{\gamma}^{jj} \right) \right) = \\
&= \mathbb{E} \left( d_i^{j-1} d_j^{j-1} \left( \sum_{\alpha} I_{\alpha}^{ij} \right) \right) + \mathbb{E} \left( d_i^{j-1} \sum_{\alpha, \gamma} I_{\alpha}^{ij} I_{\gamma}^{jj} \right) + \\
&+ \mathbb{E} \left( d_j^{j-1} \sum_{\alpha, \beta} I_{\alpha}^{ij} I_{\beta}^{ij} \right) + \mathbb{E} \left( \sum_{\alpha, \beta, \gamma} I_{\alpha}^{ij} I_{\beta}^{ij} I_{\gamma}^{jj} \right). \tag{1.5}
\end{aligned}$$

Упростим каждое из слагаемых, заменяя  $d_j^{j-1}$  на  $m$ , помня о совпадающих и несовместных индикаторах и совершая переходы к условному математическому ожиданию при условии сигма-алгебры графа на предыдущем шаге процесса.

Для первого слагаемого имеем

$$\begin{aligned}
\mathbb{E} \left( d_i^{j-1} d_j^{j-1} \left( \sum_{\alpha} I_{\alpha}^{ij} \right) \right) &= \mathbb{E} \left( d_i^{j-1} m \frac{m d_i^{j-1}}{D_{j-1}} \right) = \\
&= \mathbb{E} \left( d_i^{j-1} m^2 \frac{d_i^{j-1}}{D_{j-1}} \right) = \mathbb{E} \left( \left( d_i^{j-1} \right)^2 \right) \frac{m^2}{D_{j-1}}.
\end{aligned}$$

Для второго слагаемого имеем

$$\begin{aligned}
\mathbb{E} \left( d_i^{j-1} \sum_{\alpha, \gamma} I_{\alpha}^{ij} I_{\gamma}^{jj} \right) &= \mathbb{E} \left( d_i^{j-1} \sum_{\alpha \neq \gamma} I_{\alpha}^{ij} I_{\gamma}^{jj} \right) = \mathbb{E} \left( d_i^{j-1} m(m-1) \frac{d_i^{j-1}}{D_{j-1}} \frac{d_j^{j-1}}{D_{j-1}} \right) \\
&= \mathbb{E} \left( d_i^{j-1} m^2(m-1) \frac{d_i^{j-1}}{(D_{j-1})^2} \right) = \mathbb{E} \left( \left( d_i^{j-1} \right)^2 \right) \frac{m^2}{D_{j-1}} \frac{m-1}{D_{j-1}}.
\end{aligned}$$

Для третьего слагаемого имеем

$$\begin{aligned}
\mathbb{E} \left( d_j^{j-1} \sum_{\alpha, \beta} I_\alpha^{ij} I_\beta^{ij} \right) &= \mathbb{E} \left( d_j^{j-1} \sum_{\alpha} I_\alpha^{ij} \right) + \mathbb{E} \left( d_j^{j-1} \sum_{\alpha \neq \beta} I_\alpha^{ij} I_\beta^{ij} \right) = \\
&= \mathbb{E} \left( m \frac{m d_i^{j-1}}{D_{j-1}} \right) + \mathbb{E} \left( m \cdot m(m-1) \cdot \left( \frac{d_i^{j-1}}{D_{j-1}} \right)^2 \right) = \\
&= \mathbb{E} \left( d_i^{j-1} \right) \frac{m^2}{D_{j-1}} + \mathbb{E} \left( \left( d_i^{j-1} \right)^2 \right) \frac{m^2}{D_{j-1}} \frac{m-1}{D_{j-1}}.
\end{aligned}$$

Наконец, для четвертого слагаемого имеем

$$\begin{aligned}
\mathbb{E} \left( \sum_{\alpha, \beta, \gamma} I_\alpha^{ij} I_\beta^{ij} I_\gamma^{jj} \right) &= \mathbb{E} \left( \sum_{\alpha=\beta \neq \gamma} I_\alpha^{ij} I_\gamma^{jj} \right) + \mathbb{E} \left( \sum_{\alpha \neq \beta \neq \gamma} I_\alpha^{ij} I_\beta^{ij} I_\gamma^{jj} \right) = \\
&= \mathbb{E} \left( m(m-1) \frac{d_i^{j-1}}{D_{j-1}} \frac{d_j^{j-1}}{D_{j-1}} \right) + \mathbb{E} \left( m(m-1)(m-2) \left( \frac{d_i^{j-1}}{D_{j-1}} \right)^2 \frac{m}{D_{j-1}} \right) = \\
&= \mathbb{E} \left( d_i^{j-1} \right) \frac{m^2}{D_{j-1}} \frac{m-1}{D_{j-1}} + \mathbb{E} \left( \left( d_i^{j-1} \right)^2 \right) \frac{m^2}{D_{j-1}} \frac{(m-1)(m-2)}{(D_{j-1})^2}.
\end{aligned}$$

Подставим полученные равенства в (1.5) и приведем подобные члены:

$$\begin{aligned}
\mathbb{E} \left( e_{ij} \cdot d_i^j \cdot d_j^j \right) &= \mathbb{E} \left( \left( d_i^{j-1} \right)^2 \right) \frac{m^2}{D_{j-1}} \cdot \left( 1 + 2 \frac{m-1}{D_{j-1}} + \frac{(m-1)(m-2)}{(D_{j-1})^2} \right) + \\
&\quad + \mathbb{E} \left( d_i^{j-1} \right) \frac{m^2}{D_{j-1}} \cdot \left( 1 + \frac{m-1}{D_{j-1}} \right) =
\end{aligned}$$

(подставим  $D_{j-1} = (2j-1)m$ )

$$= \left( 1 + O \left( \frac{1}{j} \right) \right) \frac{m}{2j} \left( \mathbb{E} \left( \left( d_i^{j-1} \right)^2 \right) + \mathbb{E} \left( d_i^{j-1} \right) \right).$$

□

**Лемма 6.** *Имеет место соотношение*

$$\mathbb{E} \left( d_i^j \right) = \left( 1 + O \left( \frac{1}{i} \right) \right) \cdot m \cdot \sqrt{\frac{j}{i}}.$$

*Доказательство.* Имеем

$$\mathbb{E} \left( d_i^j \right) = \mathbb{E} \left( d_i^{j-1} + \sum_{\alpha} I_{\alpha}^{ij} \right) =$$

(сделаем переход к условному математическому ожиданию при условии сигма-алгебры графа на предыдущем шаге процесса)

$$= \mathbb{E} \left( d_i^{j-1} \left( 1 + \frac{m}{D_{j-1}} \right) \right) = \mathbb{E} \left( d_i^{j-1} \cdot \prod_{k=i-1}^{j-1} \left( 1 + \frac{m}{D_k} \right) \right).$$

Подставив  $d_i^{j-1} = m$  и воспользовавшись леммой 1, получим искомое равенство. □

**Лемма 7.** *Имеет место соотношение*

$$\mathbb{E} \left( \left( d_i^j \right)^2 \right) = \left( 1 + O \left( \frac{1}{i} \right) \right) m \cdot (m + 1) \cdot \frac{j}{i}.$$

*Доказательство.* Воспользуемся выражением (1.3):

$$\begin{aligned} \mathbb{E} \left( \left( d_i^j \right)^2 \right) &= \mathbb{E} \left( \left( d_i^{j-1} + \sum_{\alpha} I_{\alpha}^{ij} \right) \cdot \left( d_i^{j-1} + \sum_{\beta} I_{\beta}^{ij} \right) \right) = \\ &= \mathbb{E} \left( d_i^{j-1} d_i^{j-1} + d_i^{j-1} \cdot \sum_{\beta} I_{\beta}^{ij} + \sum_{\alpha} I_{\alpha}^{ij} \cdot d_i^{j-1} + \right. \\ &\quad \left. + \sum_{\alpha=\beta} I_{\alpha}^{ij} \cdot I_{\beta}^{ij} + \sum_{\alpha \neq \beta} I_{\alpha}^{ij} \cdot I_{\beta}^{ij} \right) = \end{aligned}$$

(сократим совпадающие индикаторы и сделаем переход к условному математическому ожиданию при условии сигма-алгебры графа на предыдущем шаге процесса)

$$= \mathbb{E} \left( \left( d_i^{j-1} \right)^2 + 2d_i^{j-1} \cdot \frac{m d_i^{j-1}}{D_{j-1}} + \frac{m d_i^{j-1}}{D_{j-1}} + m(m-1) \left( \frac{d_i^{j-1}}{D_{j-1}} \right)^2 \right) =$$

$$= \mathbf{E} \left( \left( d_i^{j-1} \right)^2 \right) \cdot \left( 1 + \frac{2m}{D_{j-1}} + \frac{m(m-1)}{(D_{j-1})^2} \right) + \mathbf{E} \left( d_i^{j-1} \right) \frac{m}{D_{j-1}} =$$

(применим полученное рекуррентное соотношение несколько раз)

$$= \mathbf{E} \left( \left( d_i^{j-1} \right)^2 \right) \prod_{k=i}^{j-1} \left( 1 + \frac{2m}{D_k} + \frac{m(m-1)}{(D_k)^2} \right) +$$

$$+ \sum_{k=i}^{j-1} \mathbf{E} \left( d_i^k \right) \cdot \left( \prod_{l=k}^{j-1} \left( 1 + \frac{2m}{D_l} + \frac{m(m-1)}{(D_l)^2} \right) \right) \cdot \frac{m}{D_k} =$$

(подставим  $d_i^{j-1} = m$ , воспользуемся леммой 6 для  $\mathbf{E} \left( d_i^j \right)$  и леммой 2 для

$$\prod_{l=k}^{j-1} \left( 1 + \frac{2m}{D_{k-1}} + \frac{m(m-1)}{(D_{k-1})^2} \right))$$

$$= \left( 1 + O \left( \frac{1}{i} \right) \right) m^2 \frac{j}{i} + \sum_{k=i}^{j-1} \left( 1 + O \left( \frac{1}{i} \right) \right) \cdot m \cdot \sqrt{\frac{k}{i}} \cdot \left( 1 + O \left( \frac{1}{k} \right) \right) \frac{j}{k} \cdot \frac{m}{D_k} =$$

$$= \left( 1 + O \left( \frac{1}{i} \right) \right) \left( m^2 \frac{j}{i} + \frac{j}{\sqrt{i}} m \sum_{k=i}^{j-1} \frac{1}{k} \cdot \sqrt{k} \cdot \frac{1}{2k} \right) =$$

$$= \left( 1 + O \left( \frac{1}{i} \right) \right) \left( m^2 \frac{j}{i} + m \frac{j}{\sqrt{i}} \frac{1}{\sqrt{i}} \left( 1 + O \left( \frac{1}{i^{\frac{3}{2}}} \right) \right) \right) =$$

$$= \left( 1 + O \left( \frac{1}{i} \right) \right) \left( m^2 \frac{j}{i} + m \frac{j}{\sqrt{i}} \frac{1}{\sqrt{i}} \right) = \left( 1 + O \left( \frac{1}{i} \right) \right) \cdot m(m+1) \cdot \frac{j}{i}.$$

□

Теперь докажем теорему 6.

*Доказательство.* Собрав вместе утверждения лемм 3 – 7, получаем



$$\begin{aligned}
\mathbb{E}(\#(K_3, G_m^n)) &= \sum_{1 \leq i < j < k \leq n} \mathbb{E}(e_{ij} \cdot e_{jk} \cdot e_{ik}) = \\
&= \sum_{1 \leq i < j < k \leq n} \left(1 + O\left(\frac{1}{k}\right)\right) \cdot \frac{m-1}{4m} \cdot \frac{1}{k^2} \cdot \\
&\quad \cdot \left(1 + O\left(\frac{1}{j}\right)\right) \cdot \frac{k}{j} \cdot \frac{m}{2j} \cdot \left(\mathbb{E}\left(\left(d_i^{j-1}\right)^2\right) + \mathbb{E}\left(d_i^{j-1}\right)\right) = \\
&= \sum_{1 \leq i < j < k \leq n} \left(1 + O\left(\frac{1}{j}\right)\right) \cdot \frac{m-1}{8j^2k} \cdot \left(\mathbb{E}\left(\left(d_i^{j-1}\right)^2\right) + \mathbb{E}\left(d_i^{j-1}\right)\right) = \\
&= \sum_{1 \leq i < j < k \leq n} \left(1 + O\left(\frac{1}{j}\right)\right) \cdot \frac{m-1}{8j^2k} \cdot \\
&\quad \cdot \left(\left(1 + O\left(\frac{1}{i}\right)\right) m \cdot (m+1) \cdot \frac{j}{i} + \left(1 + O\left(\frac{1}{i}\right)\right) \cdot m \cdot \sqrt{\frac{j}{i}}\right) = \\
&= \sum_{1 \leq i < j < k \leq n} \left(1 + O\left(\frac{1}{i}\right)\right) \frac{(m-1)m(m+1)}{8ijk} + \\
&+ \sum_{1 \leq i < j < k \leq n} \left(1 + O\left(\frac{1}{i}\right)\right) \frac{(m-1)m}{8i^{\frac{1}{2}}j^{\frac{3}{2}}k} = \\
&= \left(1 + O\left(\frac{1}{\ln n}\right)\right) \frac{(m-1)m(m+1)}{48} \ln^3 n.
\end{aligned}$$

Таким образом, теорема доказана. □

### 1.2.2 Обобщение на случай произвольного подграфа

Техника, применяемая при подсчете математического ожидания числа треугольников, допускает обобщение, а именно справедливы следующие теоремы о коротком и длинном спуске, которые служат инструментом для решения самых разных задач о подграфах в случайном графе в модели Барабаши – Альберт.

В обеих теоремах  $d_i^n$  – это степень вершины  $i$  в графе  $G_m^n$ ,  $e_{ij}$  – это случайная величина, равная числу ребер между вершинами  $i, j$  в графе  $G_m^n$ , а  $\xi_k$  – это произвольная функция от  $G_m^k$ , в частности, – любой многочлен от произвольного числа величин  $d_i^l, e_{pq}$ , где  $i \leq l \leq k, p \leq q \leq k$ .

**Теорема 7** (о коротком спуске). *Имеет место соотношение*

$$\begin{aligned} & \mathbb{E} \left( \xi_{n-1} \cdot (d_{\alpha_1}^m)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^m)^{a_k} \cdot e_{\beta_1 n} \cdot \dots \cdot e_{\beta_r n} \right) = \\ & = \mathbb{E} \left( \xi_{n-1} \cdot (d_{\alpha_1}^{m-1})^{a_1} \cdot \dots \cdot (d_{\alpha_k}^{m-1})^{a_k} \cdot d_{\beta_1}^{m-1} \cdot \dots \cdot d_{\beta_r}^{m-1} \right) \left( \frac{1}{n} \right)^r \cdot \Theta(1) \end{aligned}$$

при  $r \leq m$  и  $\beta_i \neq \beta_j$ . В этом соотношении зависимость от величины  $m$  имеет вид  $1 + O\left(\frac{1}{m}\right)$  и занесена в  $\Theta(1)$ .

**Теорема 8** (о длинном спуске). *Имеет место соотношение*

$$\begin{aligned} & \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^n)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^n)^{a_k} \right) = \\ & = \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^l)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^l)^{a_k} \right) \left( \frac{n}{l} \right)^{\frac{\sum a_i}{2}} \cdot \Theta(1). \end{aligned}$$

Из теорем о длинном и коротком спуске вытекает следующая теорема, аналогичная теореме 3 для LCD-модели.

**Теорема 9** (о произвольном подграфе). *Пусть задан граф  $G_0$ , степени вершин которого равны  $d_1, \dots, d_s$ . Обозначим через  $\#(d_i = k)$  число вершин в  $G_0$ , степень каждой из которых равна  $k$ . Обозначим через  $\#(G_0, G_m^n)$  число подграфов, изоморфных графу  $G_0$ , в графе  $G_m^n$ . Тогда*

$$\mathbb{E} \left( \#(G_0, G_m^n) \right) = \Theta \left( n^{\#(d_i=0)} \cdot (\sqrt{n})^{\#(d_i=1)} \cdot (\ln n)^{\#(d_i=2)} \cdot m^{\frac{\sum d_i}{2}} \right).$$

Теперь продемонстрируем применение теорем о длинном и коротком спуске на примере доказательства теоремы 9 для случая тетраэдра ( $G_0 = K_4$ ). В этом случае теорема утверждает, что математическое ожидание числа тетраэдров ограничено константой.

Далее под  $A \approx B$  подразумевается  $A = \Theta(B)$ . Количество тетраэдров на вершинах  $i, j, k, l$  равно  $e_{ij} \cdot e_{ik} \cdot e_{il} \cdot e_{jk} \cdot e_{jl} \cdot e_{kl}$ , поэтому для общего числа тетраэдров имеем

$$\begin{aligned} \mathbb{E}(\#(K_4, G_m^n)) &= \mathbb{E} \left( \sum_{i < j < k < l \leq n} e_{ij} \cdot e_{ik} \cdot e_{il} \cdot e_{jk} \cdot e_{jl} \cdot e_{kl} \right) = \quad (1.6) \\ &= \sum_{i < j < k < l \leq n} \mathbb{E}(e_{ij} \cdot e_{ik} \cdot e_{il} \cdot e_{jk} \cdot e_{jl} \cdot e_{kl}) \approx \end{aligned}$$

(применим теорему 7, считая  $\xi_{l-1} = e_{ij} \cdot e_{ik} \cdot e_{jk}$ )

$$\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \right) \mathbb{E} (e_{ij} \cdot e_{ik} \cdot e_{jk} \cdot d_i^{l-1} \cdot d_j^{l-1} \cdot d_k^{l-1}) \approx$$

(применим теорему 8, считая  $\xi_k = e_{ij} \cdot e_{ik} \cdot e_{jk}$ )

$$\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left( \frac{l}{k} \right)^{3/2} \right) \mathbb{E} (e_{ij} \cdot e_{ik} \cdot e_{jk} \cdot d_i^k \cdot d_j^k \cdot d_k^k) \approx$$

(т.к.  $m \leq d_k^k \leq 2m$ )

$$\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left( \frac{l}{k} \right)^{3/2} \cdot m \right) \mathbb{E} (e_{ij} \cdot e_{ik} \cdot e_{jk} \cdot d_i^k \cdot d_j^k) \approx$$

(применим теорему 7, считая  $\xi_{k-1} = e_{ij}$ )

$$\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left( \frac{l}{k} \right)^{3/2} \cdot \frac{1}{k^2} \cdot m \right) \mathbb{E} (e_{ij} \cdot d_i^{k-1} \cdot d_j^{k-1} \cdot d_i^{k-1} \cdot d_j^{k-1}) =$$

$$= \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left( \frac{l}{k} \right)^{3/2} \cdot \frac{1}{k^2} \cdot m \right) \mathbb{E} (e_{ij} \cdot (d_i^{k-1})^2 \cdot (d_j^{k-1})^2) \approx$$

(применим теорему 8, считая  $\xi_j = e_{ij}$ )

$$\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left( \frac{l}{k} \right)^{3/2} \cdot \frac{1}{k^2} \cdot \left( \frac{k}{j} \right)^{4/2} \cdot m \right) \mathbb{E} (e_{ij} \cdot (d_i^j)^2 \cdot (d_j^j)^2) \approx$$

(т.к.  $m \leq d_j^j \leq 2m$ )

$$\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left( \frac{l}{k} \right)^{3/2} \cdot \frac{1}{k^2} \cdot \left( \frac{k}{j} \right)^{4/2} \cdot m^3 \right) \mathbb{E} (e_{ij} \cdot (d_i^j)^2) \approx$$

(применим теорему 7)

$$\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left( \frac{l}{k} \right)^{3/2} \cdot \frac{1}{k^2} \cdot \left( \frac{k}{j} \right)^{4/2} \cdot \frac{1}{j} \cdot m^3 \right) \mathbb{E} \left( (d_i^{j-1})^3 \right) \approx$$

(применим теорему 8)

$$\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left( \frac{l}{k} \right)^{3/2} \cdot \frac{1}{k^2} \cdot \left( \frac{k}{j} \right)^{4/2} \cdot \frac{1}{j} \cdot \left( \frac{j}{i} \right)^{3/2} \cdot m^3 \right) \mathbb{E} \left( (d_i^i)^3 \right) \approx$$

(т.к.  $m \leq d_i^i \leq 2m$ )

$$\begin{aligned} &\approx \sum_{i < j < k < l \leq n} \left( \frac{1}{l^3} \cdot \left(\frac{l}{k}\right)^{3/2} \cdot \frac{1}{k^2} \cdot \left(\frac{k}{j}\right)^{4/2} \cdot \frac{1}{j} \cdot \left(\frac{j}{i}\right)^{3/2} \cdot m^6 \right) = \\ &= \sum_{i < j < k < l \leq n} \left( \frac{1}{lkji} \right)^{3/2} \cdot m^6. \end{aligned}$$

Этот ряд сходится, а значит,  $\mathbf{E}(\#(K_4, G_m^n)) \approx \text{const}$ .

Теоремы о коротком и длинном спуске позволяют оценить математическое ожидание любого многочлена от  $d_i^l$  и  $e_{pq}$ , в котором степени  $e_{pq}$  не превосходят 1. В частности, такими многочленами описывается число подграфов, но этим не ограничивается множество величин, описываемых такими многочленами.

### 1.2.3 Доказательство теоремы о коротком спуске

Перепишем, используя индикаторы  $I_\mu^{ij}$ , выражение из теоремы:

$$\begin{aligned} &\mathbf{E} \left( \xi_{n-1} \cdot (d_{\alpha_1}^n)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^n)^{a_k} \cdot e_{\beta_1 n} \cdot \dots \cdot e_{\beta_r n} \right) = \\ &= \mathbf{E} \left( \xi_{n-1} \cdot \left( d_{\alpha_1}^{n-1} + \sum_{\mu} I_{\mu}^{\alpha_1 n} \right)^{a_1} \cdot \dots \cdot \left( d_{\alpha_k}^{n-1} + \sum_{\mu} I_{\mu}^{\alpha_k n} \right)^{a_k} \cdot \right. \\ &\quad \left. \cdot \left( \sum_{\mu} I_{\mu}^{\beta_1 n} \right) \cdot \dots \cdot \left( \sum_{\mu} I_{\mu}^{\beta_r n} \right) \right). \quad (1.7) \end{aligned}$$

Вспомним, что для произведений индикаторов верны соотношения (\*). Мысленно раскроем скобки в правой части (1.7) и рассмотрим математическое ожидание каждого слагаемого. С учетом (\*) часть слагаемых сократится из-за несовместности индикаторов, а в части слагаемых сократятся совпавшие индикаторы. Независимостью мы пока не пользуемся, а потому здесь нет необходимости рассматривать условное математическое ожидание.

В результате каждое слагаемое будет иметь вид

$$\mathbf{E} \left( \xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a'_i} \cdot \prod_{j=1}^{r'} I_{\mu_j}^{\gamma_j n} \right).$$

Здесь  $0 \leq a'_i \leq a_i$ ,  $r \leq r' \leq m$ ,  $\mu_1 \neq \mu_2 \neq \dots \neq \mu_{r'}$ , а  $\{\gamma_j\} \subseteq \{\alpha_j\} \cup \{\beta_j\}$ .

Избавимся от индикаторов, совершив следующие преобразования (формула полной вероятности):

$$\begin{aligned}
& \mathbb{E} \left( \xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a'_i} \cdot \prod_{j=1}^{r'} I_{\mu_j}^{\gamma_j n} \right) = \\
& = \mathbb{E} \left( \mathbb{E} \left( \xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a'_i} \cdot \prod_{j=1}^{r'} I_{\mu_j}^{\gamma_j n} \mid G_m^{n-1} \right) \right) = \\
& = \mathbb{E} \left( \xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a'_i} \cdot \mathbb{E} \left( \prod_{j=1}^{r'} I_{\mu_j}^{\gamma_j n} \mid G_m^{n-1} \right) \right) = \mathbb{E} \left( \xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a'_i} \cdot \prod_{j=1}^{r'} \frac{d_{\gamma_j}^{n-1}}{D_{n-1}} \right).
\end{aligned}$$

Заметим, что все такие слагаемые мажорируются главным членом

$$A_0 = \frac{m!}{(m-r)!} \xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a_i} \cdot \prod_{j=1}^r \frac{d_{\beta_j}^{n-1}}{D_{n-1}},$$

получающимся при приведении подобных из слагаемых вида

$$\xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a_i} \prod_{j=1}^r I_{\mu_j}^{\beta_j n}.$$

Слагаемые из главного члена образуются, когда в каждой из скобок вида  $(d_{\alpha_i}^{n-1} + \sum_{\mu} I_{\mu}^{\alpha_i n})$  берется  $d_{\alpha_i}^{n-1}$ , а в скобках  $(\sum_{\mu} I_{\mu}^{\beta_r n})$  берется набор независимых индикаторов. Множитель  $\frac{m!}{(m-r)!}$  отвечает за число способов выбрать попарно различные индексы  $\mu_1, \dots, \mu_r$ .

Так как количество остальных слагаемых — константа, зависящая только от  $k$  и  $r$ , то и сумма всех слагаемых оценивается как  $\Theta(1) \cdot A_0$ . Поэтому можно записать следующую оценку:

$$\begin{aligned}
& \mathbb{E} \left( \xi_{n-1} \cdot (d_{\alpha_1}^n)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^n)^{a_k} \cdot e_{\beta_1 n} \cdot \dots \cdot e_{\beta_r n} \right) = \\
& = \mathbb{E} (A_0) \cdot \Theta(1) = \\
& = \mathbb{E} \left( \xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a_i} \cdot \prod_{j=1}^r \frac{d_{\beta_j}^{n-1}}{D_{n-1}} \right) \cdot \frac{m!}{(m-r)!} \cdot \Theta(1) = \\
& = \mathbb{E} \left( \xi_{n-1} \prod_{i=1}^k (d_{\alpha_i}^{n-1})^{a_i} \cdot \prod_{j=1}^r d_{\beta_j}^{n-1} \right) \left( \frac{1}{(2n-1)m} \right)^r \frac{m!}{(m-r)!} \cdot \Theta(1) = \\
& = \mathbb{E} \left( \xi_{n-1} \cdot (d_{\alpha_1}^{n-1})^{a_1} \cdot \dots \cdot (d_{\alpha_k}^{n-1})^{a_k} \cdot d_{\beta_1}^{n-1} \cdot \dots \cdot d_{\beta_r}^{n-1} \right) \left( \frac{1}{n} \right)^r \Theta(1).
\end{aligned}$$

На этом теорема доказана.

#### 1.2.4 Доказательство теоремы о длинном спуске

Основные идеи доказательства (переход к индикаторам и избавление от них при помощи формулы полной вероятности) совпадают с идеями доказательства теоремы о коротком спуске. Точно так же основной вклад даст главный член, но окажется, что количество побочных членов по порядку может совпадать с  $n$ , поэтому надо будет пользоваться более тонкой техникой для их оценки.

Будем доказывать теорему индукцией по  $S = \sum_{i=1}^k a_i$ . Предположение индукции состоит в том, что для всех таких  $S$ , что  $k \leq S < S_0$ , выполнено

$$\begin{aligned} & \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^n)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^n)^{a_k} \right) = \\ & = \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^l)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^l)^{a_k} \right) \left( \frac{n}{l} \right)^{\frac{\sum a_i}{2}} \cdot \Theta(1). \end{aligned}$$

**Доказательство базы индукции** В качестве базы возьмем случай, когда для всех  $i$  выполнено  $a_i = 1$  и  $S_0 = k + 1$ . Подставляя индикаторы, имеем

$$\begin{aligned} & \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^n) \cdot \dots \cdot (d_{\alpha_k}^n) \right) = \\ & = \mathbb{E} \left( \xi_l \cdot \left( d_{\alpha_1}^{n-1} + \sum_{\mu} I_{\mu}^{\alpha_1 n} \right) \cdot \dots \cdot \left( d_{\alpha_k}^{n-1} + \sum_{\mu} I_{\mu}^{\alpha_k n} \right) \right). \end{aligned}$$

Раскрыв скобки, получим сумму слагаемых вида

$$\mathbb{E} \left( \xi_l \cdot \prod_{i=1}^{k-t} d_{\beta_i}^{n-1} \cdot \prod_{j=1}^t I_{\mu_j}^{\gamma_j n} \right).$$

Здесь  $0 \leq t \leq k$ ,  $\mu_1 \neq \mu_2 \neq \dots \neq \mu_t$ , а  $\{\beta_i\} \sqcup \{\gamma_i\} = \{\alpha_i\}$ .

Каждое слагаемое такого вида преобразуем с помощью формулы полной вероятности. Слагаемые с одинаковым значением  $t$  совпадут, их сумма равна

$$\mathbb{E} \left( \xi_l \underbrace{A_m^t}_{\text{выбор } \mu_j} \underbrace{A_k^t}_{\text{выбор } \gamma_j} \left( \frac{1}{D_{n-1}} \right)^t \prod_{i=1}^k d_{\alpha_i}^{n-1} \right).$$

Здесь используется стандартное обозначение  $A_a^b = a!/(a-b)!$ .

Суммируя по  $t$ , мы получаем рекуррентное соотношение

$$\begin{aligned} & \mathbb{E} (\xi_l \cdot (d_{\alpha_1}^n) \cdot \dots \cdot (d_{\alpha_k}^n)) = \\ & = \mathbb{E} (\xi_l \cdot (d_{\alpha_1}^{n-1}) \cdot \dots \cdot (d_{\alpha_k}^{n-1})) \cdot \left( 1 + \sum_{t=1}^m A_m^t A_k^t \left( \frac{1}{D_{n-1}} \right)^t \right). \end{aligned}$$

Применив его  $(n-l)$  раз, получим

$$\begin{aligned} & \mathbb{E} (\xi_l \cdot (d_{\alpha_1}^n) \cdot \dots \cdot (d_{\alpha_k}^n)) = \\ & = \mathbb{E} (\xi_l \cdot (d_{\alpha_1}^l) \cdot \dots \cdot (d_{\alpha_k}^l)) \cdot \prod_{i=l}^{n-1} \left( 1 + \sum_{t=1}^m A_m^t A_k^t \left( \frac{1}{D_i} \right)^t \right). \end{aligned}$$

Для доказательства базы индукции осталось оценить величину

$$\prod_{i=l}^{n-1} \left( 1 + \sum_{t=1}^m A_m^t A_k^t \left( \frac{1}{D_i} \right)^t \right).$$

Проделаем следующие преобразования:

$$\begin{aligned} & \prod_{i=l}^{n-1} \left( 1 + \sum_{t=1}^m A_m^t A_k^t \left( \frac{1}{D_i} \right)^t \right) = \exp \left( \ln \left( \prod_{i=l}^{n-1} \left( 1 + \sum_{t=1}^m A_m^t A_k^t \left( \frac{1}{D_i} \right)^t \right) \right) \right) = \\ & = \exp \left( \sum_{i=l}^{n-1} \ln \left( 1 + \sum_{t=1}^m A_m^t A_k^t \left( \frac{1}{(2i+1)m} \right)^t \right) \right) = \\ & = \exp \left( \sum_{i=l}^{n-1} \ln \left( 1 + mk \frac{1}{(2i+1)m} + O \left( \frac{1}{i^2} \right) \right) \right) = \end{aligned}$$

(воспользуемся приближением  $\ln(1+x) = x + O(x^2)$ )

$$\begin{aligned} & = \exp \left( \sum_{i=l}^{n-1} \left( \frac{k}{(2i+1)} + O \left( \frac{1}{i^2} \right) \right) \right) = \\ & = \exp \left( \frac{k}{2} \ln \left( \frac{n}{l} \right) + O \left( \frac{1}{l} \right) \right) = \\ & = \left( \frac{n}{l} \right)^{\frac{k}{2}} \exp \left( O \left( \frac{1}{l} \right) \right) = \left( \frac{n}{l} \right)^{\frac{k}{2}} \left( 1 + O \left( \frac{1}{l} \right) \right). \end{aligned}$$

База индукции доказана. Докажем шаг индукции.

**Доказательство шага индукции** Нужно доказать, что теорема также справедлива для всех таких  $a_1, a_2, \dots, a_k$ , что  $\sum_{i=1}^k a_i = S_0$ .

Как обычно, в левую часть равенства из формулировки теоремы подставим

$$d_{\alpha_i}^n = d_{\alpha_i}^{n-1} + \sum_{\mu} I_{\mu}^{\alpha_i n}$$

и раскроем скобки. Получившиеся слагаемые можно разделить на два вида. *Хорошие* слагаемые — это те, в которых не было совпавших индикаторов, и *плохие* слагаемые — это те, в которых были совпавшие индикаторы. В доказанной нами базе индукции все слагаемые были хорошими. После применения формулы полной вероятности имеем

$$\begin{aligned} & \underbrace{\mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^n)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^n)^{a_k} \right)}_{P_n} = \\ & = \underbrace{\mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^{n-1})^{a_1} \cdot \dots \cdot (d_{\alpha_k}^{n-1})^{a_k} \right)}_{P_{n-1}} \cdot \underbrace{\left( 1 + \sum_{t=1}^m \left( A_m^t A_{S_0}^t \left( \frac{1}{D_{n-1}} \right)^t \right) \right)}_{Q_{n-1}} + \\ & + \sum_{a'_1, a'_2, \dots, a'_k} \underbrace{\mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^{n-1})^{a'_1} \cdot \dots \cdot (d_{\alpha_k}^{n-1})^{a'_k} \right) \cdot \frac{A(a'_1, \dots, a'_k)}{(D_{n-1})^{\#\{a'_i \neq a_i\}}} \left( 1 + O\left(\frac{1}{n}\right) \right)}_{R_{n-1}(a'_1, a'_2, \dots, a'_k)}. \end{aligned} \quad (1.8)$$

Здесь  $0 \leq a'_i \leq a_i$ , причем  $\sum a'_i < \sum a_i$ , а  $A(a'_1, \dots, a'_k)$  — функции, не зависящие от  $n$ , но зависящие от  $a'_1, \dots, a'_k, a_1, \dots, a_k, m, k$ . Сумма *хороших* слагаемых дала главный член  $P_{n-1} \cdot Q_{n-1}$ ; сумма *плохих* слагаемых дала побочный член  $\sum_{a'_1, a'_2, \dots, a'_k} R_{n-1}(a'_1, a'_2, \dots, a'_k)$ .

Поясним подробнее структуру слагаемых  $R_{n-1}(a'_1, a'_2, \dots, a'_k)$ . Величина  $\#\{a'_i \neq a_i\}$  равняется минимальному числу независимых индикаторов в слагаемом, в котором после применения формулы полной вероятности получается произведение  $(d_{\alpha_1}^{n-1})^{a'_1} \cdot \dots \cdot (d_{\alpha_k}^{n-1})^{a'_k}$ . Это минимальное число независимых индикаторов достигается, когда для каждого  $\alpha_i$  совпавшие индикаторы  $I(\langle \alpha_i; n \rangle)_{\mu}$  имеют одинаковое  $\mu$ . Все слагаемые с бóльшим числом независимых индикаторов имеют бóльшую степень множителя  $\frac{1}{D_{n-1}}$  и учтены за счет домножения на  $1 + O\left(\frac{1}{n}\right)$ .



Равенство (1.8) можно переписать как следующее рекуррентное соотношение:

$$P_n = P_{n-1} \cdot Q_{n-1} + \sum_{a'_1, a'_2, \dots, a'_k} R_{n-1}(a'_1, a'_2, \dots, a'_k). \quad (1.9)$$

Применив это рекуррентное соотношение  $n - l$  раз, получим

$$P_n = P_l \prod_{i=l}^{n-1} Q_i + \sum_{i=l}^{n-1} \left( \prod_{j=i}^{n-1} Q_j \sum_{a'_1, a'_2, \dots, a'_k} R_i(a'_1, a'_2, \dots, a'_k) \right). \quad (1.10)$$

Для величины  $\prod_{j=l'}^{n-1} Q_j$  справедливо следующее выражение (см. доказательство шага индукции):

$$\prod_{j=l'}^{n-1} Q_j = \prod_{j=l'}^{n-1} \left( 1 + \sum_{t=1}^m \left( A_m^t A_S^t \left( \frac{1}{D_j} \right)^t \right) \right) = \left( \frac{n}{l'} \right)^{\frac{s_0}{2}} \left( 1 + O\left( \frac{1}{l'} \right) \right). \quad (1.11)$$

Отсюда для главного члена в (1.10) имеем

$$P_l \prod_{i=l}^{n-1} Q_i = E \left( \xi_l \cdot (d_{\alpha_1}^l)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^l)^{a_k} \right) \left( \frac{n}{l} \right)^{\frac{s_0}{2}} \left( 1 + O\left( \frac{1}{l} \right) \right).$$

Для доказательства теоремы осталось оценить побочный член в (1.10). Меняя в нем порядок суммирования, получаем

$$\begin{aligned} & \sum_{i=l}^{n-1} \left( \prod_{j=i}^{n-1} Q_j \sum_{a'_1, a'_2, \dots, a'_k} R_i(a'_1, a'_2, \dots, a'_k) \right) = \\ & = \sum_{a'_1, a'_2, \dots, a'_k} \sum_{i=l}^{n-1} \left( \prod_{j=i}^{n-1} Q_j \cdot R_i(a'_1, a'_2, \dots, a'_k) \right) = \end{aligned}$$

(подставим (1.11))

$$= \sum_{a'_1, a'_2, \dots, a'_k} \sum_{i=l}^{n-1} \left( \left( \frac{n}{i} \right)^{\frac{s_0}{2}} \left( 1 + O\left( \frac{1}{i} \right) \right) \cdot R_i(a'_1, a'_2, \dots, a'_k) \right) =$$

(подставим  $R_i(a'_1, a'_2, \dots, a'_k)$ )

$$\begin{aligned}
&= \sum_{a'_1, a'_2, \dots, a'_k} A(a'_1, \dots, a'_k) \cdot \sum_{i=l}^{n-1} \left( \mathbf{E} \left( \xi_l \cdot (d_{\alpha_1}^i)^{a'_1} \cdot \dots \cdot (d_{\alpha_k}^i)^{a'_k} \right) \left( \frac{n}{i} \right)^{\frac{\sum a_j}{2}} \right. \\
&\quad \left. \cdot \left( 1 + O \left( \frac{1}{i} \right) \right) \left( \frac{1}{D_i} \right)^{\#\{a'_i \neq a_i\}} \right) = \\
&\text{(воспользуемся предположением индукции, т.к. } \sum a'_j < \sum a_j = S_0) \\
&= \sum_{a'_1, a'_2, \dots, a'_k} A(a'_1, \dots, a'_k) \Theta(1) \mathbf{E} \left( \xi_l \cdot (d_{\alpha_1}^l)^{a'_1} \cdot \dots \cdot (d_{\alpha_k}^l)^{a'_k} \right) \cdot \\
&\quad \cdot \underbrace{\sum_{i=l}^{n-1} \left( \frac{i}{l} \right)^{\frac{\sum a'_j}{2}} \left( \frac{n}{i} \right)^{\frac{\sum a_j}{2}} \left( \frac{1}{D_i} \right)^{\#\{a'_i \neq a_i\}} \left( 1 + O \left( \frac{1}{i} \right) \right)}_{(**)}. \tag{1.12}
\end{aligned}$$

Для (\*\*\*) имеем

$$(***) = \left( \frac{n^{\frac{\sum a_j}{2}}}{l^{\frac{\sum a'_j}{2}}} \right) \left( \frac{1}{2m} \right)^{\#\{a'_i \neq a_i\}} \cdot \sum_{i=l}^{n-1} \left( \frac{1}{i^{\#\{a'_i \neq a_i\} + \sum (a_j - a'_j)/2}} \left( 1 + O \left( \frac{1}{i} \right) \right) \right) =$$

(оценим степенной ряд, имея в виду, что  $\#\{a'_i \neq a_i\} + \sum (a_j - a'_j)/2 \geq 1 + 1/2$ )

$$\begin{aligned}
&= \left( \frac{n^{\frac{\sum a_j}{2}}}{l^{\frac{\sum a'_j}{2}}} \right) \left( \frac{1}{2m} \right)^{\#\{a'_i \neq a_i\}} \cdot \left( \frac{1}{l^{\#\{a'_i \neq a_i\} - 1 + \sum (a_j - a'_j)/2}} \right) \left( 1 + O \left( \frac{1}{l} \right) \right) = \\
&= \left( \frac{n}{l} \right)^{\frac{\sum a_j}{2}} \left( \frac{1}{2m} \right)^{\#\{a'_i \neq a_i\}} \left( \frac{1}{l^{\#\{a'_i \neq a_i\} - 1}} \right) \left( 1 + O \left( \frac{1}{l} \right) \right).
\end{aligned}$$

Подставим в (1.12) полученное равенство:

$$\begin{aligned}
(1.12) &= \sum_{a'_1, a'_2, \dots, a'_k} A(a'_1, \dots, a'_k) \Theta(1) \mathbf{E} \left( \xi_l \cdot (d_{\alpha_1}^l)^{a'_1} \cdot \dots \cdot (d_{\alpha_k}^l)^{a'_k} \right) \cdot \\
&\quad \cdot \left( \frac{n}{l} \right)^{\frac{\sum a_j}{2}} \left( \frac{1}{2m} \right)^{\#\{a'_i \neq a_i\}} \left( \frac{1}{l^{\#\{a'_i \neq a_i\} - 1}} \right) \left( 1 + O \left( \frac{1}{l} \right) \right) =
\end{aligned}$$

(видим, что слагаемые с  $\#\{a'_i \neq a_i\} > 1$  имеют большую степень по  $\frac{1}{l}$  и могут быть занесены в  $O\left(\frac{1}{l}\right)$  )

$$= \sum_{\substack{a'_1, a'_2, \dots, a'_k: \\ \#\{a'_i \neq a_i\}=1}} A(a'_1, \dots, a'_k) \Theta(1) \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^l)^{a'_1} \cdot \dots \cdot (d_{\alpha_k}^l)^{a'_k} \right) \cdot \left( \frac{n}{l} \right)^{\frac{\sum a_j}{2}} \left( \frac{1}{2m} \right) \left( 1 + O\left(\frac{1}{l}\right) \right). \quad (1.13)$$

Подставим

$$A(a_1, \dots, a'_t, \dots, a_k) = \underbrace{C_{a_t}^{a'_t-1}}_{\text{выбор скобок, из которых берутся индикаторы}} \cdot \underbrace{C_m^1}_{\text{выбор } \mu}.$$

Получим

$$\begin{aligned} (1.13) &= \Theta(1) \sum_{t=1}^k \sum_{a'_t < a_t} C_{a_t}^{a'_t-1} \cdot m \cdot \\ &\cdot \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^l)^{a_1} \cdot \dots \cdot (d_{\alpha_1}^l)^{a'_t} \cdot \dots \cdot (d_{\alpha_k}^l)^{a_k} \right) \left( \frac{n}{l} \right)^{\frac{\sum a_j}{2}} \left( \frac{1}{2m} \right) \left( 1 + O\left(\frac{1}{l}\right) \right) \leq \\ &\leq \Theta(1) \sum_{t=1}^k \sum_{a'_t < a_t} C_{a_t}^{a'_t-1} \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^l)^{a_1} \cdot \dots \cdot (d_{\alpha_1}^l)^{a_t} \cdot \dots \cdot (d_{\alpha_k}^l)^{a_k} \right) \cdot \\ &\cdot \left( \frac{n}{l} \right)^{\frac{\sum a_j}{2}} \left( 1 + O\left(\frac{1}{l}\right) \right) = \mathbb{E} \left( \xi_l \cdot (d_{\alpha_1}^l)^{a_1} \cdot \dots \cdot (d_{\alpha_k}^l)^{a_k} \right) \left( \frac{n}{l} \right)^{\frac{\sum a_j}{2}} \cdot \Theta(1). \end{aligned}$$

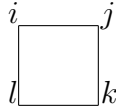
Мы получили, что вклад побочных членов не превосходит вклада главного члена. Теорема доказана.

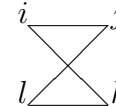
### 1.2.5 Доказательство теоремы о произвольном подграфе

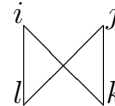
Выражения, подобные (1.1) и (1.6), могут быть записаны и для произвольного подграфа. Общий случай отличается тем, что произвольный подграф  $G_0$  может быть вложен в  $G_m^n$  разными способами, имеющими одно и то же множество вершин, но разные множества ребер. Рассмотрим, например, случай  $G_0 = K_{2,2}$ , который мало отличается от случая произвольного  $G_0$ .

В этом случае

$$\#(K_{2,2}, G_m^n) = \sum_{1 \leq i < j < k < l \leq n} \left( \underbrace{e_{ij}e_{jk}e_{kl}e_{li}}_{\downarrow} + \underbrace{e_{ik}e_{jl}e_{ij}e_{kl}}_{\downarrow} + \underbrace{e_{ik}e_{jk}e_{jl}e_{il}}_{\downarrow} \right).$$







Перейдем теперь к случаю произвольного  $G_0$ . Обозначим через  $r$  количество вершин в  $G_0$ . Пусть  $\Omega$  — это множество всех возможных отображений, под действием каждого из которых граф  $G_0$  вкладывается в клику размера  $r$  с вершинами  $\{1, \dots, r\}$ . Пусть  $E(\omega(G_0))$  — это множество ребер графа  $\omega(G_0)$  для  $\omega \in \Omega$ . Тогда

$$\mathbb{E}(\#(G_0, G_m^n)) = \mathbb{E} \left( \sum_{\omega \in \Omega} \left( \sum_{1 \leq i_1 < \dots < i_r \leq n} \left( \prod_{(p,q) \in E(\omega(G_0))} e_{i_p i_q} \right) \right) \right).$$

Будем применять к этому выражению теоремы о коротком и длинном спуске, подобно тому, как делали это для тетраэдра. Заметим, что в итоге каждый множитель  $e_{i_p i_q}$  перейдет в множитель  $\sqrt{\frac{m}{i_p}} \cdot \sqrt{\frac{m}{i_q}} \cdot \Theta(1)$ . Действительно, пусть, например,  $i_p > i_q$ , тогда в одном из коротких спусков  $e_{i_p i_q}$  перейдет в  $\frac{1}{i_p} \cdot d_{i_q}^{i_p-1} \cdot \Theta(1)$ . После длинных спусков от  $p$  до  $t_1$ , от  $t_1$  до  $t_2$ ,  $\dots$ , от  $t_n$  до  $q$  получим, что  $\frac{1}{i_p} \cdot d_{i_q}^{i_p-1} \cdot \Theta(1)$  перейдет в  $\frac{1}{i_p} \cdot \sqrt{\frac{i_p}{i_{t_1}} \frac{i_{t_1}}{i_{t_2}} \dots \frac{i_{t_n}}{i_q}} \cdot d_{i_q}^{i_q} \cdot \Theta(1) = \sqrt{\frac{m}{i_p}} \cdot \sqrt{\frac{m}{i_q}} \cdot \Theta(1)$ , т.к.  $m \leq d_{i_q}^{i_q} \leq 2m$ . Таким образом, имеем

$$\begin{aligned} \mathbb{E}(\#(G_0, G_m^n)) &= \Theta(1) \cdot \sum_{\omega \in \Omega} \left( \sum_{1 \leq i_1 < \dots < i_r \leq n} \left( \prod_{(p,q) \in E(\omega(G_0))} \sqrt{\frac{m}{i_p}} \cdot \sqrt{\frac{m}{i_q}} \right) \right) = \\ &= \Theta(1) \cdot \sum_{\omega \in \Omega} \left( \sum_{1 \leq i_1 < \dots < i_r \leq n} \left( \prod_{p \in \{1, \dots, r\}} \left( \sqrt{\frac{m}{i_p}} \right)^{d_{\omega(G_0)}(p)} \right) \right) = \end{aligned}$$

$$\begin{aligned}
&= \Theta(1) \cdot \sum_{\omega \in \Omega} \left( \int_{i_1=1}^n \cdots \int_{i_r=1}^n \left( \prod_{p \in \{1 \dots r\}} \left( \sqrt{\frac{m}{i_p}} \right)^{d_{\omega(G_0)}(p)} \right) \right) = \\
&= \Theta(1) \cdot \sum_{\omega \in \Omega} \left( \prod_{p \in \{1 \dots r\}} \int_{i_p=1}^n \left( \sqrt{\frac{m}{i_p}} \right)^{d_{\omega(G_0)}(p)} \right) = \\
&= \Theta(1) \cdot \prod_{\tilde{p} \in \{1 \dots r\}} \int_{i_{\tilde{p}}=1}^n \left( \sqrt{\frac{m}{i_{\tilde{p}}}} \right)^{d_{G_0}(\tilde{p})}.
\end{aligned}$$

Все величины, не зависящие от  $m$  и  $n$ , мы заносили в  $\Theta(1)$ . После взятия интеграла получаем требуемое

$$\Theta \left( n^{\#(d_i=0)} \cdot (\sqrt{n})^{\#(d_i=1)} \cdot (\ln n)^{\#(d_i=2)} \cdot m^{\frac{\sum d_i}{2}} \right).$$

Теорема доказана.

## 1.3 Обобщенное предпочтительное присоединение

В этом разделе мы предлагаем общий подход к моделям предпочтительного присоединения и доказываем теоремы, нацеленные на упрощение анализа их свойств.

### 1.3.1 Определение $PA$ -класса

Пусть  $G_m^n$  ( $n \geq n_0$ ) – это граф с  $n$  вершинами  $\{1, \dots, n\}$  и  $mn$  ребрами, построенный в результате следующего случайного процесса. Мы стартуем в момент времени  $n_0$  с произвольного графа  $G_m^{n_0}$  с  $n_0$  вершинами и  $mn_0$  ребрами. На  $(n+1)$ -ом шаге ( $n \geq n_0$ ) мы строим граф  $G_m^{n+1}$  из графа  $G_m^n$  путем добавления новой вершины  $n+1$  и  $m$  ребер, соединяющих эту вершину и некоторые  $m$  вершин из множества  $\{1, \dots, n, n+1\}$ . Обозначим через  $d_v^n$  степень вершины  $v$  в графе  $G_m^n$ . Если для некоторых констант  $A$  и  $B$  выполняются следующие условия

$$\mathbb{P}(d_v^{n+1} = d_v^n \mid G_m^n) = 1 - A \frac{d_v^n}{n} - B \frac{1}{n} + O\left(\frac{(d_v^n)^2}{n^2}\right), \quad 1 \leq v \leq n, \quad (1.14)$$

$$\mathbb{P}(d_v^{n+1} = d_v^n + 1 \mid G_m^n) = A \frac{d_v^n}{n} + B \frac{1}{n} + O\left(\frac{(d_v^n)^2}{n^2}\right), \quad 1 \leq v \leq n, \quad (1.15)$$

$$\mathbb{P}(d_v^{n+1} = d_v^n + j \mid G_m^n) = O\left(\frac{(d_v^n)^2}{n^2}\right), \quad 2 \leq j \leq m, \quad 1 \leq v \leq n, \quad (1.16)$$

$$\mathbb{P}(d_{n+1}^{n+1} = m + j) = O\left(\frac{1}{n}\right), \quad 1 \leq j \leq m, \quad (1.17)$$

то мы говорим, что случайный граф  $G_m^n$  – это модель из  $PA$ -класса ( $PA$  от preferential attachment). Условие (1.17) означает, что вероятность образовать петлю мала. Как мы увидим далее, частные детали модели, вроде того, что разрешены ли петли и мультиребра или нет, являются не важными для многих свойств модели.

Так как на каждом шаге мы добавляем  $m$  ребер, суммируя по всем вершинам равенства (1.15)-(1.17) (с соответствующими коэффициентами) и пренебрегая остаточными членами, мы получаем  $2mA + B = m$ . Можно доказать, что в данном случае сумма остаточных членов равна 0, но для простоты мы полагаем  $2mA + B = m$ . Далее, мы имеем  $0 \leq A \leq 1$  (для (1.15) нам нужно  $mA + B \geq 0$ , и мы положили  $2mA + B = m$ , поэтому  $A \leq 1$ ).

Здесь хочется подчеркнуть, что мы определили не одну модель, а целый класс моделей. Даже задание конкретных значений параметров  $A$  и  $m$  не определяет полностью процедуру построения графа. То, чего не хватает в определении, это конкретное вероятностное распределение на наборах из  $m$  вершин, с которыми будет соединена добавляемая вершина. Таким образом, существует множество моделей с разными свойствами, удовлетворяющих условиям (1.14–1.17). Например, LCD-модель, модель Холм–Кима и RAN-модель принадлежат  $PA$ -классу с параметрами  $A = 1/2$  и  $B = 0$ . Модели Бакли–Остуса и Мори также принадлежат  $PA$ -классу с параметрами  $A = \frac{1}{2+\beta}$  и  $B = \frac{m\beta}{2+\beta}$ . Еще один пример подробно рассмотрен в параграфах 1.3.4 и 1.3.6. Заметим, что наш класс моделей шире класса Барабаши–Алберт, так как в нем мы имеем настраиваемый параметр степенного закона распределения степеней вершин.

В математическом анализе моделей сложных сетей имеется тенденция рассматривать конкретные модели или параметризованные семейства моделей. Мы же в следующих двух разделах получим результаты об общих свойствах всего  $PA$ -класса, заданного в терминах соотношений на вероятности событий, оставляя тем самым большую свободу в точном определении модели.

### 1.3.2 Степенной закон распределения степеней вершин

В параграфах 1.3.2 и 1.3.3 мы используем следующие обозначения. Под **whp** (“with high probability”), мы имеем в виду, что для некоторой последовательности событий  $A_n$  выполнено  $P(A_n) \rightarrow 1$ , когда  $n \rightarrow \infty$ . Мы говорим, что  $a_n \sim b_n$ , если  $a_n = (1 + o(1))b_n$ , и  $a_n \approx b_n$ , если  $C_0b_n \leq a_n \leq C_1b_n$  для некоторых констант  $C_0, C_1 > 0$ . Мы говорим, что **whp**  $a_n \sim b_n$ , если  $\exists \phi : \phi = o(1)$  и  $P(a_n = (1 + \phi(n))b_n) \rightarrow 1$ , когда  $n \rightarrow \infty$ .

Несмотря на то, что нам не задано точное распределение на наборах вершин, в которые добавляемая вершина проводит ребра, мы все же можем описать распределение степеней вершин в графе.

Сначала мы оценим  $N_n(d)$  — количество вершин степени  $d$  в графе  $G_m^n$ . Мы докажем следующий результат о математическом ожидании  $\mathbf{E}N_n(d)$  случайной величины  $N_n(d)$ .

**Теорема 10.** Пусть  $d \geq m$ . Тогда

$$\mathbf{E}N_n(d) = c(m, d) \left( n + O\left(d^{2+\frac{1}{A}}\right) \right),$$

где

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

а  $\Gamma(x)$  — это гамма-функция.

Во-вторых, мы докажем, что количество вершин степени  $d$  плотно сконцентрировано около своего математического ожидания.

**Теорема 11.** Для каждой целочисленной ненулевой функции  $d = d(n)$  мы имеем

$$\mathbf{P}\left(|N_n(d) - \mathbf{E}N_n(d)| \geq d \sqrt{n} \ln n\right) = O\left(n^{-\ln n}\right),$$

откуда для любого  $\delta > 0$  существует такая функция  $\varphi(n) = o(1)$ , что **whp** для любого  $d \leq n^{\frac{A-\delta}{4A+2}}$  выполнено

$$|N_n(d) - \mathbf{E}N_n(d)| \leq \varphi(n) \mathbf{E}N_n(d).$$

Эти две теоремы означают, что распределение степеней вершин подчиняется (асимптотически) степенному закону с параметром  $1 + \frac{1}{A}$ .

Теорема 10 доказывается индукцией по  $d$  и  $n$ . Легко видеть, что при условии графа  $G_m^n$  мы можем выразить условное математическое ожидание числа вершин степени  $d$  в графе  $G_m^{n+1}$  (т.е.,  $\mathbf{E}(N_{n+1}(d) \mid G_m^n)$ ) в терминах  $N_n(d), N_n(d-1), \dots, N_n(d-m)$ . Здесь мы используем тот факт, что вероятность ребра между вершинами  $n+1$  и  $v$  зависит только от степени вершины  $v$  (см. (1.14)). Используя формулу полной вероятности, мы получаем рекуррентное соотношение для  $\mathbf{E}N_{n+1}(d)$  и доказываем утверждение теоремы 10 по индукции.

Мы используем неравенство Азумы–Хеффдинга, чтобы доказать концентрационный результат из теоремы 11. Для этого мы рассматриваем мартингал  $X_i(d) = \mathbf{E}(N_n(d) \mid G_m^i)$ ,  $i = 0, \dots, n$ . Полные доказательства этих теорем помещены в раздел 1.3.8.

### 1.3.3 Кластерный коэффициент

В этом разделе мы рассмотрим кластерный коэффициент моделей из  $PA$ -класса. Существует два популярных определения кластерного коэффициента. *Глобальный кластерный коэффициент*  $C_1(n)$  графа  $G_m^n$  — это отношение



утроенного числа треугольников к числу пар примыкающих ребер в графе  $G_m^n$ . Средний локальный кластерный коэффициент определяется следующим образом:  $C_2(n) = \frac{1}{n} \sum_{i=1}^n C(i)$ , где  $C(i)$  — это локальный кластерный коэффициент вершины  $i$ :  $C(i) = \frac{T^i}{P_2^i}$ , где  $T^i$  — это количество ребер между соседями вершины  $i$  и  $P_2^i$  — это общее число разных пар соседей. Результаты касательно кластерных коэффициентов для некоторых классических моделей (LCD и Мори) упомянуты в разделе 1.1.

Здесь мы обобщаем эти результаты. Сначала мы изучаем случайную величину  $P_2(n)$ , равную количеству  $P_2$  (пар смежных ребер) в случайном графе  $G_m^n$ , для произвольной модели из  $PA$ -класса.

**Теорема 12.** *Для каждой модели из  $PA$ -класса мы имеем*

- (1) Если  $2A < 1$ , то **whp**  $P_2(n) \sim \left(2m(A + B) + \frac{m(m-1)}{2}\right) \frac{n}{1-2A}$ ,
- (2) Если  $2A = 1$ , то **whp**  $P_2(n) \sim \left(2m(A + B) + \frac{m(m-1)}{2}\right) n \ln n$ ,
- (3) Если  $2A > 1$ , то для любого  $\varepsilon > 0$  **whp**  $n^{2A-\varepsilon} \leq P_2(n) \leq n^{2A+\varepsilon}$ .

Доказательство теоремы 12 в некотором приближении приведено в разделе 1.3.8. Мы доказываем теорему 12 не полностью строго математически, однако используемая нами аппроксимация точнее метода среднего поля, часто применяемого в физических статьях. Более того, она согласуется с эмпирическими результатами, полученными в разделе 1.3.6. Здесь стоит отметить, что поведение  $P_2(n)$  в масштабно-инвариантных графах обычно определяется показателем степенного закона  $\gamma$ . Действительно, мы имеем  $P_2(n) = \sum_{d=1}^{d_{\max}} N_n(d) \frac{d(d-1)}{2} \approx \sum_{d=1}^{d_{\max}} nd^{2-\gamma}$ , где  $d_{\max}$  — это максимальная степень вершины в графе  $G_m^n$ . Поэтому если  $\gamma > 3$ , то функция  $P_2(n)$  линейна по  $n$ . Однако, если  $\gamma \leq 3$ , то  $P_2(n)$  сверхлинейна.

Далее, мы изучаем случайную величину  $T(n)$ , равную числу треугольников в  $G_m^n$ . Заметим, что для любой модели из  $PA$ -класса мы имеем  $T(n) = O(n)$ , так как на каждом шаге мы добавляем не более  $\frac{m(m-1)}{2}$  треугольников. Если мы соединим этот факт с предыдущим наблюдением, мы увидим, что если  $\gamma \leq 3$ , то в любой модели предпочтительного присоединения (в которой исходящие степени вершин равны  $m$ ) глобальный кластерный коэффициент стремится к нулю с ростом  $n$ .

Наша цель — найти модели с постоянным кластерным коэффициентом. Давайте рассмотрим подкласс  $PA$ -класса со следующим свойством:

$$\mathbb{P}(d_i^{n+1} = d_i^n + 1, d_j^{n+1} = d_j^n + 1 \mid G_m^n) = e_{ij} \frac{D}{mn} + O\left(\frac{d_i^n d_j^n}{n^2}\right). \quad (1.18)$$

Здесь  $e_{ij}$  — это количество ребер между вершинами  $i$  и  $j$  в графе  $G_m^n$ , а  $D$  — положительная константа. Заметим, что это свойство по-прежнему полностью не определяет зависимости между добавляемым ребрами.

**Теорема 13.** Пусть  $G_m^n$  удовлетворяет условию (1.18). Тогда **whp**  $T(n) \sim Dn$ .

Доказательство этой теоремы естественно. Ниже мы приводим рассуждение, которое формально верно, но в котором опущены подробные выкладки. Мы опускаем подробные выкладки, поскольку для нас важна лишь теорема 14, которая также будет приведена ниже, а она будет следствием теорем 12 и 13, причем теорема 12 доказана только в некотором приближении.

Математическое ожидание числа треугольников, которые мы добавляем на каждом шаге равняется  $D + o(1)$ . Действительно, методом математической индукции, с учетом условий (1.14–1.17), несложно получить, что максимальная степень  $\Delta_n$  растет как  $n^A$ . Поэтому сумма побочных членов  $O\left(\frac{d_i^n d_j^n}{n^2}\right)$  по всем парам *соединенных ребром* вершин равна  $o(1)$ . В итоге  $ET(n) = Dn + o(n)$ . Концентрация случайной величины  $T(n)$  около своего математического ожидания следует из неравенства Азумы–Хеффдинга.

Как следствие теорем 12 и 13, мы получаем следующий результат о глобальном кластерном коэффициенте  $C_1(n)$  графа  $G_m^n$ .

**Теорема 14.** Пусть случайный граф  $G_m^n$  принадлежит PA-классу и удовлетворяет (1.18). Тогда

$$(1) \text{ Если } 2A < 1, \text{ то } \mathbf{whp} \ C_1(n) \sim \frac{3(1-2A)D}{(2m(A+B) + \frac{m(m-1)}{2})},$$

$$(2) \text{ Если } 2A = 1, \text{ то } \mathbf{whp} \ C_1(n) \sim \frac{3D}{(2m(A+B) + \frac{m(m-1)}{2}) \ln n},$$

$$(3) \text{ Если } 2A > 1, \text{ то для любого } \varepsilon > 0 \ \mathbf{whp} \ n^{1-2A-\varepsilon} \leq C_1(n) \leq n^{1-2A+\varepsilon}.$$

Теорема 14 показывает, что в некоторых случаях ( $2A \geq 1$ ) глобальный кластерный коэффициент  $C_1(n)$  стремится к нулю с ростом размера графа, даже если выполнено условие (1.18). Средний локальный кластерный коэффициент  $C_2(n)$  в этом случае ведет себя по-другому. Действительно, из теорем 10 и 11 следует, что **whp** количество вершин степени  $m$  в графе  $G_m^n$  больше  $cn$  для некоторой положительной константы  $c$ . Математическое ожидание числа треугольников, добавляемых на каждом шаге, равно  $D + o(1)$ . Поэтому **whp**

$$C_2(n) \geq \frac{1}{n} \sum_{i: \deg(i)=m} C(i) \geq \frac{2cD}{m(m+1)}.$$

В следующем параграфе мы предложим нетривиальную модель из  $PA$ -класса, обладающую более реалистичным поведением глобального кластерного коэффициента и рядом других интересных свойств.

### 1.3.4 Полиномиальная модель

В этом параграфе мы рассмотрим *полиномиальную модель* случайного графа, которая принадлежит общему  $PA$ -классу, определенному выше. Применяя наши теоретические результаты к полиномиальной модели, мы увидим, что эта модель очень гибкая: в ней можно настраивать и параметр степенного закона распределения степеней вершин, и глобальный кластерный коэффициент.

**Определение полиномиальной модели** Сейчас мы определим *полиномиальную модель*. Так же, как и в случайном процессе из параграфа 1.3.1, мы строим граф  $G_m^n$  шаг за шагом. На  $(n + 1)$ -ом шаге граф  $G_m^{n+1}$  образуется из графа  $G_m^n$  путем добавления вершины  $n + 1$  и последовательного проведения  $m$  ребер (кратные ребра и петли разрешаются).

Будем говорить, что ребро  $ij$  направлено от  $i$  к  $j$ , если  $i \geq j$ , таким образом исходящая степень каждой вершины равна  $m$ . Мы также будем говорить, что  $i$  и  $j$  — это соответственно *начало* и *конец* ребра  $ij$ . Рассмотрим три способа провести новые ребра из вершины  $n + 1$ . Сначала мы случайно выбираем в графе  $G_m^n$  ребро равномерно и независимо, и после этого возможны три варианта:

- Preferential attachment (PA): провести ребро из вершины  $n + 1$  в *конец* выбранного ребра
- Uniform (U): провести ребро из вершины  $n + 1$  в *начало* выбранного ребра
- Triangle formation (TF): провести два ребра из вершины  $n + 1$  в *начало* и *конец* выбранного ребра

Определим теперь, как провести все  $m$  ребер из вершины  $n + 1$ . Рассмотрим набор таких неотрицательных параметров  $\{\alpha_{k,l}\}$  для  $0 \leq k \leq m/2$  и  $0 \leq l \leq m - 2k$ , что  $\sum_{k,l} \alpha_{k,l} = 1$ . Эти параметры полностью определяют модель. В начале  $(n + 1)$ -го шага с вероятностью  $\alpha_{k,l}$  мы выбираем некоторые  $k = k_0$  и  $l = l_0$ , затем проводим  $l_0$  ребер по правилу PA,  $2k_0$  ребер по правилу TF и  $(m - l_0 - 2k_0)$  ребер по правилу U. Полиномиальная модель построена. Из определения следует, что граф в такой модели может быть сгенерирован

на компьютере за линейное время. Более того, эта модель принадлежит  $PA$ -классу. Действительно, посредством простых вычислений можно показать, что условия (1.14–1.17) выполняются.

Объясним, почему мы называем описанную модель полиномиальной. Обозначим через  $\widehat{d}_i^n = d_i^n - m$  входящую степень вершины  $i$  в графе  $G_m^n$ . Напомним, что через  $e_{ij}$  мы обозначаем количество ребер между вершинами  $i$  и  $j$ . Для любых  $k, l$ , таких, что  $0 \leq k \leq m/2$  и  $0 \leq l \leq m - 2k$ , положим

$$M_{k,l}^{n,m}(i_1, \dots, i_m) = \frac{1}{n^{m-l-2k}} \prod_{x=1}^k \frac{e_{i_{2x}i_{2x-1}}}{2mn} \prod_{y=2k+1}^{2k+l} \frac{\widehat{d}_{i_y}^n}{mn}.$$

Это моном, зависящий от  $\widehat{d}_{i_y}^n$  и  $e_{i_{2x}i_{2x-1}}$ . Легко видеть, что

$$\begin{aligned} P(\text{ребра } e_1, \dots, e_m \text{ идут в вершины } i_1, \dots, i_m, \text{ соответственно}) &= \\ &= \sum_{k=0}^{m/2} \sum_{l=0}^{m-2k} \alpha_{k,l} M_{k,l}^{n,m}(i_1, \dots, i_m). \end{aligned} \quad (1.19)$$

Многие из моделей являются частными случаями полиномиальной модели. Если мы рассмотрим полином  $\prod_{y=1}^m \frac{\widehat{d}_{i_y}^n + m}{2mn}$ , то мы получим модель, практически идентичную LCD-модели. Модель Бакли–Остуса также может быть интерпретирована в терминах полиномиальной модели.

**Свойства** Легко проверить, что параметры  $\alpha_{k,l}$  из (1.19) и  $A$  из (1.14) связаны следующим образом:

$$A = \sum \alpha_{k,l} \frac{l+k}{m}. \quad (1.20)$$

Это значит, что мы можем получить любое значение  $A \in [0, 1]$  и любой параметр в степенном законе  $\gamma \in (2, \infty)$ . Также заметим, что  $D = \sum_{k,l} k \alpha_{k,l}$ .

В следующем разделе мы экспериментально проанализируем некоторые свойства графов в полиномиальной модели. Мы сгенерируем графы и сравним их свойства с теоретическими предсказаниями.

### 1.3.5 Описание модели, изученной эмпирически

Мы эмпирически изучим полиномиальную модель с  $m = 2p$  и вероятностью провести ребра в вершины  $i_1, \dots, i_{2m}$ , равной

$$\prod_{k=1}^p \left( \alpha \frac{\widehat{d}_{i_{2k}}^n \widehat{d}_{i_{2k-1}}^n}{(mn)^2} + \beta \frac{e_{i_{2k}i_{2k-1}}}{2mn} + \frac{\delta}{n^2} \right).$$

Здесь нам нужно, чтобы  $\alpha, \beta, \delta \geq 0$  и  $\alpha + \beta + \delta = 1$ , поэтому, мы имеем только три независимых параметра:  $m$ ,  $\alpha$  и  $\beta$ . Заметьте, что здесь мы записали модель в симметричной форме, так как забыли о порядке добавляемых ребер.

Наши теоретические результаты дают нам определенные ожидания касательно свойств возникающих при генерации графов. Из (1.20) мы получаем, что  $A = \alpha + \frac{\beta}{2}$ ,  $B = m(\delta - \alpha)$ ,  $D = p\beta = \frac{m\beta}{2}$ , а значит, согласно теореме 10 и теореме 14, мы ожидаем, что

$$C_1(n) \sim \frac{3(1 - 2\alpha - \beta)\beta}{5m - 1 - 2(2m - 1)(2\alpha + \beta)}, \quad \gamma = 1 + \frac{2}{2\alpha + \beta}. \quad (1.21)$$

### 1.3.6 Эмпирические результаты

**Распределение степеней вершин и кластерный коэффициент** Сначала мы изучили два случайных графа с  $n = 10^7$ ,  $m = 2$  и  $A = 0.4$ , взяв  $\alpha = 0.4, \beta = 0$  для первого и  $\alpha = 0, \beta = 0.8$  для второго. Наблюдаемые распределения степеней вершин практически идентичны и имеют ожидаемый параметр  $\gamma = 3.5$ , см. Рис. 1.1а.

В обоих случаях мы также изучили поведение глобального и среднего локального кластерных коэффициентов у сгенерированных графов. Мы строили 40 графов для каждого  $n = \lceil 10^{1+0.06i} \rceil, i = 0, \dots, 100$ , см. Рис. 1.1бс. В первом случае мы наблюдаем  $C_1(n) \rightarrow 0, C_2(n) \rightarrow 0$  (т.к.  $\beta = 0$ ), а во втором случае  $C_1(n) \rightarrow \frac{2}{15}$  (как и ожидалось ввиду (1.21)) и  $C_2(n) \rightarrow \text{const} > 0$ .

Мы также сгенерировали графы с  $n = 10^6, m = 2$  и меняли  $A$  (мы взяли  $\beta = 0.5$  и  $\alpha \in (0, 0.5)$ ). Другими словами, мы фиксировали вероятность добавления треугольника и меняли параметр степенного распределения. Полученные результаты показаны на Рис. 1.2а. Поведения кластерных коэффициентов разительно отличаются. Если  $A$  растет, то растет и  $P_2(n)$  (поэтому  $C_1(n) \rightarrow 0$ ), также растет количество вершин малой степени, которые имеют большой локальный кластерный коэффициент (поэтому  $C_2(n)$  растет).

Чтобы еще больше продемонстрировать разницу между глобальным и средним локальным кластерными коэффициентами, мы сгенерировали граф с  $m = 2, \alpha = 0.5, \beta = 0.2$  и меняли  $n$  (Рис. 1.2б). В этом случае мы имеем  $A = \alpha + \frac{\beta}{2} > 0.5$  и  $C_1(n) \rightarrow 0$ , как и ожидалось. Как бы то ни было, для локального кластерного коэффициента мы получаем  $C_2(n) \rightarrow \text{const} > 0$ .

**Сравнение с другими моделями** Приведенная таблица резюмирует сравнение полиномиальной модели с остальными моделями предпочтительного присоединения, упомянутыми в тексте:

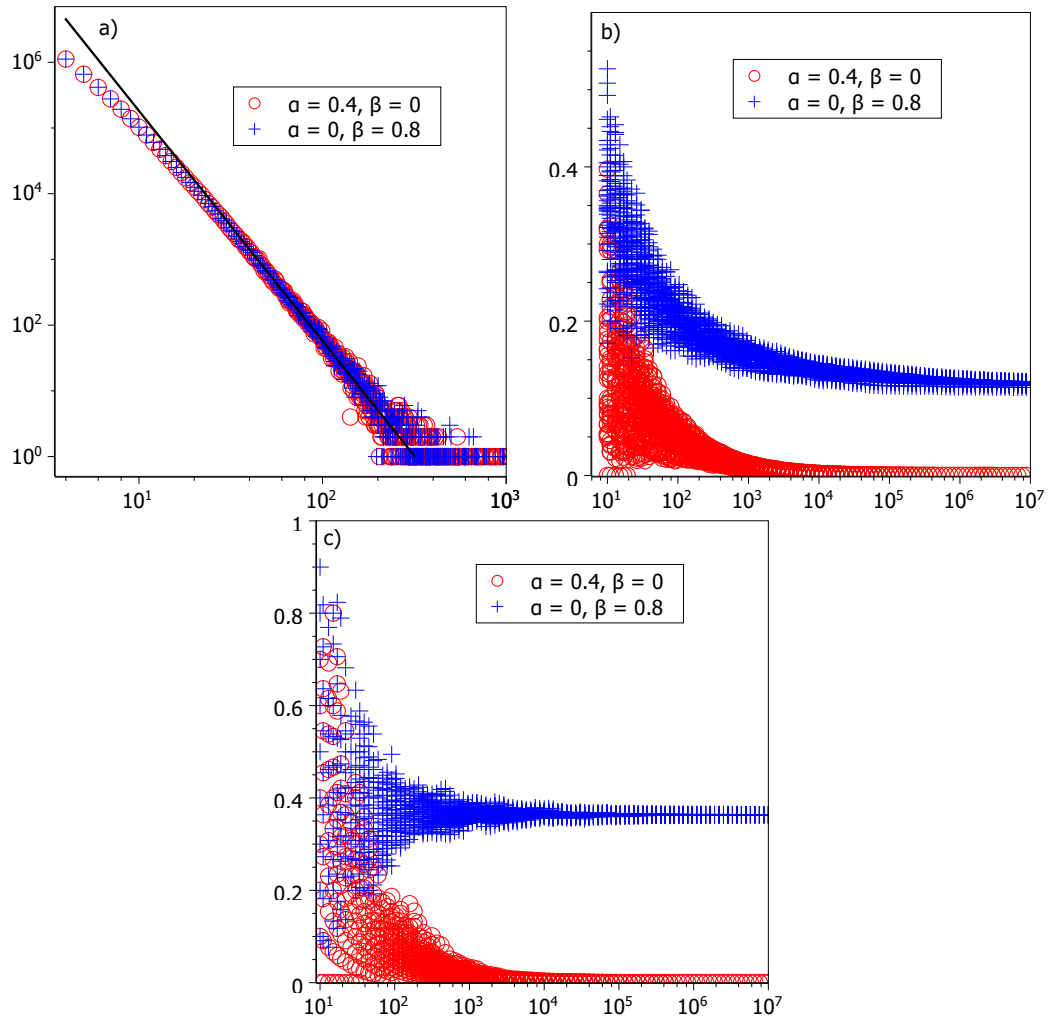


Рис. 1.1: а) Распределение степеней вершин для  $n = 10^7$  и  $m = 2$ . б) Глобальный кластерный коэффициент для  $m = 2$  в зависимости от  $n$ . в) Средний локальный кластерный коэффициент для  $m = 2$  в зависимости от  $n$ .

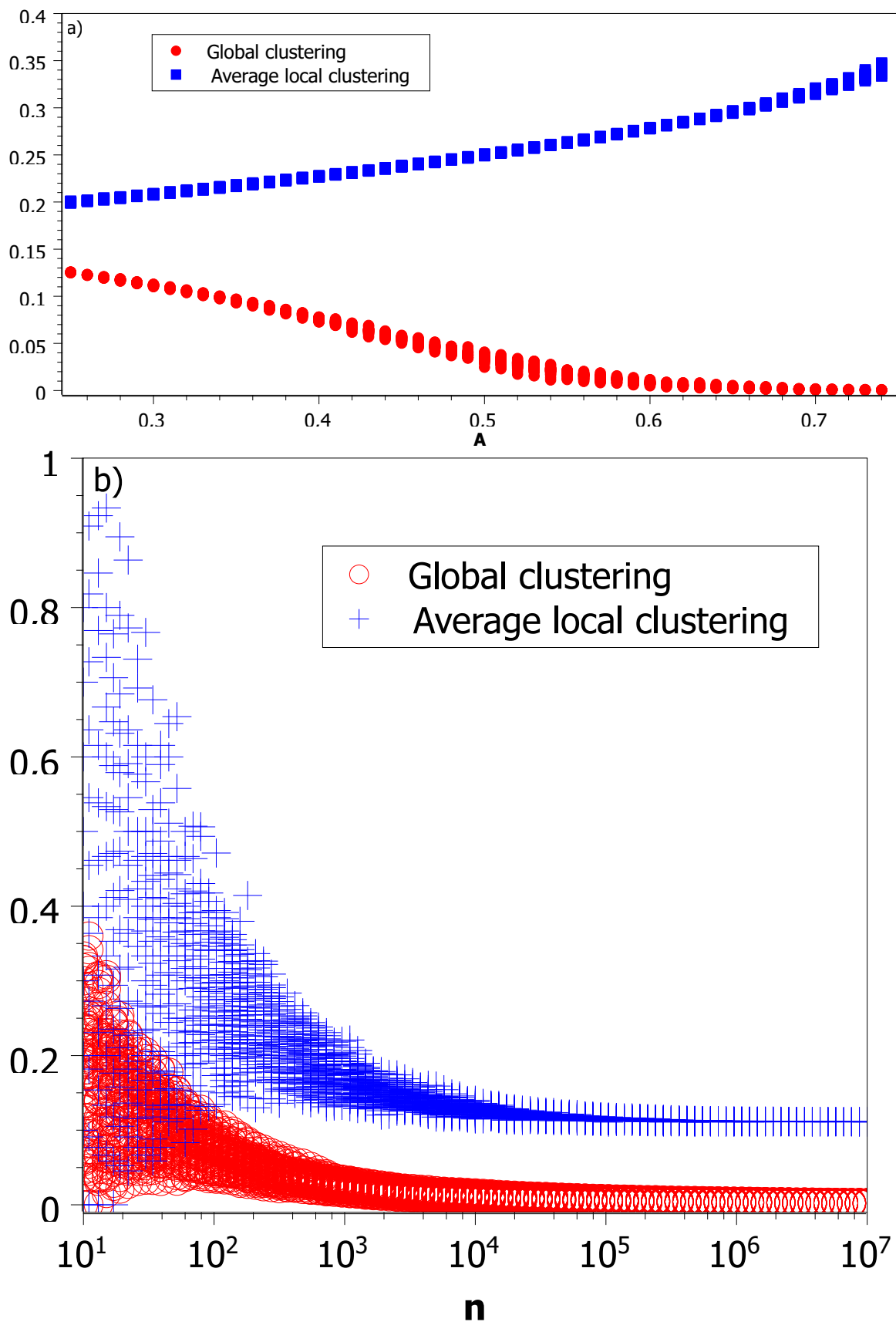


Рис. 1.2: а) Глобальный и средний локальный кластерные коэффициенты для  $n = 10^6$ ,  $m = 2$ ,  $\beta = 0.5$  в зависимости от  $A$ . б) Глобальный и средний локальный кластерные коэффициенты для  $m = 2$ ,  $\alpha = 0.5$ ,  $\beta = 0.2$  в зависимости от  $n$ .

|          | $A$                               | $D$                  | $\gamma$      | $C_1(n)$                    | $C_2(n)$        |
|----------|-----------------------------------|----------------------|---------------|-----------------------------|-----------------|
| LCD      | 1/2                               | 0                    | 3             | $\rightarrow 0$             | $\rightarrow 0$ |
| Мори     | $1/(2 + \beta)$                   | 0                    | $(2, \infty)$ | $\rightarrow 0$             | $\rightarrow 0$ |
| Холм–Ким | 1/2                               | $m_t$                | 3             | $\rightarrow 0$             | $\rightarrow 0$ |
| RAN      | 1/2                               | 3                    | 3             | $\rightarrow 0$             | $> 0$           |
| Полином. | $\sum \alpha_{k,l} \frac{l+k}{m}$ | $\sum k\alpha_{k,l}$ | $(2, \infty)$ | $> 0$ при $A < \frac{1}{2}$ | $> 0$           |

Из приведенных моделей полиномиальная является единственной, в которой можно контролировать параметр степенного закона и в то же время иметь ненулевой глобальный кластерный коэффициент.

### 1.3.7 Обсуждение

В этом разделе мы ввели  $PA$ -класс, который обобщает предыдущие подходы к предпочтительному присоединению. Мы доказали, что любая модель из  $PA$ -класса имеет степенной закон распределения степеней вершин, причем возможны разные значения параметра закона. Мы также оценили кластерные коэффициенты для моделей из этого класса. Затем мы привели одну конкретную модель из предложенного класса, которая имеет настраиваемый параметр распределения степеней и кластерный коэффициент. Компьютерная симуляция иллюстрирует наши теоретические результаты. Мы также отметили, что поведение среднего локального и глобального кластерных коэффициентов разительно отличается для моделей предпочтительного присоединения.

В то время, как в моделях предпочтительного присоединения можно настроить распределение степеней вершин так, чтобы соответствовать реальным сетям, с кластерным коэффициентом в некоторых случаях наблюдаются проблемы. Для большинства реальных сетей параметр  $\gamma$  распределения степеней вершин принадлежит  $[2, 3]$ . Но, как мы показали в параграфе 1.3.3, если  $\gamma \leq 3$ , то в моделях предпочтительного присоединения глобальный кластерный коэффициент стремится к нулю с ростом графа, что не согласуется с поведением большинства реальных сетей. Причина этого состоит в том, что на каждом шаге мы добавляем постоянное число ребер, и поэтому число треугольников растет слишком медленно.

К радости, существует множество способов преодолеть это обстоятельство. Например, Купер предложил модель, в которой количество ребер, добавляемых на каждом шаге, является случайной величиной [21]. Вместе с Пралатом он также рассматривал модель типа Барабаши–Альберт, в которой вершина в момент времени  $t$  создает  $t^c$  ребер [23]. Модели предпочтительного присоединения со случайными начальными степенями были рассмотрены в [24].



Также есть модели, в которых добавляются ребра между старыми вершинами (например, [22]). Использование одной из этих идей в отношении  $PA$ -класса является темой для будущих исследований.

### 1.3.8 Доказательства теорем

#### Доказательство теоремы 10

В этих доказательствах мы используем обозначение  $\theta(\cdot)$  для остаточных членов. Через  $\theta(X)$  мы обозначаем любую такую функцию, что  $|\theta(X)| < X$ . Нам также понадобятся следующие обозначения, мотивированные соотношениями (1.14–1.17):

$$\mathbf{P}(d_v^{n+1} = d \mid d_v^n = d) = 1 - A\frac{d}{n} - B\frac{1}{n} + O\left(\frac{d^2}{n^2}\right), \quad (1.22)$$

$$p_n^1(d) := \mathbf{P}(d_v^{n+1} = d + 1 \mid d_v^n = d) = A\frac{d}{n} + B\frac{1}{n} + O\left(\frac{d^2}{n^2}\right), \quad (1.23)$$

$$p_n^j(d) := \mathbf{P}(d_v^{n+1} = d + j \mid d_v^n = d) = O\left(\frac{d^2}{n^2}\right), \quad 2 \leq j \leq m. \quad (1.24)$$

$$p_n := \sum_{k=1}^m \mathbf{P}(d_{n+1}^{n+1} = m + k) = O\left(\frac{1}{n}\right). \quad (1.25)$$

Заметим, что остаточный член в  $p_n^j(d)$  может зависеть от  $v$ . Мы опускаем  $v$  в обозначении  $p_n^j(d)$ , чтобы не загромождать формулы.

Положим  $p_n(d) = \sum_{j=1}^m p_n^j(d)$ . Заметим, что  $\frac{Ad+B+1}{Ad-A+B}p_n^1(d-1) - p_n(d) = \frac{1}{n} + O\left(\frac{d^2}{n^2}\right)$ . В ходе доказательства мы будем несколько раз пользоваться этим равенством.

Мы хотим доказать, что  $\mathbf{E}N_i(d) = c(m, d) \left(i + \theta\left(Cd^{2+\frac{1}{A}}\right)\right)$  для некоторой константы  $C$  и некоторой функции  $\theta$ . Нам нужны следующие равенства:

$$\mathbf{E}(N_{i+1}(m) \mid N_i(m)) = N_i(m) (1 - p_i(m)) + 1 - p_i, \quad (1.26)$$

$$\begin{aligned} \mathbf{E}(N_{i+1}(d) \mid N_i(d), N_i(d-1), \dots, N_i(d-m)) &= N_i(d) (1 - p_i(d)) + \\ &+ N_i(d-1)p_i^1(d-1) + \sum_{j=2}^m N_i(d-j)p_i^j(d-j) + O(p_i). \end{aligned} \quad (1.27)$$

Доказательство ведется индукцией сначала по  $d$  и затем по  $i$ . Сначала мы докажем теорему для  $d = m$  и всех  $i$ . Затем, мы докажем, что если теорема верна для  $d = d_0$  и всех  $i$ , то она верна и для  $d = d_0 + 1$  и всех  $i$ .

Рассмотрим случай  $d = m$ . Для константного числа  $i$  мы, очевидно, имеем  $\mathbf{E}N_i(m) = \frac{i}{Am+B+1} + \theta(C_1)$  для некоторого  $C_1$ . Предположим, что  $\mathbf{E}N_i(m) = \frac{i}{Am+B+1} + \theta(C_1)$ . Из (1.26) мы получаем

$$\begin{aligned} \mathbf{E}N_{i+1}(m) &= \mathbf{E}N_i(m) (1 - p_i(m)) + 1 - p_i = \\ &= \left( \frac{i}{Am+B+1} + \theta(C_1) \right) (1 - p_i(m)) + 1 + \theta(C_2/i) = \\ &= \frac{i+1}{Am+B+1} + \theta(C_1) (1 - p_i(m)) + \theta \left( \frac{C_3}{i} \right) \frac{1}{Am+B+1} + \theta(C_2/i). \end{aligned}$$

Остается показать, что

$$C_1 p_i(m) \geq \frac{C_3}{i(Am+B+1)} + \theta(C_2/i).$$

Мы имеем  $p_i(m) \geq \frac{mA+B}{i} - \frac{C_0}{i^2}$ . Это дает нам

$$C_1(Am+B) \geq \frac{C_1 C_0}{i} + \frac{C_3}{Am+B+1} + C_2.$$

Это равенство выполняется для больших  $i$  и  $C_1$ . Это завершает доказательство для случая  $d = m$ .

Напомним, что доказательство ведется индукцией по  $d$  и  $i$ . Рассмотрим случай  $d > m$  и предположим, что мы можем доказать теорему для всех меньших степеней. Теперь мы сделаем шаг индукции по  $d$ , используя индукцию по  $i$ .

Мы имеем  $N_i(d) \leq \frac{2mi}{d}$ , поэтому  $N_i(d) = O(ic(m, d)d^{1/A})$ . В частности, для  $i < 2C_7 d^2$ , мы имеем

$$\mathbf{E}N_i(d) = c(m, d) \left( i + \theta \left( Cd^{2+1/A} \right) \right)$$

для некоторого  $C$ . Здесь, константа  $C_7$  зависит только от параметров модели и имеет такой большой индекс, так как ее значение будет определено только в последующих выкладках. Таким образом, база индукции по  $i$  верна и мы можем сделать шаг индукции, предполагая, что  $i \geq 2C_7 d^2$ .

Из (1.27) и предположения индукции мы получаем

$$\begin{aligned}
\mathbf{E}N_{i+1}(d) &= \mathbf{E}N_i(d) (1 - p_i(d)) + \mathbf{E}N_i(d-1)p_i^1(d-1) + \\
&\quad + \sum_{j=2}^m \mathbf{E}N_i(d-j)p_i^j(d-j) + O(p_i) = \\
&= c(m, d) \left( i + \theta \left( Cd^{2+1/A} \right) \right) (1 - p_i(d)) + \\
&\quad + c(m, d-1) \left( i + \theta \left( C(d-1)^{2+1/A} \right) \right) p_i^1(d-1) + \theta \left( \frac{C_4 c(m, d) d^2 i d^{1/A}}{i^2} \right) = \\
&= c(m, d)(i+1) + c(m, d-1)ip_i^1(d-1) - \\
&\quad - c(m, d)ip_i(d) - c(m, d) + c(m, d)\theta \left( Cd^{2+1/A} \right) (1 - p_i(d)) + \\
&\quad + \frac{c(m, d)(Ad + B + 1)}{Ad - A + B} \theta \left( C(d-1)^{2+1/A} \right) p_i^1(d-1) + \theta \left( \frac{C_4 c(m, d) d^2 d^{1/A}}{i} \right) = \\
&= c(m, d)(i+1) + c(m, d)\theta \left( Cd^{2+1/A} \right) (1 - p_i(d)) + \\
&\quad + \frac{c(m, d)(Ad + B + 1)}{Ad - A + B} \theta \left( C(d-1)^{2+1/A} \right) p_i^1(d-1) + \theta \left( \frac{C_5 c(m, d) d^2 d^{1/A}}{i} \right).
\end{aligned}$$

Нам нужно доказать, что существует такая константа  $C$ , что

$$Cd^{2+1/A}p_i(d) \geq \frac{C(Ad + B + 1)}{Ad - A + B} (d-1)^{2+1/A} p_i^1(d-1) + \frac{C_5 d^{2+1/A}}{i},$$

$$\begin{aligned}
Cd^{2+1/A}p_i(d) &\geq \frac{C(Ad + B + 1)}{Ad - A + B} \left( d^{2+1/A} - (2 + 1/A)d^{1+1/A} + C_6 d^{1/A} \right) \cdot \\
&\quad \cdot p_i^1(d-1) + \frac{C_5 d^{2+1/A}}{i},
\end{aligned}$$

$$\begin{aligned}
\frac{Cd^{2+1/A}}{i} \left( 2A + \frac{(B-A)(2A+1)}{Ad} + O\left(\frac{d}{i^2}\right) \right) &\geq Cd^{2+1/A} O\left(\frac{d^2}{i^2}\right) + \\
&\quad + \frac{C(Ad + B + 1)}{Ad - A + B} C_6 d^{1/A} \left( A \frac{d-1}{i} + B \frac{1}{i} + O\left(\frac{d^2}{i^2}\right) \right) + \frac{C_5 d^{2+1/A}}{i}, \\
\frac{Cd^{2+1/A}}{i} &\geq \frac{C_7 Cd^{4+1/A}}{i^2} + \frac{C_8 Cd^{1+1/A}}{i} + \frac{C_9 d^{2+1/A}}{i}.
\end{aligned}$$

Константа  $C_7$  возникает из  $O\left(\frac{d^2}{i^2}\right)$ . При  $i > 2C_7 d^2$  это неравенство выполняется для больших  $C \geq C_{10}$  и  $d \geq d_1$ . Для константного числа  $d < d_1$  существует

такая функция  $f(d) > 0$ , что

$$f(d)d^{2+1/A}p_i(d) \geq f(d)\frac{Ad + B + 1}{Ad - A + B}(d - 1)^{2+1/A}p_i^1(d - 1) + \frac{C_5d^{2+1/A}}{i}.$$

Таким образом, окончательно получаем, что  $C = \max\{C_{10}, \max_{d < d_1}\{f(d)\}\}$ . Это завершает доказательство.

### Доказательство теоремы 11

Для доказательства теоремы 11 нам понадобится неравенство Азумы–Хеффдинга:

**Теорема** (Азума, Хеффдинг). Пусть мартингал  $(X_i)_{i=0}^n$  таков, что  $|X_i - X_{i-1}| \leq c_i$  для любого  $1 \leq i \leq n$ . Тогда

$$\mathbf{P}(|X_n - X_0| \geq x) \leq 2e^{-\frac{x^2}{2\sum_{i=1}^n c_i^2}}$$

для всех  $x > 0$ .

Зафиксируем некоторые  $n$  и  $d$  и рассмотрим случайные величины  $X_i(d) = \mathbf{E}(N_n(d) \mid G_m^i)$ ,  $i = 0, \dots, n$ . Случайная величина  $\mathbf{E}(N_n(d) \mid G_m^i)$  является средним числом вершин степени  $d$ , которое мы могли бы иметь на шаге  $n$  процесса  $G_m^i$ , если бы мы зафиксировали первые  $i$  шагов эволюции и разрешили остальным  $n - i$  шагам быть произвольными. Заметим, что  $X_0(d) = \mathbf{E}N_n(d)$  и  $X_n(d) = N_n(d)$ . Легко увидеть, что  $X_i(d)$  является мартингалом.

Ниже мы докажем, что для всех  $i = 0, \dots, n - 1$  выполнено

$$|X_{i+1}(d) - X_i(d)| \leq Md,$$

где  $M > 0$  — это некоторая константа. Теорема немедленно следует из этого утверждения. В самом деле, положим  $c_i = Md$  для всех  $i$ . Тогда из неравенства Азумы–Хеффдинга следует, что

$$\mathbf{P}(|N_n(d) - \mathbf{E}N_n(d)| \geq d\sqrt{n} \ln n) \leq 2 \exp\left\{-\frac{n d^2 \ln^2 n}{2n M^2 d^2}\right\} = O(n^{-\ln n}).$$

Предположим, мы задались некоторым  $\delta > 0$ . Если  $d \leq n^{\frac{A-\delta}{4A+2}}$ , то значение  $\frac{n}{d^{1+1/A}}$  значительно больше величины  $d \ln n \sqrt{n}$ . Это в точности то, что нам надо.

Нам осталось оценить величину  $|X_{i+1}(d) - X_i(d)|$ . Доказательство проводится непосредственным подсчетом.

Зададимся  $0 \leq i \leq n - 1$  и некоторым графом  $G_m^i$ . Заметим, что

$$\begin{aligned} & | \mathbf{E} (N_n(d) \mid G_m^{i+1}) - \mathbf{E} (N_n(d) \mid G_m^i) | \leq \\ & \leq \max_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{ \mathbf{E} \left( N_n(d) \mid \tilde{G}_m^{i+1} \right) \right\} - \min_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{ \mathbf{E} \left( N_n(d) \mid \tilde{G}_m^{i+1} \right) \right\}. \end{aligned}$$

Положим  $\hat{G}_m^{i+1} = \arg \max \mathbf{E}(N_n(d) \mid \tilde{G}_m^{i+1})$ ,  $\bar{G}_m^{i+1} = \arg \min \mathbf{E}(N_n(d) \mid \tilde{G}_m^{i+1})$ . Мы хотим оценить разницу  $\mathbf{E}(N_n(d) \mid \hat{G}_m^{i+1}) - \mathbf{E}(N_n(d) \mid \bar{G}_m^{i+1})$ .

Для  $i + 1 \leq t \leq n$  положим

$$\delta_t^i(d) = \mathbf{E}(N_t(d) \mid \hat{G}_m^{i+1}) - \mathbf{E}(N_t(d) \mid \bar{G}_m^{i+1}).$$

Мы хотим доказать, что  $\delta_n^i(d) \leq Md$ , для этого мы по индукции докажем, что  $\delta_t^i(d) \leq Md$  для  $1 \leq i \leq n$  и  $i \leq t \leq n$ .

Во-первых, заметим, что если  $t \leq C_{11}d^2$ , то мы имеем  $\delta_t^i(d) \leq \frac{2mt}{d} \leq Md$  для некоторой константы  $M$ .

Зафиксируем  $G_m^i$ . Графы  $\hat{G}_m^{i+1}$  и  $\bar{G}_m^{i+1}$  получены из графа  $G_m^i$  путем добавления вершины  $i + 1$  и  $m$  ребер. Поэтому  $\delta_{i+1}^i(d) \leq 2m$ .

Теперь рассмотрим  $t: i \leq t \leq n, t > C_{11}d^2$ . Заметим, что

$$\begin{aligned} \mathbf{E} (N_{t+1}(m) \mid G_m^i) &= \mathbf{E} (N_t(m) \mid G_m^i) (1 - p_t(m)) + 1 + O(1/t), \\ \mathbf{E} (N_{t+1}(d) \mid G_m^i) &= \mathbf{E} (N_t(d) \mid G_m^i) (1 - p_t(d)) + \\ &+ \mathbf{E} (N_t(d-1) \mid G_m^i) p_t^1(d-1) + \sum_{j=2}^m \mathbf{E} (N_t(d-j) \mid G_m^i) p_t^j(d-j) + \\ &+ O(1/t), \quad d \geq m + 1. \end{aligned}$$

Мы получали подобные равенства при доказательстве теоремы 10, см. (1.26–1.27). Заменим в этих равенствах граф  $G_m^i$  графом  $\hat{G}_m^i$  или графом  $\bar{G}_m^i$ . Вычитая равенства с  $\bar{G}_m^i$  из равенств с  $\hat{G}_m^i$ , мы получаем (для  $d > m$ )

$$\begin{aligned} \delta_{t+1}^i(d) &= \delta_t^i(d) (1 - p_t(d)) + \delta_t^i(d-1) p_t^1(d-1) + O \left( \frac{\mathbf{E} N_t(d) d^2}{t^2} \right) + O \left( \frac{1}{t} \right) = \\ &= \delta_t^i(d) (1 - p_t(d)) + \delta_t^i(d-1) p_t^1(d-1) + \theta \left( \frac{C_{12}d}{t} \right). \quad (1.28) \end{aligned}$$

Здесь мы используем тот факт, что  $\mathbf{E} N_t(d) = O(td^{-1-1/A} + d) = O(t/d)$ . По предположению индукции и для достаточно большого  $M$  получаем, что

$$\begin{aligned} \delta_{t+1}^i(d) &\leq Md(1 - p_t(d)) + M(d-1)p_t^1(d-1) + \frac{C_{12}d}{t} \leq \\ &\leq Md - \frac{MA(2d-1)}{t} - \frac{MB}{d} + \frac{C_{13}Md^3}{t^2} + \frac{C_{12}d}{t} \leq Md. \end{aligned}$$

Таким образом, шаг индукции верен. Это завершает доказательство теоремы 11.

### Доказательство теоремы 12

Заметим, что

$$P_2(n) = \sum_{d=m}^{\infty} N_n(d) \frac{d(d-1)}{2}.$$

Подобно приближению среднего поля, мы предполагаем, что случайная величина  $P_2(n)$  сконцентрирована около своего математического ожидания, для которого справедливо следующие рекуррентное соотношение

$$\mathbf{E}P_2(i+1) = \mathbf{E}P_2(i) \left(1 + \frac{2A}{i}\right) + 2m(A+B) + \frac{m(m-1)}{2}.$$

Оно мотивировано следующими выкладками

$$\begin{aligned} \mathbf{E}P_2(i+1) &= \sum_{d=m}^{\infty} \mathbf{E}N_{i+1}(d) \frac{d(d-1)}{2} = \mathbf{E}P_2(i) + \frac{m(m-1)}{2} + \sum_{d=m}^{\infty} \mathbf{E}N_i(d) p_i(d) d \sim \\ &\sim \mathbf{E}P_2(i) + \frac{m(m-1)}{2} + \sum_{d=m}^{\infty} \frac{(Ad+B)d \mathbf{E}N_i(d)}{i} = \mathbf{E}P_2(i) \left(1 + \frac{2A}{i}\right) + \frac{m(m-1)}{2} + \\ &+ \sum_{d=m}^{\infty} \frac{(A+B)d \mathbf{E}N_i(d)}{i} = \mathbf{E}P_2(i) \left(1 + \frac{2A}{i}\right) + 2m(A+B) + \frac{m(m-1)}{2}. \end{aligned}$$

Таким образом, мы получаем

$$\begin{aligned} \mathbf{E}P_2(n) &\sim \left(2m(A+B) + \frac{m(m-1)}{2}\right) \sum_{t=1}^n \prod_{i=t+1}^n \left(1 + \frac{2A}{i}\right) \sim \\ &\sim \left(2m(A+B) + \frac{m(m-1)}{2}\right) \sum_{t=1}^n \frac{n^{2A}}{t^{2A}}. \end{aligned}$$

Если  $2A < 1$ , то

$$\mathbf{E}P_2(n) \sim \left(2m(A+B) + \frac{m(m-1)}{2}\right) \frac{n}{1-2A}.$$

Если  $2A = 1$ , то

$$\mathbf{E}P_2(n) \sim \left(2m(A+B) + \frac{m(m-1)}{2}\right) n \ln(n).$$

Если  $2A > 1$ , то

$$EP_2(n) = O(n^{2A}) .$$

## Глава 2

# Свойства медиа-веба и модели с устареванием

Вторая глава этой работы основана на статье автора [54].

В разделах 2.1 и 2.2 мы обсудим базовые модели и экспериментальные результаты, которые мотивировали нашу работу. В разделе 2.3, основываясь на результатах экспериментов, мы определим наш класс моделей с устареванием. Мы теоретически проанализируем некоторые свойства этих моделей в разделе 2.4, тогда как в разделе 2.5 мы проверим каждую из наших моделей эмпирически посредством оценивания правдоподобия реальных данных при условии, что данные появились в соответствии с этой моделью.

### 2.1 Базовые модели

Как уже обсуждалось во введении и разделе 1.1, одна из первых попыток предложить реалистичную математическую модель Веба была предпринята Барабаши и Альберт в [5]. Главная идея была в том, чтобы учесть тот факт, что новые страницы часто ссылаются на популярные старые, а мерой популярности выступала текущая степень вершины.

Позднее Бианкони и Барабаши заметили, что в реальных сетях узлы получают ребра не только из-за своей большой степени (популярности), но также и благодаря своим внутренним свойствам [7]. Например, новая веб-страница с действительно интересным и качественным материалом может быстро получить много новых ссылок и стать популярнее старых страниц. Мотивированные этим наблюдением Бианкони и Барабаши расширили модели предпочтительного присоединения, введя присущее каждому узлу качество или *приспособленность (fitness)* узла. Когда новый узел добавляется в сеть, из него проводится ребро в некоторый уже существующий узел с вероятностью, пропорциональной произведению качества (приспособленности) и текущей



степени выбранного узла. Строго математически эта модель анализировалась в [11].

В контексте нашего исследования, главный недостаток этих моделей состоит в том, что в них уделяется слишком много внимания старым страницам и нереалистично объясняется, как появляются ссылки на недавно созданные страницы. Также заметим, что такие высокодинамичные части веба, как социальный веб и медиа-веб, имеют специфическое поведение и поэтому нуждаются в отдельных моделях (см. раздел 2.2). Так, в [35] была подробно исследована эволюция социальных сетей, или социального веба. Основываясь на своих результатах, авторы предложили модель, качество которой было проверено путем детального эмпирического анализа, включавшего использование метода максимального правдоподобия для настройки параметров модели. Мы, в свою очередь, предлагаем модель медиа-веба, т.е. высокодинамической части веба, где ежедневно появляется множество новых страниц, связанных с медиа-контентом: новостями, постами в блогах и форумах. При этом для поиска более реалистичной модели мы используем как теоретический анализ, так и метод максимального правдоподобия, который позволяет настраивать и количественно сравнивать модели.

Основная идея состоит в том, чтобы соединить предпочтительное присоединение и приспособленности с фактором устаревания страниц. Это значит, что страницы получают ссылки согласно своей *привлекательности*, которая определяется как произведение входящей степени страницы, ее *внутреннего качества* (некоторой константы, связанной с каждой страницей) и возраста (новые страницы активнее получают новые ссылки).

## 2.2 Свойство устаревания медиа-веба

### 2.2.1 Данные

Мы использовали публичные данные MemeTracker [1], которые покрывают довольно значительный период активности медиа-веба размером в 9 месяцев. Заметим, что исходящие ссылки извлекались из основной части страницы (меню и рекламная обвязка игнорировались). Детали о том, как собирались данные, могут быть найдены в [34].

Из этих данных мы отобрали только те ссылки, которые не выводят за пределы коллекции, т.е. только те, для которых известно время появления и источник, и цели ссылки. Затем мы отфильтровали ссылки, у которых время создания цели было меньше времени создания источника. Такие невозмож-

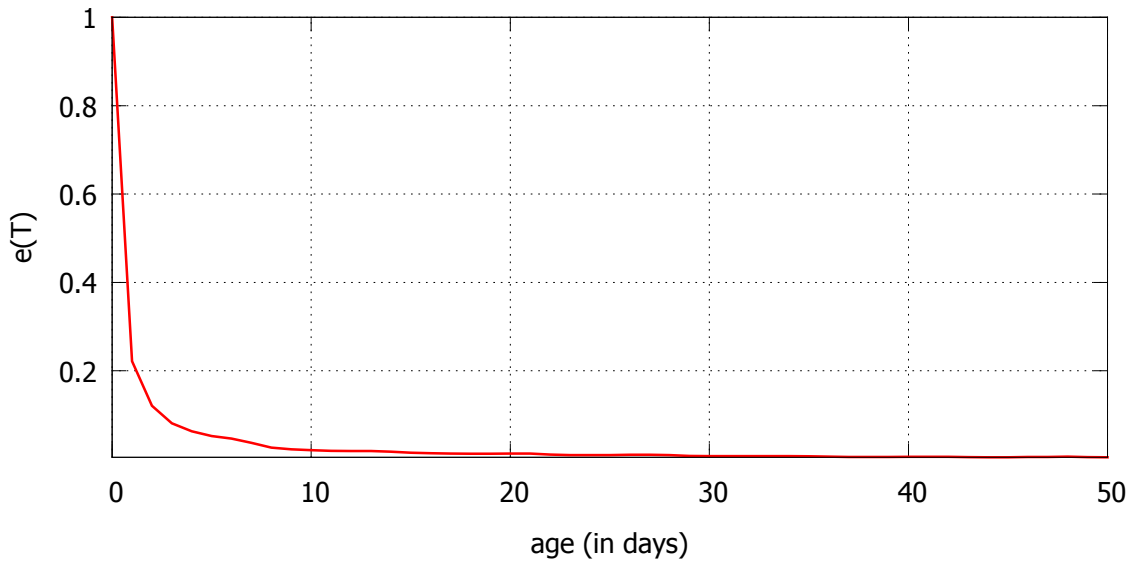


Рис. 2.1: Свойство устаревания

ные в реальности ссылки могут встречаться, потому что в данных есть шум и не всем временам создания, приведенным в данных, можно полностью доверять. В итоге мы получили около 18М ссылок и 6.5М документов, которые использовали в последующих экспериментах.

### 2.2.2 Свойство устаревания

Определим *свойство устаревания* для графа, развивающегося во времени. Обозначим через  $e(T)$  долю ребер, соединяющих страницы с разницей возрастов больше  $T$ . Мы проанализировали поведение  $e(T)$  и увидели, что медиа-страницы стремятся ссылаться на страницы близкого возраста. А именно, мы построили график зависимости  $e(T)$  для наших данных (см. Рис. 2.1) и заметили, что  $e(T)$  убывает экспоненциально быстро, что, как мы увидим дальше, не соответствует моделям предпочтительного присоединения (параграф 2.4.2).

## 2.3 Модель медиа-веба

В этом разделе мы опишем нашу модель.

Предположим, что мы имеем фиксированный набор хостов  $H_1, \dots, H_n$  и каждый хост, т.е. множество страниц,  $H_i$  характеризуется скоростью появления на нем новых страниц  $\lambda_i$ . Мы стартуем с пустого графа, т.е. в начале имеется  $n$  пустых множеств  $H_1, \dots, H_n$ . Далее мы предполагаем, что новые

страницы на хосте  $H_i$  появляются согласно пуассоновскому процессу с параметром  $\lambda_i$ .<sup>1</sup> Пуассоновские процессы для разных хостов независимы.

Мы предполагаем, что каждая страница  $p$  при появлении получает качество  $q_p$  и исходящую степень  $m_p$ . Случайные величины  $q_p$  и  $m_p$  независимы в совокупности и одинаково распределены для  $p \in H_i$ , т.е. распределение этих величин зависит только от хоста, которому принадлежит страница.

Когда новая страница  $p$  появляется на хосте  $H_i$ , она получает качество  $q_p$  и проводит взаимно независимо  $m_p$  исходящих ссылок в уже существующие страницы. Цель каждой ссылки выбирается следующим образом. Сначала выбирается целевой хост  $k$  с вероятностью  $\rho_{ik}$  ( $\sum_{k=1}^n \rho_{ik} = 1$ ). Затем вероятность выбрать страницу  $r$  на хосте  $H_k$  полагается пропорциональной *привлекательности*  $f$  страницы  $r$ , которая является некоторой функцией  $d_r$  (текущей степени страницы  $r$ ),  $q_r$  (качества страницы  $r$ ) и  $a_r$  (текущего возраста страницы  $r$ ). Рассматривается следующее семейство функций привлекательности:

$$f_{\vec{\alpha}, \tau_k}(d, q, a) = q^{\alpha_1} \cdot d^{\alpha_2} \cdot e^{-\frac{a\alpha_3}{\tau_k}},$$

где  $\tau_k$  контролирует скорость убывания привлекательности медиа-страниц на хосте  $H_k$ , а  $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3) \in \{0, 1\}^3$ .

Например,  $f(d, q, a) = d$  (т.е.  $\vec{\alpha} = (0, 1, 0)$ ) приводит к предпочтительному присоединению, тогда как  $f(d, q, a) = q \cdot d$  (т.е.  $\vec{\alpha} = (1, 1, 0)$ ) приводит к модели приспособления (fitness model). Мы изучим разные варианты и покажем, какие из них лучшим образом отражают поведение медиа-веба.

## 2.4 Теоретический анализ предложенной модели

### 2.4.1 Распределение входящей степени

В [7, 10, 15] уже анализировались модели без *фактора устаревания* (т.е., без множителя  $e^{-\frac{a}{\tau_k}}$  в функции привлекательности). В параграфе 2.4.2 мы покажем, что для того, чтобы отражать свойство устаревания медиа-веба, нам необходим фактор устаревания. Поэтому мы предполагаем здесь, что функция привлекательности имеет фактор устаревания.

Выберем некоторую страницу  $p \in H_k$ . Обозначим через  $d_p(q_p, t, t_p)$  входящую степень страницы  $p$  в момент времени  $t$ , если она была создана в момент времени  $t_p$  и имела качество  $q_p$ . Определим также для каждого хоста

---

<sup>1</sup>Пуассоновский процесс часто используется для того, чтобы моделировать последовательность независимых событий постоянной во времени интенсивности.

$H_k$  среднюю привлекательность его страниц в момент времени  $t$ :

$$W_k(t) = \mathbb{E} \sum_{p \in H_k} f(d_p(q_p, t, t_p), q_p, t - t_p). \quad (2.1)$$

Мы покажем в этом параграфе, что  $W_k(t) \rightarrow W_k$ , когда  $t \rightarrow \infty$ , где  $W_k$  — это некоторая положительная константа.

Пусть  $M_k$  — это средняя исходящая степень страниц  $p \in H_k$ . Тогда  $N_k = \sum_i \lambda_i M_i \rho_{ik}$  — средняя скорость появления ссылок, указывающих на хост  $H_k$ .

**Теорема 15.** Пусть  $p \in H_k$  — это страница с качеством  $q_p$  и временем создания  $t_p$ , тогда в приближении среднего поля мы имеем

$$(1) \text{ Если } f = q \cdot d \cdot e^{-\frac{a}{\tau_k}}, \text{ то } d_p(q_p, t, t_p) = e^{\frac{N_k \tau_k q_p}{W_k} \left(1 - e^{-\frac{t_p - t}{\tau_k}}\right)}.$$

$$(2) \text{ Если } f = q \cdot e^{-\frac{a}{\tau_k}}, \text{ то } d_p(q_p, t, t_p) = \frac{N_k \tau_k q_p}{W_k} \left(1 - e^{-\frac{t_p - t}{\tau_k}}\right).$$

Из теоремы 15 следует, что в первом случае, чтобы иметь степенной закон распределения случайной величины  $d_p$ , качество  $q_p$  должно быть распределено экспоненциально. В первом случае для каждого хоста параметр степенного закона равняется  $\frac{N_k \tau_k \mu}{W_k}$ , где  $\mu$  — это параметр экспоненциального распределения. Интересно отметить, что этот параметр  $\mu$  не влияет на параметр распределения степеней вершин. Действительно, если мы умножим  $\mu$  на некоторую константу, то  $W_k$  также умножится на эту же константу и она сократится (см. (2.1)). Поэтому мы можем менять параметр распределения степеней вершин, только меняя  $N_k$  и  $\tau_k$ . Проблема состоит в том, что константа  $W_k$  не явно зависит от  $N_k$  и  $\tau_k$  (см. выражение (2.3) в доказательстве). Таким образом, невозможно найти аналитические выражения для  $N_k$  и  $\tau_k$ , которые обеспечили бы желаемое распределение степеней вершин.

Во втором случае степенной закон  $q_p$  приводит к степенному закону  $d_p$  с тем же параметром. Поэтому в этом случае можно получить реалистичное распределение входящей степени.

В обоих случаях нельзя исключить качество страницы, так как, если мы не имеем его в функции привлекательности, то решение не зависит от  $q_p$  и нельзя получить степенной закон для распределения степеней вершин.

Чтобы проиллюстрировать результаты теоремы 15, мы сгенерировали графы согласно нашей модели для разных функций привлекательности  $f$ . Полученные результаты показаны на Рис. 2.2.

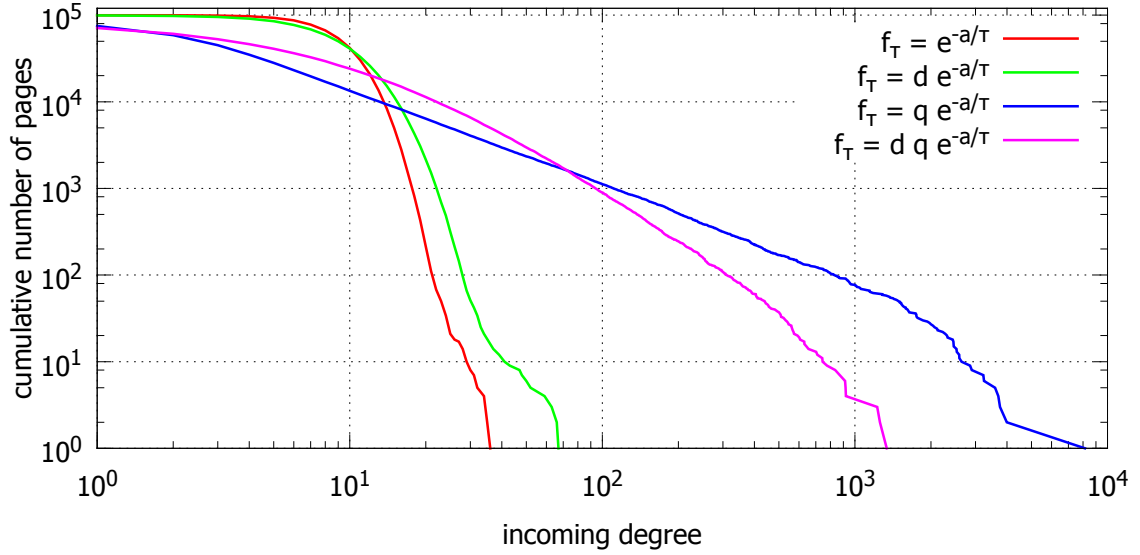


Рис. 2.2: Распределение входящей степени для каждой из моделей

*Доказательство.* В приближении среднего поля мы имеем следующее дифференциальное уравнение:

$$\frac{\partial d_p(q_p, t, t_p)}{\partial t} = N_k \frac{f(d_p(q_p, t, t_p), q_p, t - t_p)}{W_k(t)},$$

здесь  $p \in H_k$ .

Если  $f(d, q, a) = q \cdot d \cdot e^{-\frac{a}{\tau_k}}$ , то

$$\frac{\partial d_p(q_p, t, t_p)}{\partial t} = N_k \frac{q_p \cdot d_p(q_p, t, t_p) \cdot e^{-\frac{t-t_p}{\tau_k}}}{W_k(t)}. \quad (2.2)$$

Ниже мы покажем, что для каждого  $k$  величина  $W_k(t)$  стремится к некоторой положительной константе  $W_k$ :  $\lim_{t \rightarrow \infty} W_k(t) = W_k$ .

Мы поэтому имеем следующее решение уравнения (2.2):

$$d_p = e^{\frac{N_k \tau_k q_p}{W_k} \left(1 - e^{-\frac{t_p-t}{\tau_k}}\right)} \xrightarrow{t \rightarrow \infty} e^{\frac{N_k \tau_k q_p}{W_k}}.$$

Если  $f(d, q, a) = q \cdot e^{-\frac{a}{\tau_k}}$ , то посредством похожих, но даже более легких вычислений, мы получаем

$$d_p = \frac{N_k \tau_k q_p}{W_k} \left(1 - e^{-\frac{t_p-t}{\tau_k}}\right) \xrightarrow{t \rightarrow \infty} \frac{N_k \tau_k q_p}{W_k}.$$

Проверим теперь, что  $\lim_{t \rightarrow \infty} W_k(t)$  действительно константа. Рассмотрим случай  $f(d, q, a) = q \cdot d \cdot e^{-\frac{a}{\tau_k}}$ . Пусть  $\rho_k(q)$  — плотность распределения  $q_p$  для

$p \in H_k$ . Тогда

$$\begin{aligned}
W_k(t) &= \int_0^\infty \left( \int_0^t \lambda_k q \rho_k(q) d(q, t, x) \cdot e^{-\frac{t-x}{\tau_k}} dx \right) dq = \\
&= \int_0^\infty \left( \int_0^t \lambda_k q \rho_k(q) e^{\frac{N_k \tau_k q}{W_k} \left(1 - e^{-\frac{x-t}{\tau_k}}\right)} \cdot e^{-\frac{x-t}{\tau_k}} dx \right) dq = \\
&= \int_0^\infty \frac{\lambda_k W_k}{N_k} \left( e^{\frac{N_k \tau_k q}{S} \left(1 - e^{-\frac{t}{\tau_k}}\right)} - 1 \right) \rho_k(q) dq.
\end{aligned}$$

Таким образом, в итоге для  $W_k$  мы имеем следующее уравнение:

$$W_k = \lim_{t \rightarrow \infty} W_k(t) = \underbrace{\frac{\lambda_k W_k}{N_k} \left( \int_0^\infty e^{\frac{N_k \tau_k q}{W_k}} \rho_k(q) dq - 1 \right)}_{F_k(W_k)}. \quad (2.3)$$

Существует единственное решение уравнения (2.3). Чтобы показать это, мы сначала проверяем, что функция  $y = F_k(x)$  является монотонной:

$$F'_k(x) = \frac{\lambda_k}{N_k} \left( \int_0^\infty e^{\frac{N_k \tau_k q}{x}} \left( 1 - \frac{N_k \tau_k q}{x} \right) \rho_k(q) dq - 1 \right) \leq 0,$$

так как

$$e^{\frac{N_k \tau_k q}{x}} \left( 1 - \frac{N_k \tau_k q}{x} \right) \leq 1 \text{ и } \int_0^\infty \rho_k(q) dq = 1.$$

Также  $F_k(x) \rightarrow \tau_k \lambda_k \mathbf{E}_{p \in S_k} q_p$ , когда  $x \rightarrow \infty$ , и  $F_k(x) \rightarrow \infty$ , когда  $x \rightarrow 0$ . Из этого следует, что  $y = x$  и  $y = F_k(x)$  имеют единственное пересечение. Другими словами, уравнение (2.3) имеет единственное решение.

Аналогично мы можем показать, что  $\lim_{t \rightarrow \infty} W_k(t) = W_k$  для функции привлекательности  $f = q \cdot e^{-\frac{a}{\tau_k}}$ .

□

## 2.4.2 Свойство устаревания

В этом параграфе мы покажем, что нам нужен фактор устаревания  $e^{-\frac{a}{\tau_k}}$  в формуле для функции привлекательности  $f$ . Мы докажем, что благодаря фактору устаревания, количество ребер, которые соединяют ребра с разницей возрастов больше  $T$ , убывает экспоненциально по  $T$ . А именно, мы докажем следующую теорему.

**Теорема 16.** Для  $f = q \cdot d \cdot e^{-\frac{a}{\tau_k}}$  или  $f = q \cdot e^{-\frac{a}{\tau_k}}$  в приближении среднего поля мы имеем

$$e(T) = (1 + o(1)) \sum_k N_k C_k e^{-\frac{T}{\tau_k}},$$

где  $C_k$  — это некоторые константы.

*Доказательство.* Чтобы оценить  $e(T)$ , нам сначала понадобится оценить для всех страниц, созданных в период времени  $T$  на хосте  $H_k$ , среднюю привлекательность в момент времени  $t$ :

$$\mathcal{W}_k(T, t) = \mathbb{E} \sum_{\substack{p \in H_k \\ |t-t_p| < T}} f(d_p(q_p, t, t_p), q_p, t - t_p).$$

Мы покажем, что если  $t > T$ , то эта функция не зависит от  $t$ .

Мы можем проанализировать функцию  $\mathcal{W}_k(T, t)$ , используя технику из раздела 2.4.1. Рассмотрим случай  $f(d, q, a) = q \cdot d \cdot e^{-\frac{a}{\tau_k}}$ :

$$\begin{aligned} \mathcal{W}_k(T, t) &= \\ &= \int_0^\infty \left( \int_{t-T}^t \lambda_k q \rho_k(q) d(q, t, x) \cdot e^{-\frac{t-x}{\tau_k}} dx \right) dq = \\ &= \int_0^\infty \left( \int_{t-T}^t \lambda_k q \rho_k(q) e^{\frac{N_k \tau_k q}{W_k} \left(1 - e^{-\frac{x-t}{\tau_k}}\right)} \cdot e^{-\frac{x-t}{\tau_k}} dx \right) dq = \\ &= \frac{\lambda_k W_k}{N_k} \int_0^\infty \left( e^{\frac{N_k \tau_k q}{W_k} \left(1 - e^{-\frac{T}{\tau_k}}\right)} - 1 \right) \rho_k(q) dq. \end{aligned}$$

Мы доказали, что  $\mathcal{W}_k(T, t)$  не зависит от  $t$ , и теперь мы будем использовать обозначение  $\mathcal{W}_k(T) = \mathcal{W}_k(T, t)$ . Также

$$\begin{aligned} W_k - \mathcal{W}_k(T) &= \\ &= \frac{\lambda_k W_k}{N_k} \int_0^\infty \left( 1 - e^{-\frac{N_k \tau_k q}{W_k} e^{-\frac{T}{\tau_k}}} \right) e^{\frac{N_k \tau_k q}{W_k}} \rho_k(q) dq = \\ &= (1 + o(1)) \frac{\lambda_k W_k}{N_k} e^{-\frac{T}{\tau_k}} \int_0^\infty \frac{N_k \tau_k q}{W_k} e^{\frac{N_k \tau_k q}{W_k}} \rho_k(q) dq = (1 + o(1)) C_k e^{-\frac{T}{\tau_k}}, \end{aligned}$$

где константа  $C_k$  не зависит  $T$ .

Заметим, что доля ссылок, которые идут на хост  $H_k$  и имеют разницу возрастов меньше  $T$ , есть  $\frac{W_k - \mathcal{W}_k(T)}{W_k}$ . Таким образом, используя величину  $N_k$ , которая равняется средней скорости появления ссылок на хост  $H_k$  (см. параграф 2.4.1), мы можем выписать следующее выражение для  $e(T)$ :

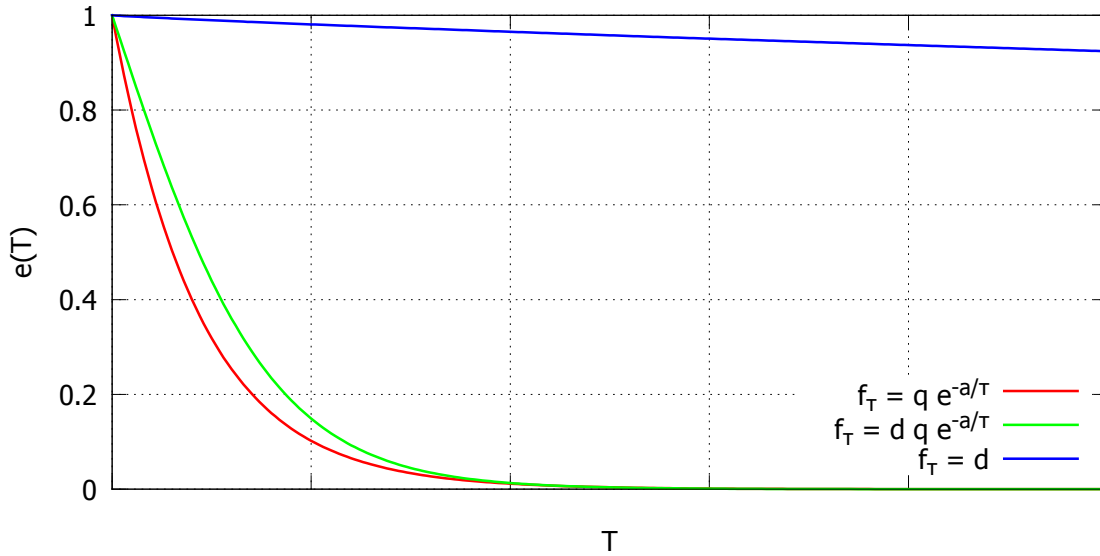


Рис. 2.3: Свойство устаревания в модели

$$e(T) = \sum_k N_k \frac{W_k - \mathcal{W}_k(T)}{W_k} = (1 + o(1)) \sum_k \frac{N_k C_k}{W_k} e^{-\frac{T}{\tau_k}}.$$

Подобный анализ может быть проведен для случая  $f(d, q, a) = q \cdot e^{-\frac{a}{\tau_k}}$ . В этом случае мы получаем

$$\mathcal{W}_k(T) = \int_0^\infty \left( \int_{t-T}^t \lambda_k q \rho_k(q) \cdot e^{-\frac{t-x}{\tau_k}} dx \right) dq = \lambda_k \tau_k \left( 1 - e^{-\frac{T}{\tau_k}} \right) \mathbf{E}_{p \in H_k} q_p,$$

где  $\mathbf{E}_{p \in H_k} q_p$  — среднее качество страниц на хосте  $H_k$ . Дальнейшие рассуждения похожи на приведенные выше. □

Чтобы проиллюстрировать полученные результаты, мы построили график функции  $e(T)$  для разных функций привлекательности (см. Рис. 2.3). Заметим, что для функции привлекательности, равной  $d$  (предпочтительное присоединение), величина  $e(T)$  убывает слишком медленно.

## 2.5 Эмпирический анализ предложенной модели

Идея использовать метод максимального правдоподобия для того, чтобы сравнивать разные графовые модели, была предложена в [6]. После этого метод был использован для нескольких моделей [35, 36]. Мотивированные этими работами мы также используем метод максимального правдоподобия для



сравнения модели, которую мы предложили, с моделями предпочтительного присоединения и приспособления.

### 2.5.1 Оценивание параметров

Для того, чтобы оценить правдоподобие данных, мы сначала должны оценить параметры наших моделей. Заметим, что здесь мы не пытаемся найти наилучшие оценки для параметров. Вместо этого мы предлагаем использовать некоторые простые и естественные оценки, которых, однако, уже достаточно, чтобы показать улучшения, достигаемые с помощью предложенных моделей.

**Межхостовые вероятности.** Мы оценили матрицу  $\rho_{ij}$ , посчитав долю ссылок, идущих с хоста  $H_i$  на хост  $H_j$ . Заметим, что 74% ссылок являются внутривхостовыми.

**Оценка параметра  $\tau$ .** Для того, чтобы оценить параметр  $\tau_k$  для каждого хоста  $H_k$ , мы рассматриваем гистограмму разниц возрастов цели и источника для всех ссылок. Пусть  $x_i$  ( $i \geq 0$ ) — это количество ссылок “длиной” больше  $i$  дней, но меньше  $i + 1$  дней. Если предположить экспоненциальное убывание, то для  $i < j$  мы имеем  $\frac{x_i}{x_j} = e^{\frac{(i-j)T}{\tau_k}}$ , т.е.,  $\tau_k = \frac{(i-j)T}{\ln \frac{x_i}{x_j}}$ , где  $T$  — это интервал времени размером в один день. Поэтому мы возьмем

$$\tau_k = \sum_{\substack{0 \leq i < j < 10: \\ x_i \neq 0, x_j \neq 0}} \frac{(i-j)T}{\binom{10}{2} \ln \frac{x_i}{x_j}}.$$

Мы оставили только первые 10 дней, так как в действительности хвост распределения несколько тяжелее экспоненциального, а нам важно иметь хорошие оценки на начальном отрезке, т.е. когда появляется большинство новых входящих ссылок.

**Оценка качества.** Зная  $d$  — конечную степень узла, мы можем использовать теорему 15, чтобы найти его качество, т.е. мы имеем  $q = \frac{Wd}{N_k \tau_k}$  в случае  $f = qe^{\frac{-a}{T}}$  и  $q = \frac{W \ln d}{N_k \tau_k}$  в случае  $f = dqe^{\frac{-a}{T}}$ . Заметим, что множитель  $\frac{W}{N_k \tau_k}$  является общим для всех страниц, созданных на хосте  $H_k$ , и поэтому он может быть сокращен. Таким образом, в итоге мы используем следующие оценки:  $q = d$  и  $q = \ln d$  соответственно.

### 2.5.2 Правдоподобие

Для того, чтобы проверить модели, мы опять используем данные, описанные в параграфе 2.2.1, и оцениваем для каждой модели правдоподобие

Таблица 2.1: Логарифм правдоподобия: средний логарифм правдоподобия ребра.

| $d$   | $q$   | $e^{\frac{-a}{\tau}}$ | $dq$  | $de^{\frac{-a}{\tau}}$ | $qe^{\frac{-a}{\tau}}$ | $dqe^{\frac{-a}{\tau}}$ |
|-------|-------|-----------------------|-------|------------------------|------------------------|-------------------------|
| -6.11 | -5.56 | -5.34                 | -6.08 | -5.50                  | <b>-5.17</b>           | -5.45                   |

Таблица 2.2: Абсолютная победа: доля ребер на которых модель побеждает все остальные.

| $d$  | $q$  | $e^{\frac{-a}{\tau}}$ | $dq$ | $de^{\frac{-a}{\tau}}$ | $qe^{\frac{-a}{\tau}}$ | $dqe^{\frac{-a}{\tau}}$ |
|------|------|-----------------------|------|------------------------|------------------------|-------------------------|
| 0.03 | 0.07 | 0.28                  | 0.07 | 0.07                   | <b>0.30</b>            | 0.16                    |

данных при условии, что они появились в соответствии с этой моделью. Мы делаем это следующим образом.

Мы добавляем ребра по одному в соответствии с их историческим порядком и считаем вероятность каждого ребра при условии рассматриваемой модели. Сумма логарифмов полученных вероятностей даст нам логарифм правдоподобия графа. Мы нормализуем сумму на количество ребер и получаем Таблицу 2.1.

Здесь мы видим, что наиболее вероятная модель имеет функцию привлекательности  $f = qe^{\frac{-a}{\tau}}$ . Как бы то ни было, так как времена создания страниц подвержены шуму и не всегда надежны (см. параграф 2.2.1), эти результаты могут быть не репрезентативными (например, если вероятность какого-нибудь ребра слишком мала, она сильно повлияет на общее правдоподобие). Поэтому в дополнение к подсчету логарифма правдоподобия, который может сильно зависеть от выбросов, мы также проводим анализ вероятностей отдельных ребер, т.е. мы пытаемся понять, какая из моделей лучше, изучая отдельные ребра. Мы считаем, что такой более глубокий анализ позволяет уменьшить влияние выбросов при сравнении моделей. Насколько нам известно, такой анализ при использовании метода максимального правдоподобия для сравнения графовых моделей выполняется впервые.

Согласно разным моделям ребра имеют разные вероятности, та из моделей, которая приписывает ребру наибольшую вероятность, называется *побеждающей* на этом ребре (см. Таблицу 2.2). Также для каждой пары моделей  $M_1$  и  $M_2$  мы вычислили процент ребер, имеющих большую вероятность согласно  $M_1$ , чем согласно  $M_2$  (см. Таблицу 2.3). Из обеих таблиц ясно видно, что фактор устаревания играет очень важную роль.

Затем для каждой из моделей мы упорядочили вероятности ребер в убывающем порядке (см. Рис. 2.4), менее вероятные ребра находятся справа. Однако, так как из-за малого зазора между графиками разница была видна не

Таблица 2.3: Поединки: значение в  $(a, b)$  — это доля ребер, на которых  $a$  побеждает  $b$ .

|                         | $d$  | $q$  | $e^{\frac{-a}{\tau}}$ | $dq$ | $de^{\frac{-a}{\tau}}$ | $qe^{\frac{-a}{\tau}}$ | $dqe^{\frac{-a}{\tau}}$ |
|-------------------------|------|------|-----------------------|------|------------------------|------------------------|-------------------------|
| $d$                     | -    | 0.22 | 0.30                  | 0.43 | 0.18                   | 0.22                   | 0.19                    |
| $q$                     | 0.78 | -    | 0.38                  | 0.76 | 0.41                   | 0.23                   | 0.40                    |
| $e^{\frac{-a}{\tau}}$   | 0.70 | 0.62 | -                     | 0.69 | 0.54                   | 0.40                   | 0.53                    |
| $dq$                    | 0.57 | 0.24 | 0.31                  | -    | 0.24                   | 0.23                   | 0.17                    |
| $de^{\frac{-a}{\tau}}$  | 0.82 | 0.59 | 0.44                  | 0.76 | -                      | 0.39                   | 0.43                    |
| $qe^{\frac{-a}{\tau}}$  | 0.78 | 0.77 | 0.60                  | 0.77 | 0.61                   | -                      | 0.62                    |
| $dqe^{\frac{-a}{\tau}}$ | 0.81 | 0.60 | 0.47                  | 0.83 | 0.57                   | 0.38                   | -                       |

достаточно отчетливо, мы нормализовали вероятности, поделив вероятность каждого  $k$ -ого ребра в соответствующем упорядоченном списке на вероятность  $k$ -ого ребра в упорядоченном списке модели предпочтительного присоединения (см. Рис 2.5).

Видно, что модель с  $f = qe^{\frac{-a}{\tau}}$  опять показывает наилучшие результаты в наших тестах. Это значит, что в медиа-вебе вероятность цитирования страницы определяется скорее качеством, чем ее текущей популярностью (т.е. входящей степенью). Наконец, роль  $\rho_{ij}$  показана на Рис. 2.6.

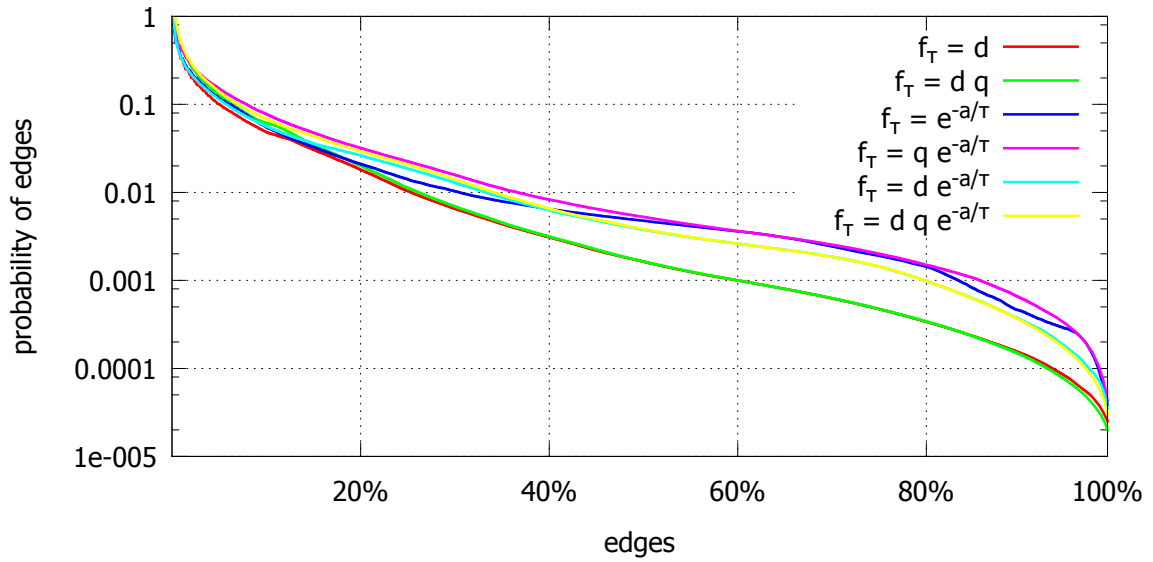


Рис. 2.4: Распределение вероятности ребер

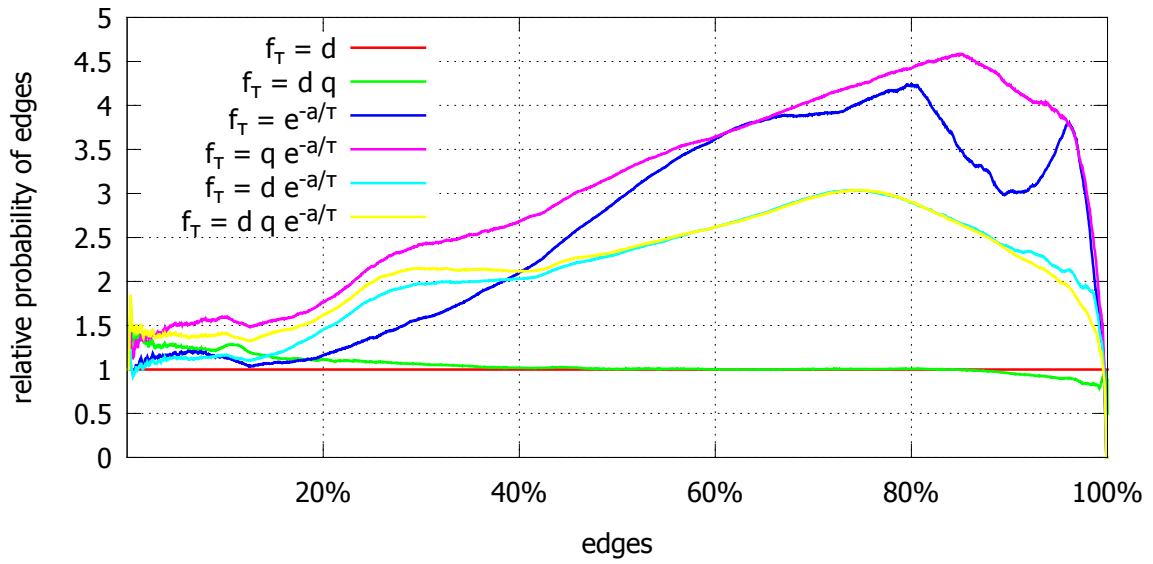


Рис. 2.5: Распределение вероятности ребер относительно модели предпочтительного присоединения

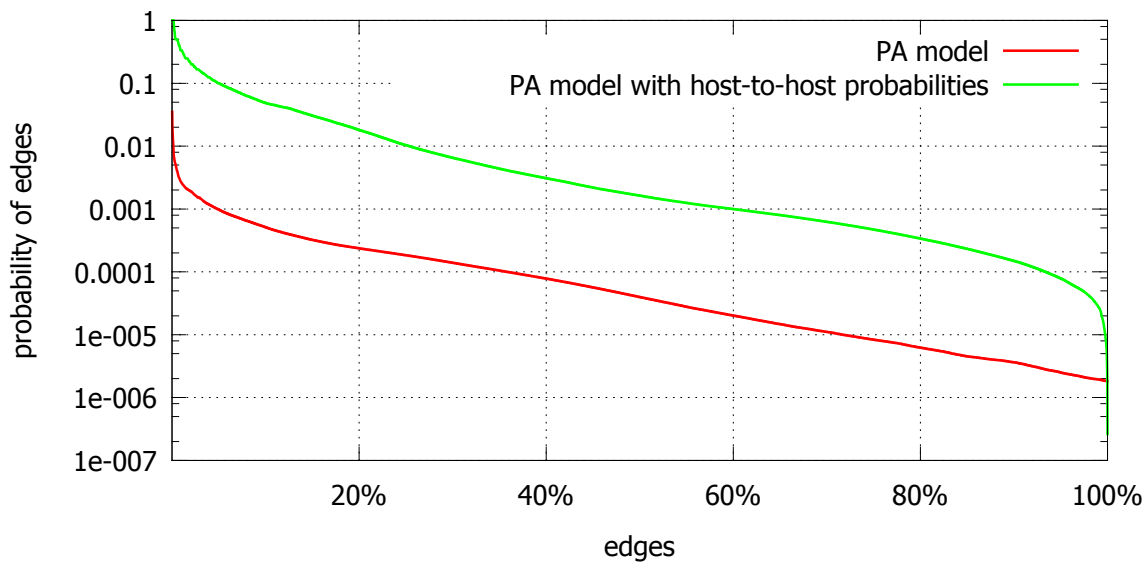


Рис. 2.6: Влияние межхостовых вероятностей на примере модели предпочтительного присоединения (РА)

## Глава 3

# Приложение моделей к задаче обхода эфемерных страниц поисковым роботом

В разделе 3.1 мы формализуем проблему обхода эфемерных страниц поисковым роботом. Мы проверяем гипотезу о том, что большинство эфемерных страниц может быть найдено на небольшом множестве источников контента в разделе 3.2, а в разделе 3.3 предлагаем эффективный алгоритм мониторинга источников контента, который учитывает быстрое убывание пользовательского интереса к эфемерным страницам. Наконец, в разделе 3.4 мы экспериментально анализируем качество работы предложенного алгоритма и обсуждаем полученные результаты в разделе 3.5.

Эта глава основана на статье автора [55], а один из вариантов предложенного алгоритма был успешно внедрен в компании Яндекс.

### 3.1 Формализация проблемы

Во введении мы обсуждали основные задачи поискового робота, а также специфику задачи обхода эфемерных страниц. В этом разделе мы формализуем рассматриваемую проблему, введя подходящую метрику. Мы сначала введем метрику, а затем покажем, как она связана с моделью медиа-веба, предложенной в главе 2.3.

Мы имеем дело с эфемерными страницами, т.е. такими страницами, к которым пользовательский интерес возрастает в течении часов после появления, но длится лишь несколько дней (см. Рис. 3.1). Медиа-веб, который обсуждался в главе 2.3, богат подобными страницами.

Предположим, что для каждой страницы  $i$  мы знаем убывающую функцию  $P_i(\Delta t)$ , которая есть “польза” от ее скачивания с задержкой  $\Delta t$  секунд после ее появления  $t_i$  (под пользой можно иметь ввиду количество “кликов”

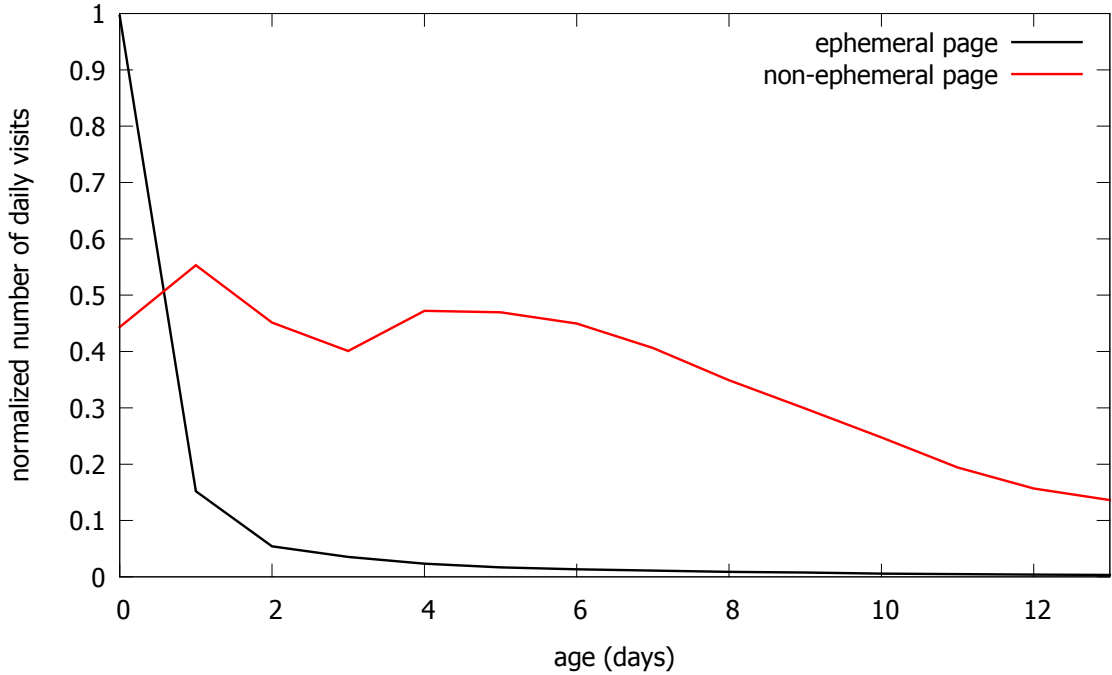


Рис. 3.1: Типичные паттерны пользовательского интереса для эфемерных и неэфемерных страниц

или “показов”, которые страница соберет в поисковой выдаче). Если в итоге каждая страница  $i$  была скачана с задержкой  $\Delta t_i$ , мы можем определить *динамическое качество* поискового робота:

$$Q_T(t) = \frac{1}{T} \sum_{i: t_i + \Delta t_i \in [t-T, t]} P_i(\Delta t_i). \quad (3.1)$$

Иными словами, динамическое качество — это средняя польза, которую приносит поисковый робот во временном окне размера  $T$ . Динамическое качество может быть полезно для того, чтобы понять влияние недельных и суточных трендов на текущую полезность поискового робота.

Определим теперь *среднее качество* поискового робота. Естественно ожидать, что, если выбрать временное окно  $T$  достаточно большим, влияние временных трендов пользовательского интереса на динамическое качество поискового робота усреднится. Иными словами, функция  $Q_T(t)$  стремится к константе, когда  $T$  увеличивается. Таким образом, в предположении стационарности, мы можем рассмотреть *среднее качество* поискового робота:

$$Q = \lim_{T \rightarrow \infty} Q_T(t) \quad (3.2)$$

которое не зависит от  $t$ .

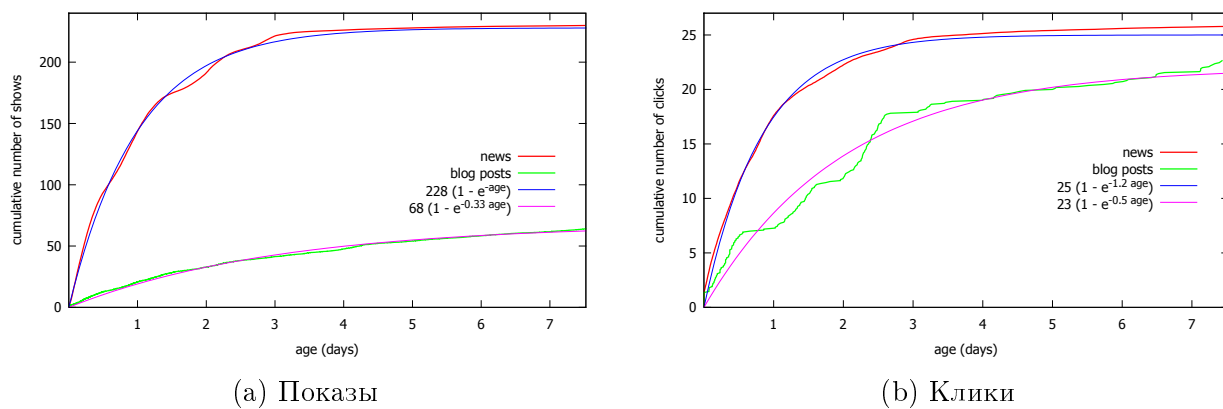


Рис. 3.2: Среднее количество общих кликов и показов в зависимости от возраста страницы

В этой работе под пользой  $P_i(\Delta t)$  от скачивания страницы  $i$  в момент времени  $t_i + \Delta t$  мы будем иметь в виду общее количество кликов пользователей, которое она получит в поисковой выдаче после момента скачивания (мы пренебрегаем временем индексации). Таким образом, мы можем оценить, насколько страница соответствует текущим интересам пользователей. Альтернативно мы могли бы использовать показы вместо кликов, но клики в меньшей степени зависят от текущего алгоритма ранжирования. К тому же, как мы увидим дальше, клики и показы очень коррелированы, поэтому не так важно, кого из них выбрать.

Итак, наша целевая метрика определена, мы будем оптимизировать ее, чтобы найти оптимальное расписание переобхода источников контента, о которых мы подробно поговорим в следующем разделе. Заметим, что на практике эта метрика может быть вычислена только с некоторой задержкой. Дело в том, что для того, чтобы учесть пользу от недавно скачанной страницы, мы должны подождать, пока она перестанет быть интересной пользователям и они прекратят на нее кликать.

Таким образом, в момент обнаружения ссылки поисковый робот не может знать ее истинного качества, т.е. сколько раз пользователи в будущем кликнут на страницу, на которую введет ссылка, если сейчас он эту страницу скачает. Однако, для принятия решения о скачивании роботу придется каким-то образом оценить это будущее качество. Итак, мы не знаем функцию  $P_i(\Delta t)$  для только что появившихся страниц, но мы можем попытаться предсказать ее. Естественно ожидать, что страницы похожей природы имеют схожее поведение пользовательского интереса. Чтобы продемонстрировать это, на Рис. 3.2a и Рис. 3.2b мы построили среднее количество общих кликов и показов в зависимости от возраста страницы для всех страниц, опубликованных в течение недели на случайно выбранном новостном сайте и блоге.



Мы можем заметить, что большинство кликов и показов приходится на первую неделю жизни страницы и что полученная зависимость довольно хорошо описывается функцией  $P(1 - e^{-\mu \cdot \Delta t})$ , где  $P$  — это общее число кликов (показов), полученных страницей за ее время жизни. Мы, таким образом, предлагаем следующую аппроксимацию для  $P_i(\Delta t)$  (т.е. количества будущих кликов):

$$P_i(\Delta t) \approx P_i \cdot e^{-\mu_i \cdot \Delta t},$$

где *скорость устаревания*  $\mu_i$  и *польза*  $P_i$  зависят от источника контента и могут быть оценены из исторических данных (см. детали в параграфе 3.3.2). Мы используем эту аппроксимацию в разделе 3.3 для поиска оптимального расписания переобхода источников контента и скачивания новых страниц.

Заметим, что используемая нами аппроксимация функции полезности фактически является лучшим вариантом функции привлекательности, который мы нашли в главе 2.3, а именно  $f(d, q, a) = q \cdot e^{-\frac{a}{\tau}}$ . Таким образом, используя несколько другую терминологию, мы приходим к одинаковым выводам. Также в главе 2.3 мы предполагали, что новые страницы в медиа-вебе появляются на хостах, причем скорость появления зависит от хоста, в следующем разделе мы уточним механизм их происхождения.

## 3.2 Источники контента

В этом разделе мы покажем, что большинство нового эфемерного (быстроустаревającego) контента может быть найдено на небольшом количестве источников, и предложим метод их нахождения, подходящий для нашей задачи.

Наша гипотеза состоит в том, что большинство новых эфемерных страниц появляется в вебе на небольшом множестве источников контента (главные страницы хостов, рубрики новостных сайтов и т.д.), и поисковому роботу по этому достаточно регулярно переобходить это небольшое множество источников, чтобы быстро обнаруживать новые эфемерные страницы.

Для того, чтобы проверить эту гипотезу, нам нужно исследовать эволюцию ссылочного графа веба во времени. Мы не могли использовать для проверки нашей гипотезы логи текущего поискового робота. Дело в том, что ссылки с источников на эфемерные страницы короткоживущие, поэтому поисковый робот должен достаточно часто переобходить источники, чтобы не пропустить эти ссылки. Текущий же робот не был заточен под решение рассматриваемой проблемы и поэтому, пропуская ссылки, обладал не точным

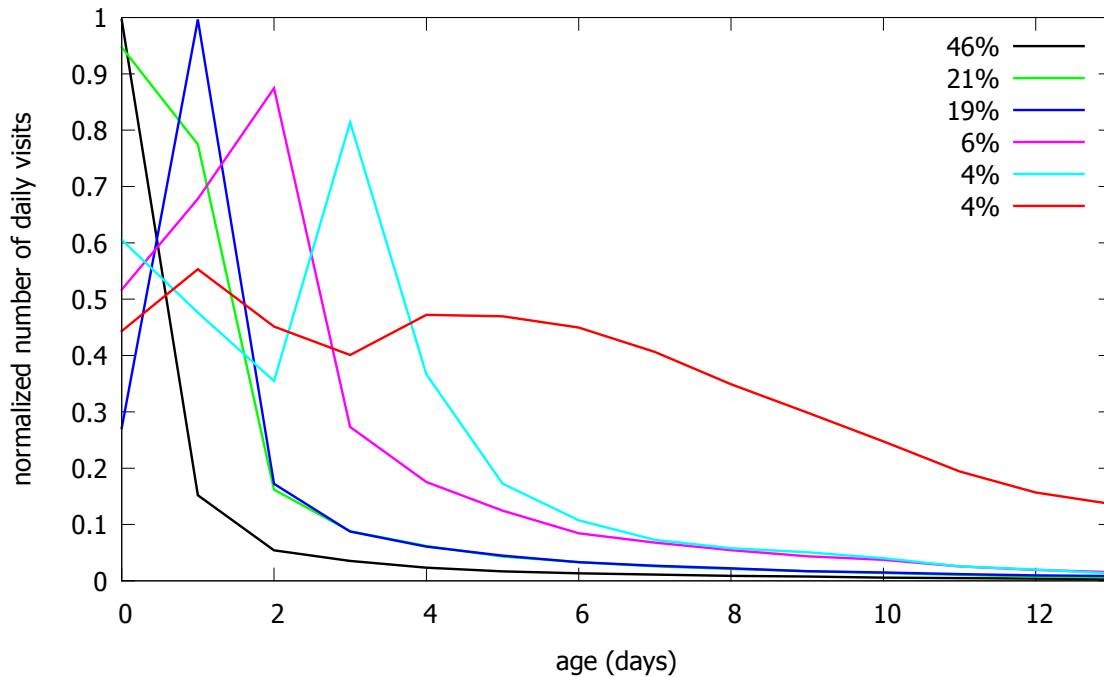


Рис. 3.3: Паттерны пользовательского интереса к новым страницам.

знанием об эволюции ссылочного графа.

Вместо этого мы взяли логи Яндекс.Бара, который применяется для того, чтобы отслеживать визиты пользователей на новые страницы. Используя эти логи, мы можем легко отслеживать момент появления новых страниц, интересных хотя бы одному пользователю Яндекс.Бара, узнавать, какие источники контента ссылаются на него, а также следить за эволюцией пользовательского интереса к страницам во времени. Эти данные репрезентативно описывают ссылочный граф веба, так как Яндекс.Баром пользуются миллионы людей в России и странах СНГ. Заметим, что мы используем эти данные только для проверки нашей гипотезы, сам алгоритм не должен на них полагаться, так как интересно разработать подход, применимый и для тех стран, где Яндекс.Бара пока нет.

Используя данные Яндекс.Бара, мы случайно выбрали  $50 \cdot 10^3$  страниц из тех, что появились в течение одной недели и были посещены хотя бы одним пользователем. Эти страницы были распределены по  $\sim 1.6 \cdot 10^3$  разным хостам. Для каждой страницы мы нашли визиты на нее в течение двух недель после ее появления. Далее, используя вектор признаков (нормализованный так, чтобы максимальное значение равнялось 1) размерности 14 (по числу дней), мы кластеризовали эти страницы на 6 кластеров с помощью метода k-средних. Заметим, что, когда мы пытались использовать меньшее число

кластеров, отдельный кластер неэфемерных страниц был не виден. Наконец, мы получили только  $\sim 4\%$  неэфемерных страниц. Таким образом, процент эфемерных страниц (посещенных хотя бы однажды за период наблюдений) равен  $96\%$ , что очень значительно. Центроиды этих кластеров показаны на Рис. 3.3 (на Рис. 3.1 мы показали только два из них).

В большинстве случаев логи Яндекс.Бара также содержат запись о странице, с которой пользователь попал на целевую страницу. Таким образом, логи Яндекс.Бара также дают информацию о ссылочной структуре веба, причем время первого перехода хорошо аппроксимирует время появления ссылки. Мы извлекли все такие *ссылки* (пользовательские переходы) из логов, собранных за три недели (в течение дополнительных двух недель большинство страниц устарело и прекратило получать новые ссылки), и получили  $\sim 750 \cdot 10^3$  ссылок.

Используя эти ссылки, мы изучили процент покрываемых страниц в зависимости от числа выбранных источников контента. Мы хотим найти наименьшее множество источников контента, ссылающихся на большинство появившихся новых эфемерных страниц, и поэтому предлагаем следующий жадный алгоритм. Мы сначала выбираем источник, покрывающий наибольшее число новых страниц, потом удаляем его и все покрытые страницы, потом таким же образом выбираем второй источник и т.д.. На Рис. 3.4 мы видим, что всего  $3 \cdot 10^3$  источников достаточно, чтобы покрыть  $80\%$  новых страниц, этот факт подтверждает нашу гипотезу о малом количестве источниках контента. Интересно, что  $42\%$  из этих  $3 \cdot 10^3$  источников — главные страницы хостов, а  $44\%$  являются рубриками, на которые можно перейти с главной страницы хоста. Таким образом,  $86\%$  источников контента находятся на удалении не более 1 хопа от главной страницы хоста. Это натолкнуло нас на идею использовать следующий простой метод для нахождения источников контента:

1. скачать главную страницу каждого интересующего хоста, а также все страницы, ссылки на которые будут найдены на главной странице;
2. выбрать главную страницу и все найденные страницы старше нескольких дней в качестве источников контента (молодые страницы скорее всего не будут источниками контента).

Эту процедуру можно запускать периодически, чтобы обновлять список источников. Заметим, что в поиске источников контента нас больше всего волнует полнота, так как наш алгоритм сам оптимизирует точность, прекращая обходить некачественные источники контента (см. раздел 3.3).

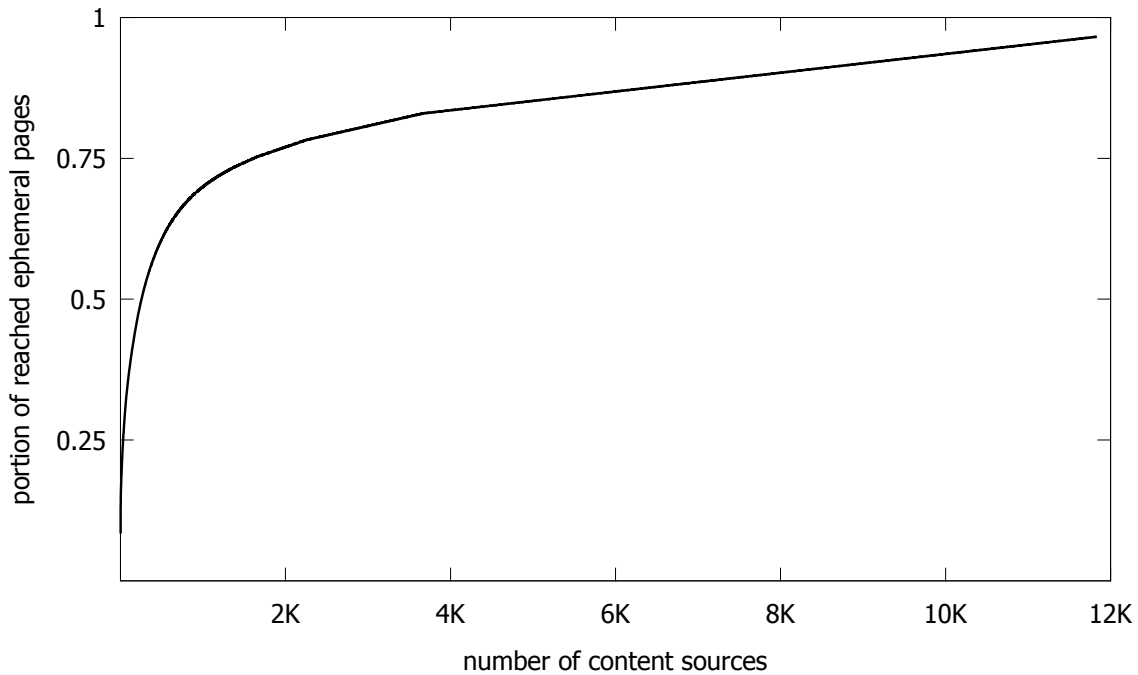


Рис. 3.4: Доля покрываемых новых эфемерных страниц в зависимости от среднего числа источников контента, взятого на хосте.

### 3.3 Оптимальный обход источников

В этом разделе мы предполагаем, что нам дан сравнительно небольшой список источников контента, которые генерируют большинство качественного эфемерного контента (в разделе 3.2 обсуждалось, как можно найти такой список). Наши текущие задачи следующие: (1) найти оптимальное расписание переобхода источников контента, которое позволит быстро находить появляющиеся новые качественные страницы, и (2) понять, как распределить ресурсы между обходом новых страниц и переобходом источников контента.

Сначала мы формализуем нашу проблему в виде оптимизационной задачи и находим ее оптимальное решение. Затем мы описываем практический алгоритм, основанный на этом решении.

#### 3.3.1 Теоретический анализ

Предположим, что нам дано множество источников контента  $S_1, \dots, S_n$ . Заметим, что скорость появления нового контента может отличаться от источника к источнику. Например, обычно публикуется гораздо больше новостей о политике, чем об искусстве, т.е. обычно разные категории новостных сайтов с разной скоростью публикуют новый контент. Пусть  $\lambda_i$  — это *скорость появления новых ссылок* на источнике  $S_i$ , т.е. среднее количество ссы-

лок на новые страницы, появляющихся в секунду.

Рассмотрим алгоритм, который переобходит источник  $S_i$  каждые  $I_i$  секунд, находит ссылки на новые страницы, а затем скачивает все эти новые страницы. Мы хотим найти такое расписание переобхода источников, которое максимизировало бы среднее качество  $Q$  (см. выражение (3.2)), т.е., оптимальные значения  $I_i$ . Предположим, что наша инфраструктура позволяет обходить  $N$  страниц в секунду ( $N$  может быть не целым). Это ограничение на ресурсы приводит к следующему ограничению на интервалы переобхода:

$$\sum_i \frac{1 + \lambda_i I_i}{I_i} \leq N.$$

В среднем количество страниц, найденных после переобхода на источнике  $S_i$ , равняется  $\lambda_i I_i$ , поэтому каждые  $I_i$  секунд нам приходится обходить  $1 + \lambda_i I_i$  страниц (сам источник и все новые страницы, найденные на нем). Очевидно, что оптимальное решение потребует расходовать все имеющиеся ресурсы:

$$\sum_i \frac{1}{I_i} = N - \sum_i \lambda_i. \quad (3.3)$$

И мы хотим максимизировать среднее качество, т.е.,

$$Q = \sum_i \frac{1}{I_i} \sum_{j: p_j \in S_i \wedge t_j \in [0, I_i]} P_j(\Delta t_j) \rightarrow \max.$$

Заметим, что это выражение есть в точности средняя общая польза, приносимая роботом в единицу времени.

Источники контента могут отличаться по качеству, т.е., некоторые источники могут обеспечивать пользователей более качественными ссылками. Предположим теперь, что страницы, найденные на одном источнике, в среднем имеют похожее поведение убывания пользы со временем, и поэтому заменяем  $P_j(\Delta t_j)$  приближением  $P_i e^{-\mu_i \Delta t_j}$ , которое обсуждалось в разделе 3.1. Мы считаем, что общая польза  $P_i$  и скорость убывания пользы  $\mu_i$  являются параметрами источника  $S_i$ . Таким образом, мы получаем:

$$\begin{aligned} Q &= \sum_i \frac{P_i}{I_i} \sum_{j=0}^{\lambda_i I_i - 1} e^{-\mu_i \frac{j}{\lambda_i}} = \\ &= \sum_i \frac{P_i}{I_i} \frac{1 - e^{-\mu_i I_i}}{1 - e^{-\frac{\mu_i}{\lambda_i}}} = \sum_i p_i x_i \left(1 - e^{-\mu_i/x_i}\right), \end{aligned}$$

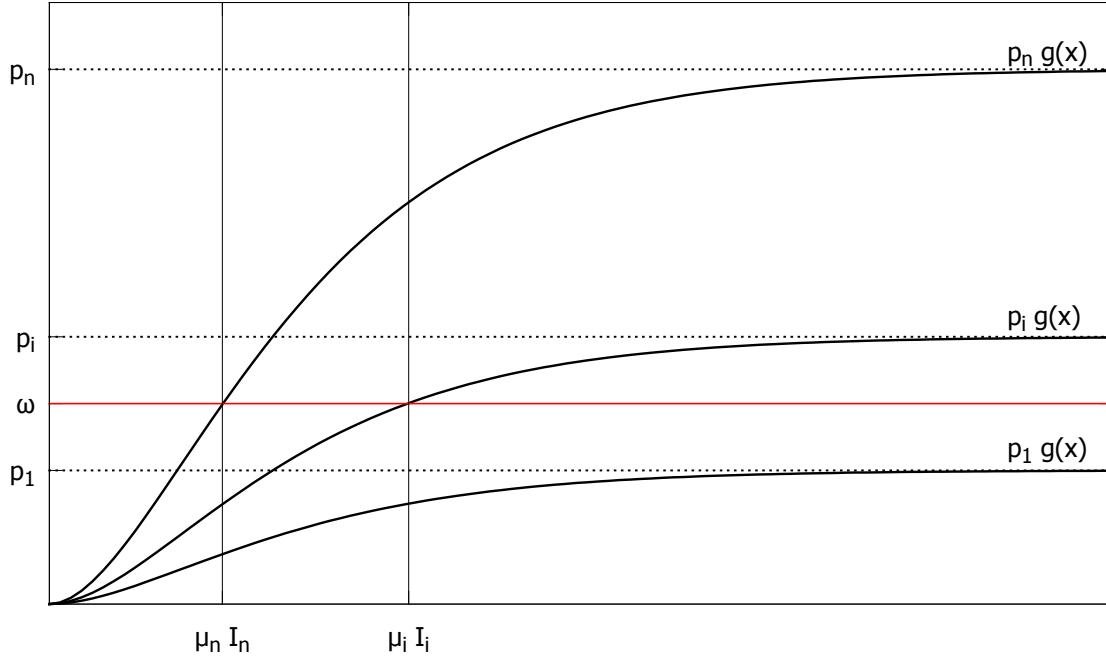


Рис. 3.5: Оптимизация  $I_i$

где  $p_i = \frac{P_i}{1 - e^{-\frac{\mu_i}{\lambda_i}}}$  и  $x_i = \frac{1}{I_i}$ . Без ограничения общности, предполагаем, что  $p_1 \leq \dots \leq p_n$ . Теперь для максимизации  $Q(x_1, \dots, x_n)$  при условии (3.3) мы используем метод множителей Лагранжа:

$$\begin{cases} p_i (1 - e^{-\mu_i/x_i}) - \frac{\mu_i p_i}{x_i} e^{-\mu_i/x_i} = \omega, & i = 1, \dots, n, \\ \sum_i x_i = N - \sum_i \lambda_i, \end{cases}$$

где  $\omega$  — это множитель Лагранжа.

Вспоминая, что  $I_i = \frac{1}{x_i}$ , получаем

$$\begin{cases} p_i (1 - (1 + \mu_i I_i) e^{-\mu_i I_i}) = \omega, & i = 1, \dots, n, \\ \sum_i \frac{1}{I_i} = N - \sum_i \lambda_i. \end{cases} \quad (3.4)$$

Функция  $g(x) = (1 - (1 + x)e^{-x})$  монотонно возрастает для  $x > 0$ , причем  $g(0) = 0$  и  $g(+\infty) = 1$ . Следовательно, для любого  $\omega$  ( $0 < \omega < p_i$ ) существует единственное  $\mu_i I_i = g^{-1}(\frac{\omega}{p_i})$ , как показано на Рис. 3.5. Обратную функцию  $g^{-1}$  можно легко вычислить, используя бинарный поиск. Так как,  $\mu_i I_i$  монотонно возрастающая функция  $\omega$ , то  $\sum_i \frac{1}{I_i}$  является монотонной функцией  $\omega$ , и мы опять используем бинарный поиск (см. алгоритм 1) для того, что удовлетворить условию  $\sum_i \frac{1}{I_i} = N - \sum_i \lambda_i$ .

---

**Algorithm 1:** Поиск оптимального расписания переобхода

---

**input** : общие полезности  $P_i$ , скорости убывания полезности  $\mu_i$ , скорости появления новых ссылок  $\lambda_i$ , количество скачиваний в секунду  $N$ , точность  $\varepsilon$

**output:** оптимальные периоды переобхода  $I_i$

$$\omega_l \leftarrow 0; \omega_u \leftarrow p_n = \frac{P_n}{1 - e^{-\frac{\mu_n}{\lambda_n}}};$$

**while**  $\left| \sum_{i:I_i \neq \infty} \frac{1}{I_i} - N + \sum_{i:I_i \neq \infty} \lambda_i \right| > \varepsilon$  **do**

$$\omega \leftarrow \frac{\omega_u + \omega_l}{2};$$

$$I_i \leftarrow \frac{1}{\mu_i} g^{-1} \left( \frac{\omega}{p_i} \right) = \frac{1}{\mu_i} g^{-1} \left( \omega \frac{1 - e^{-\frac{\mu_i}{\lambda_i}}}{P_i} \right);$$

**if**  $\sum_{i:I_i \neq \infty} \frac{1}{I_i} < N - \sum_{i:I_i \neq \infty} \lambda_i$  **then**

$$\omega_u \leftarrow \omega;$$

**else**

$$\omega_l \leftarrow \omega;$$

---

Пусть  $\omega_l$  и  $\omega_u$  — это, соответственно, нижняя и верхняя граница  $\omega$ . На первом шаге мы можем положить  $\omega_l = 0$ , а  $\omega_u = p_n$ . Действительно,  $p_n$  — очевидная верхняя граница для  $\omega$ , так как в этом случае мы совсем отказываемся от обхода источников. На каждом шаге мы рассматриваем значение  $\omega = \frac{\omega_u + \omega_l}{2}$ . Для этого значения  $\omega$  мы пересчитываем интервалы  $I_i = \frac{1}{\mu_i} g^{-1} \left( \frac{\omega}{p_i} \right)$ . Заметим, что, если мы получим  $\omega > p_j$  для некоторого  $j$ , то  $I_j = \infty$  и мы решаем никогда не обходить этот источник. После этого, если  $\sum_{i:I_i \neq \infty} \frac{1}{I_i} < N - \sum_{i:I_i \neq \infty} \lambda_i$ , то мы можем взять текущую верхнюю границу  $\omega_u = \omega$ . Иначе мы полагаем  $\omega_l = \omega$ . Мы продолжаем в том же духе, пока не достигнем желаемой точности  $\varepsilon$ .

Полученное значение  $\omega$  может быть интерпретировано как ограничение, которое мы накладываем на минимальную полезность источника контента. Мы даже можем найти требование на минимальное количество страниц, которые должен уметь обходить поисковый робот, чтобы нам не пришлось полностью отказаться от обхода какого-то из источников.

Мы полностью решили оптимизационную задачу для метрики, предложенной в разделе 3.1, решение (3.4) является теоретически оптимальным (мы будем использовать название *ECHO-based* для полученного алгоритма, где ECHO — это аббревиатура для Ephemeral Content Holistic Ordering). Однако, нам еще потребуются некоторые усилия, чтобы сделать описанный алгоритм практически применимым. Для каждого из источников нам надо оценить следующие параметры: среднюю общую полезность  $P_i$ , скорость убывания

полезности  $\mu_i$  и скорость появления новых ссылок  $\lambda_i$ . В следующем параграфе мы опишем конкретный практический алгоритм, основанный на нашем теоретическом анализе.

### 3.3.2 Реализация

Мы используем результаты предыдущих разделов, чтобы получить конкретный алгоритм обхода Интернета. Сначала мы используем процедуру, описанную в разделе 3.2, чтобы получить множество источников контента. Далее, для того, чтобы применить алгоритм 1 для нахождения оптимального расписания переобхода источников, для каждого источника нам надо знать среднюю общую полезность  $P_i$ , скорость убывания полезности  $\mu_i$  и скорость появления новых ссылок  $\lambda_i$ . Мы предлагаем оценивать все эти параметры динамически, используя историю переобхода источников и поисковые логи. Так как эти параметры постоянно меняются, мы должны периодически переоценивать периоды переобхода  $I_i$  (см. алгоритм 1), т.е., обновлять расписание переобхода. Ясно, что, чем чаще мы переоцениваем  $I_i$ , тем лучше результаты мы получим, и выбор периода переоценивания зависит от доступных вычислительных ресурсов.

Таким образом, мы сначала обсудим, как оценивать эти параметры, а потом, что делать в случае отклонения источников контента от идеалистического поведения, предполагаемого при теоретическом анализе.

#### Оценивание средней общей полезности $P_i$ и скорость убывания полезности $\mu_i$

В этой части используем поисковые логи для того, чтобы анализировать историю кликов пользователей на новые страницы в поисковой выдаче. Мы аппроксимируем среднее общее число кликов на страницу в зависимости от ее возраста экспоненциальной функцией (см. Рис. 3.2b).

Предположим, что мы построили накопительную гистограмму (с корзиной размером  $D$  минут) количества кликов на все страницы, найденные на некотором источнике контента в зависимости от их возраста. Пусть  $s_i$  — это количество раз, которое все  $N$  найденных на источнике страниц были кликнуто в течении первых  $iD$  минут после их появления. Тогда  $s_i/N$  — это число кликов, которое в среднем страницы на выбранном источнике получают в первые  $iD$  минут своей жизни.

Мы теперь можем использовать метод наименьших квадратов, т.е., нам



нужно найти:

$$\arg \min_{\mu, P} F(P, \mu) = \arg \min_{\mu, P} \sum_i \left( P (1 - e^{-\mu i D}) - \frac{s_i}{N} \right)^2. \quad (3.5)$$

Другими словами, мы хотим найти значения параметров  $\mu$  и  $P$ , которые минимизируют квадратичное отклонение среднего числа кликов на страницу и его приближения  $P (1 - e^{-\mu i D})$ . Найти аналитическое решение (3.5) представляет сложным, но мы можем использовать градиентный спуск:

---

**Algorithm 2:** Estimate profit decay function

---

**input** : размер корзины гистограммы  $D$ , общее число кликов  $s_i$ , количество найденных новых страниц  $N$ , точность  $\varepsilon$ , величина шага  $\gamma$ , начальные значения  $P_{init}$  и  $\mu_{init}$

**output:** средняя полезность  $P$ , скорость убывания полезности  $\mu$

$P_{old} \leftarrow 0; P \leftarrow P_{init}; \mu_{old} \leftarrow 0; \mu \leftarrow \mu_{init};$

**while**  $\max\{|P_{old} - P|, |\mu_{old} - \mu|\} > \varepsilon$  **do**

$P_{old} \leftarrow P;$   
 $\mu_{old} \leftarrow \mu;$   
 $P \leftarrow P_{old} - \gamma \frac{\partial F}{\partial P}(P_{old}, \mu_{old});$   
 $\mu \leftarrow \mu_{old} - \gamma \frac{\partial F}{\partial \mu}(P_{old}, \mu_{old});^4$

---

С точки зрения практической реализации важно понять, как часто надо обрабатывать поисковые логи, чтобы переоценить значения  $\mu_i$  и  $P_i$ , так как эта операция является достаточно дорогой. Мы обозначим *период обработки логов* через  $L$ . В разделе 3.4 мы анализируем влияние  $L$  на качество алгоритма.

### Оценка скорости появления новых страницы $\lambda_i(t)$

Скорость появления новых ссылок  $\lambda_i(t)$  меняется в течение дня или недели. Поэтому мы динамически оцениваем эту скорость для каждого источника контента. Для этого мы используем исторические данные: а именно, информацию о количествах страниц, найденных на источнике за последние  $T$  переобходов. Мы анализируем влияние  $T$  на качество алгоритма в разделе 3.4.

### Расписание

Наконец, чтобы применить наш алгоритм, нам необходимо решить следующую проблему: в реальности количество ссылок, которые мы найдем на

---

<sup>4</sup>  $\frac{\partial F}{\partial P}(P, \mu) = 2 \sum_i \left( P (1 - e^{-\mu i D}) - \frac{s_i}{N} \right) (1 - e^{-\mu i D})$   
 $\frac{\partial F}{\partial \mu}(P, \mu) = 2 \sum_i \left( P (1 - e^{-\mu i D}) - \frac{s_i}{N} \right) i P D e^{-\mu i D}$

источнике после переобхода, может отличаться от ожидаемого  $\lambda_i I_i$ . Мы можем найти больше ссылок, чем ожидали, и, если мы будем скачивать их все, то можем отклониться от расписания. Поэтому мы не можем одновременно строго следовать расписанию и скачивать все найденные страницы. Мы предлагаем следующие два способа решения этой проблемы.

**ЕСНО-newpages.** Для того, чтобы не терять клики, мы сразу скачиваем новые страницы после обнаружения ссылки на них. Если еще необойденных новых страниц не осталось, мы пытаемся догнать расписание. Мы всегда обходим тот источник контента, который больше всего отклоняется от расписания, т.е. с наибольшим значением  $I'_i/I_i$ , где  $I'_i$  — это время, прошедшее после последнего переобхода  $i$ -го источника контента.

**ЕСНО-schedule.** Мы всегда переобходим источники согласно периодам  $I_i$ , и, когда у нас есть свободные ресурсы, мы обходим найденные новые страницы (наиболее свежие сначала).

В следующем разделе мы экспериментально сравним два этих подхода.

### Возможная архитектура

Мы завершаем этот раздел описанием возможной продакшн-архитектуры для нашего алгоритма (см. Рис. 3.6), тем самым, подчеркивая практическую значимость описываемого здесь подхода. Вначале создается список источников контента в соответствии с разделом 3.2. Эту процедуру можно запускать периодически, чтобы обновлять список источников контента. Затем *Scheduler* (планировщик) находит оптимальное расписание переобхода источников контента, а *Fetcher* (сборщик) обходит источники в соответствии с этим расписанием. *Scheduler* динамически оценивает скорость появления новых ссылок на источниках и также оценивает функцию убывания полезности для новых страниц источника, используя информацию о кликах, получаемых из поисковых логов. Таким образом, мы динамически подстраиваемся под текущие интересы пользователей, а также учитываем временные тренды в скорости появления новых ссылок на источниках контента.

## 3.4 Эксперименты

В этом разделе мы на реальных данных сравним наш алгоритм с некоторыми другими подходами.

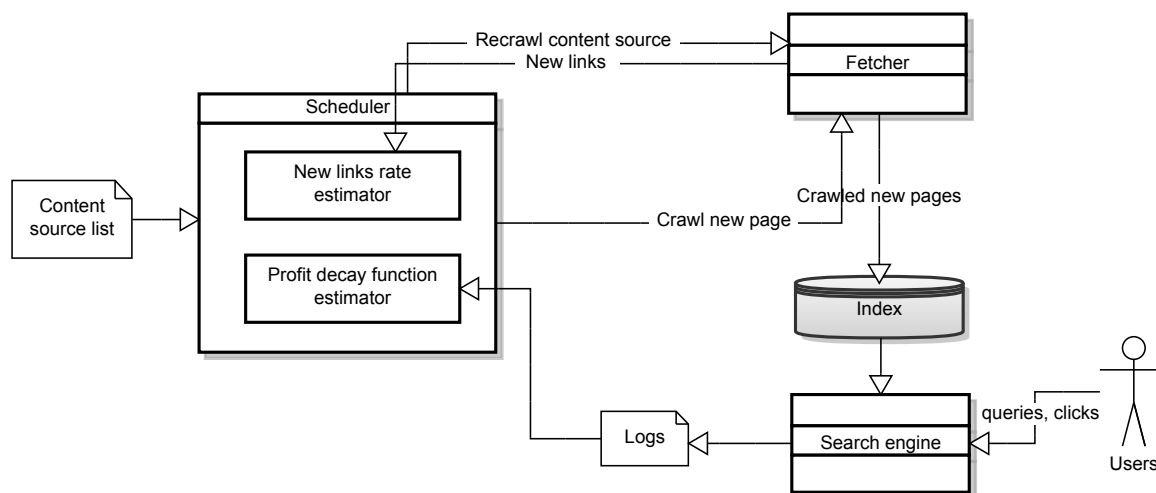


Рис. 3.6: Возможная продакшн-архитектура

### 3.4.1 Данные

Мы выбрали несколько сайтов, которые составляют достаточно большое репрезентативное подмножество веба, и использовали их для наших экспериментов.

А именно, мы выбрали 100 наиболее посещаемых российских новостных сайтов и 50 российских блогов, используя открытую информацию из доверенных источников<sup>5</sup>. Мы считаем, что это множество достаточно репрезентативно для нашей задачи, так как порождает 5-6% из  $\sim 500 \cdot 10^3$  новых страниц (посещенных хотя бы одним пользователем Яндекс.Бара), которые ежедневно появляются в российском вебе. Для каждого сайта мы применили процедуру из раздела 3.2 и получили  $\sim 3 \cdot 10^3$  источников контента.

Затем мы переобходим эти источники каждые 10 минут в течение трех недель (что достаточно часто, чтобы находить все ссылки на них до того, как они исчезнут). Время обнаружения новой ссылки, таким образом, отличается не больше чем на 10 минут от реального времени ее появления в вебе. Мы считали страницы, найденные при первом обходе источников старыми (каждый источник был обойден  $\sim 3 \cdot 10^3$  раз) и обнаружили  $\sim 415 \cdot 10^3$  новых страниц в течение этих 3 недель. Следя за тем, когда новые ссылки появляются и пропадают, мы создали динамический граф Интернета. Этот граф содержит  $\sim 2.4$ М уникальных ссылок.

Дополнительно мы использовали поисковые логи Яндекса, чтобы полу-

<sup>5</sup>

<http://liveinternet.ru/rating/ru/media/>  
<http://blogs.yandex.ru/top/>

чить клики пользователей на найденные страницы, собранные за эти 3 недели и еще 1 неделю, необходимую для того, чтобы большинство найденных страниц устарело. Мы обнаружили, что  $\sim 20\%$  этих страниц были кликнуты хотя бы один раз за эти 4 недели.

### 3.4.2 Упрощения предложенного алгоритма

Мы сравним алгоритм из раздела 3.3 с некоторыми другими подходами. Отметим, что так как рассматриваемая проблема новая, то нам не известны какие-то хорошие классические подходы к ее решению, но мы рассмотрим несколько естественных идей:

- **Breadth-first search (BFS)** Мы переобходим источники контента последовательно в каком-то фиксированном порядке. После переобхода источника мы обходим все найденные на нем ссылки, которые не были обойдены до этого.

Мы также упрощаем наш алгоритм разными способами, чтобы понять 1) важность единого упорядочения обхода и 2) полезность кликового сигнала.

- **Fixed-quota** Этот алгоритм похож на *ECHO*, но мы используем фиксированную квоту  $\frac{1}{2}$  для переобхода источников и обхода новых страниц, т.е., мы с вероятностью  $\frac{1}{2}$  либо обходим источник, наиболее отклонившийся от расписания, либо самую свежую новую страницу.
- **Frequency** Этот алгоритм также похож на *ECHO*, но мы не используем клики из поисковых логов, т.е., мы считаем, что все источники имеют одинаковое качество и отличаются только в скорости публикации новых ссылок.

Мы также предлагаем следующее упрощение нашего алгоритма, которое может быть гораздо удобнее в практической реализации.

- **ECHO-greedy** Мы игнорируем скорость убывания полезности страниц и обходим источник с самой большой ожидаемой суммарной полезностью найденных страниц, т.е., с наибольшим значением  $\lambda_i P_i I'_i$ , где  $I'_i$  — это время, прошедшее с последнего переобхода источника,  $\lambda_i$  — это скорость появления новых страниц, а  $P_i$  — средняя общая полезность страниц источника. Затем мы обходим все найденные новые страницы и повторяем процедуру.

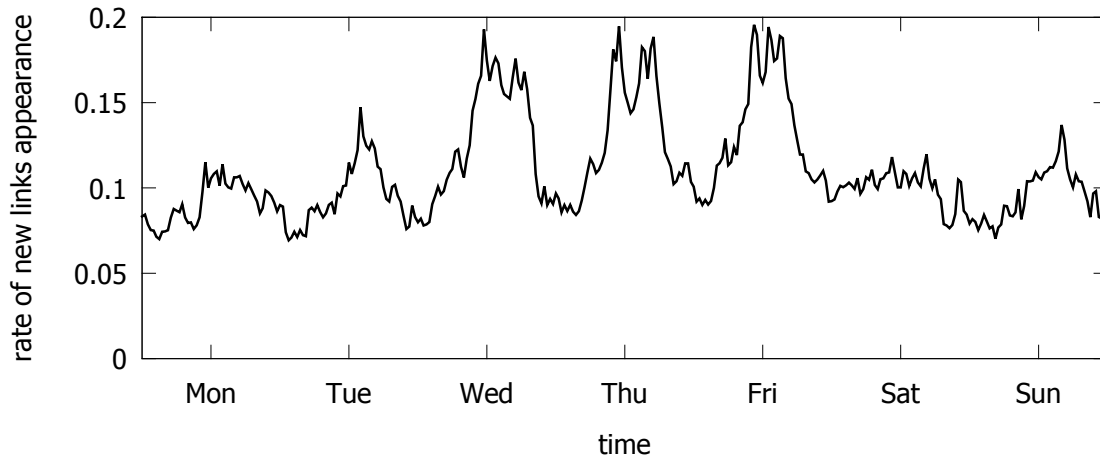


Рис. 3.7: Оценивание скорости появления нового контента.

### 3.4.3 Результаты

#### Схема эксперимента

В этом параграфе, используя реальные данные, мы исследуем влияние параметров на качество нашего алгоритма и сравниваем его с другими подходами, изложенными в параграфе 3.4.2. Мы симулируем работу каждого алгоритма на динамическом графе, описанном в параграфе 3.4.1.

В наших экспериментах по изучению влияния параметров мы использовали скорость обхода, равную  $N = 0.1$  страница в секунду. Как показано на Рис. 3.7, эта скорость достаточна для обхода значительной части появляющихся новых страниц, но не слишком велика, чтобы BFS алгоритм смог обходить все новые страницы (нереалистичная ситуация). Мы также попробовали скорости обхода  $N = 0.05$  и  $N = 0.2$ , чтобы лучше понять влияние этой скорости.

#### Влияние параметров

Мы применяем алгоритм 1 для переоценивания периодов обхода  $I_i$  каждые 30 минут, это достаточно часто, так как меньшие периоды практически не влияют на качество, но также и реалистично с точки зрения практической реализации. Мы взяли размер корзинок  $D$ , используемый в алгоритме 2, равным 20 минутам, что достаточно для робастных оценок  $P_i$  и  $\mu_i$ , т.к. обычно функция убывания полезности  $P_i(\Delta t)$  незначительно меняется в течение таких малых периодов времени. Мы не исследуем здесь эти параметры детально, потому что выбор параметров, величина которых меньше этих ре-

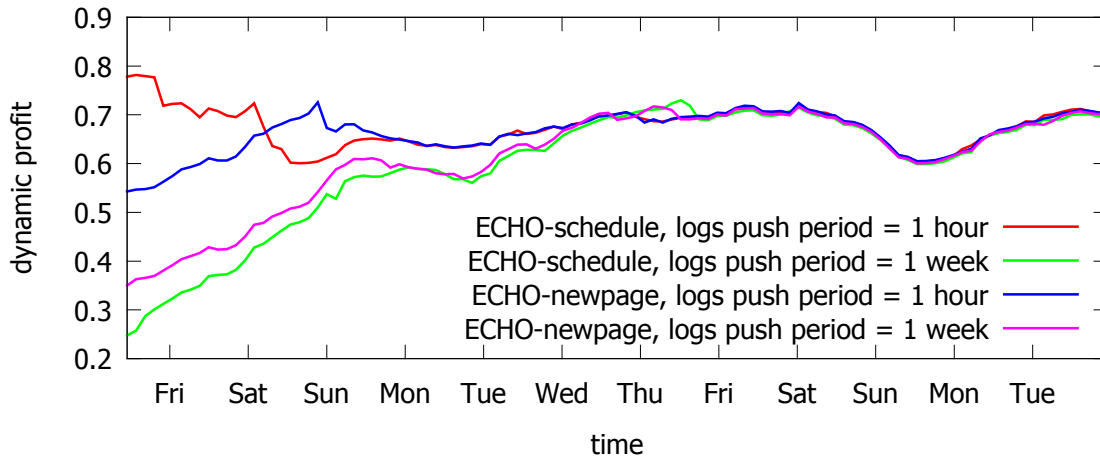


Рис. 3.8: Динамическое качество для временного окна в 1 неделю.

листочных оценок, оказывает пренебрежимое влияние на качество алгоритма в сравнении с остальными параметрами.

Кроме того, нам необходимы начальные значения для  $P_i$ , так как вначале мы не имеем исторических данных. Заметим, что лучше использовать *пессимистичные* начальные значения, так как это позволит избежать слишком частого обхода некачественных источников, в то время пока мы не накопили достаточно исторических данных для точных оценок. Мы не можем использовать  $P_{default} = 0$ , так как согласно алгоритму 1, мы не обходим источники нулевой полезности, поэтому мы использовали маленькое ненулевое значение  $P_{default} = 0.01$ .

Мы сравниваем два варианта ECHO алгоритма, описанного в параграфе 3.3.2, используя разные значения: 1) размера истории переобходов источников, нужных для оценки  $\lambda_i(t)$  (от 3 до 10 переобходов), и 2) периода обработки поисковых логов  $L$  (мы пробуем 1 час, 12 часов, 24 часа, и 1 неделю). Интересно, но для обоих вариантов алгоритма мы не заметили разницы в конечном качестве, поэтому мы заключили, что влиянием этих параметров на конечное качество можно пренебречь. Однако, период обработки поисковых логов имеет действительно большое влияние во время “разогрева” тогда, когда накоплено не достаточно исторической информации, причем меньшие значения периода приводят к лучшим результатам (см. Рис. 3.8). Мы не нашли чего-то интересного касательно размера истории.

Отметим также, что в оптимальном расписании, найденном ECHO алгоритмом, примерно 70% источников практически не обходятся, таким образом, алгоритм экономит ресурсы, не расходуя их на источники с низким качеством.

Таблица 3.1: Среднее динамическое качество для временного окна в 1 неделю.

| алгоритм      | N = 0.05         | N = 0.10         | N = 0.20         |
|---------------|------------------|------------------|------------------|
| Frequency     | 0.014 ± 0.004    | 0.39 ± 0.04      | 0.61 ± 0.06      |
| BFS           | 0.24±0.04        | 0.46±0.03        | 0.62±0.03        |
| Fixed-quota   | 0.43±0.04        | 0.59±0.03        | 0.69±0.03        |
| ЕСНО-greedy   | 0.60±0.03        | 0.68±0.03        | 0.69±0.03        |
| ЕСНО-schedule | 0.52±0.02        | <b>0.69±0.03</b> | <b>0.71±0.03</b> |
| ЕСНО-newpages | <b>0.62±0.04</b> | <b>0.69±0.03</b> | <b>0.71±0.03</b> |
| Оценка сверху | <i>0.72</i>      | <i>0.72</i>      | <i>0.72</i>      |

### Сравнение с другими алгоритмами

Итак, в последующих экспериментах мы берем историю обхода размером 7 и период обсчета логов длиной в 1 час (случайно, согласно обсуждению в параграфе 3.4.3), и сравниваем ЕСНО подход с другими подходами для трех разных скоростей обхода. Для того, чтобы сравнить наши алгоритмы, мы во время последней недели (после периода разогрева) каждые две минуты считали динамическое качество с окном в 1 неделю (достаточно большое для компенсации дневных трендов). В Таблице 3.1 показаны полученные средние значения и их стандартные отклонения. Мы также включили в таблицу оценку сверху для возможного качества алгоритма, которую мы получили, запустив BFS алгоритм с ресурсами, достаточными для обхода новых страниц сразу после их появления. Эта оценка сверху, таким образом, не зависит от скорости обхода и равняется 0.72.

ЕСНО-newpages показывает наилучшие результаты, которые очень близки к оценке сверху, хотя скорость обхода и гораздо меньше скорости появления новых ссылок. Это значит, что наш алгоритм эффективно расходует ресурсы и вначале обходит наиболее качественные страницы и источники. Заметим, что наименьшая скорость обхода, которая позволяет BFS достичь качества в 99% от оценки сверху, равна 1 страница в секунду (это значение измерено, но в таблице не представлено), в то время как ЕСНО-newpages и ЕСНО-schedule достигают того же результата при скорости обхода 0.2 страницы в секунду.

Заметим, что качество ЕСНО-greedy также высоко. Этот факт может служить хорошей мотивацией для использования ЕСНО-greedy в реальных поисковых системах, где простота реализации является важным требованием. Во-первых, ему требуется только очередь приоритетов, тогда как другие ЕСНО алгоритмы для обновления расписания переобхода источников используют бинарный поиск (см. алгоритм 1). Во-вторых, он не использует скорость убы-

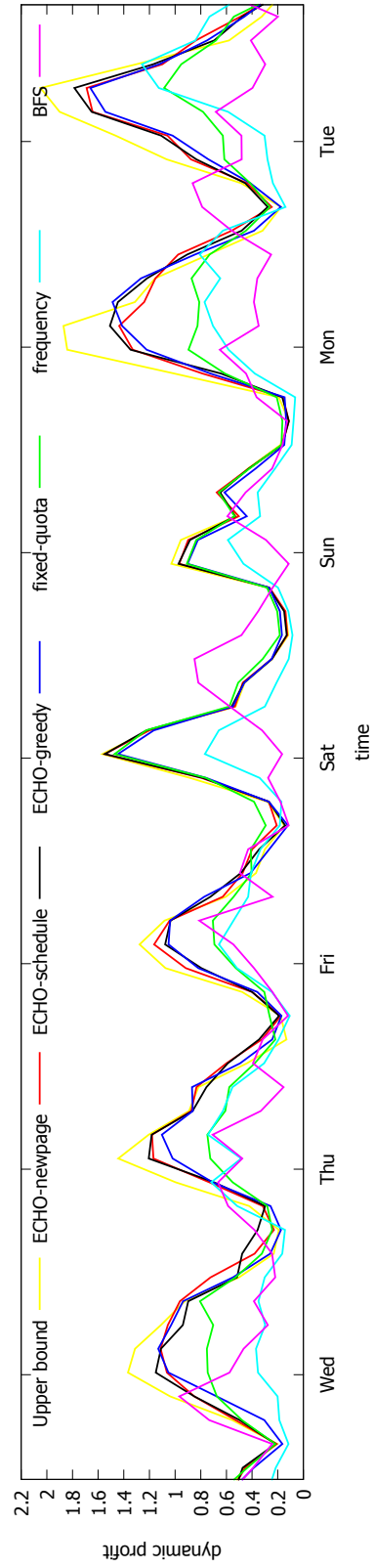


Рис. 3.9: Динамическое качество для временного окна в 5 часов.



вания полезности страниц  $\mu_i$ , тогда как  $P_i$  — это просто среднее количество кликов на страницы с  $i$ -ого источника контента, которое можно вычислить проще, чем методом градиентного спуска из алгоритма 2.

В заключение рассмотрим один показательный пример (с  $N = 0.1$ ), который демонстрирует преимущество ЕСНО алгоритмов (см. Рис. 3.9). Большинство времени ЕСНО алгоритмы выигрывают по качеству, но интересно, что в ночное время BFS показывает лучшие результаты. Это получается потому, что BFS “догоняет” остальные алгоритмы, обходя те страницы, которые другие алгоритмы уже обошли ранее. Также видно, что алгоритм с фиксированной квотой на обход и переобход неплохо работает в выходные, так как в эти дни появляется меньше новых страниц, а значит, используемой скорости обхода достаточно, чтобы обходить почти все хорошие страницы без дополнительных оптимизаций.

### 3.5 Обсуждение

Большинство работ по обходу Интернета поисковым роботом посвящены либо обнаружению новых страниц, либо обновлению изменившихся уже известных страниц. Оба эти направления до некоторой степени связаны с нашей проблемой, однако не являются ее решением.

**Политики обновления** Задача политик обновления — скачивать изменившиеся известные страницы для того, чтобы поддерживать поисковый индекс свежим. Обычно такие политики основаны на некоторой модели, которая предсказывает изменения, происходящие в Вебе. В первых роботах [12, 13, 16, 17, 18] анализ изменений страниц производился в предположении равномерного во времени пуассоновского процесса, т.е., предполагалось, что скорость изменений постоянна во времени. Однако в [12] было замечено, что в скорости изменения страниц существуют дневные и недельные тренды. Затем в [37] был предложен метод, основанный на исторических данных, который учитывал эти тренды. Более сложный подход, основанный на машинном обучении, используется в [43], где учитываются содержание страницы, величина наблюдаемых изменений и некоторые другие признаки.

Для нашей специфической задачи политики обновления могут использоваться для обнаружения новых ссылок, появляющихся на уже известных страницах (источниках контента). Таким образом, нам важны только те изменения страниц, после которых ссылки на новые страницы могут быть найдены. Интересно, что это упрощает оценку скорости изменений, так как, имея

две последовательные версии страницы, легко понять, что появились две новые ссылки, тогда как гораздо тяжелее понять, изменился ли текст страницы однажды или дважды. Этот факт позволяет нам использовать простой метод оценки скорости появления новых ссылок, который, однако, отражает временные тренды. Конечно, здесь можно применить и более сложные методы (например, использовать машинное обучение [43]), но это вне круга вопросов, подробно рассматриваемых в этой работе.

Более того, в действительности, мы мониторим изменения источника контента, чтобы не пропустить ссылки на новые страницы. И в этом смысле мы ближе к проблеме создания эффективного агрегатора RSS фидов, который мониторит RSS фиды с целью не пропустить новые посты [46, 47]. Читалка RSS, описанная в этих статьях, выясняет общие закономерности появления постов в каждом из RSS фидов с целью найти эффективный алгоритм планирования, который оптимизирует обработку RSS, чтобы быстро доставлять информацию до пользователей.

Эта RSS читалка использует краткое описание из RSS фида, когда предоставляет контент пользователям, и, таким образом, ей нет нужды в том, чтобы обходить сами страницы с новостями. Нам же, напротив, приходится скачивать полностью упоминаемые в RSS новости, чтобы их проиндексировать и сделать доступными в поисковой системе. Таким образом, хотя политики мониторинга RSS до некоторой степени похожи на проблему поиска новых эфемерных страниц на источниках контента, использовать их прямо “из коробки” для нашей проблемы нельзя. Проблема состоит в том, что мы вынуждены тратить заметное количество ресурсов на скачивание найденных страниц, и, если мы хотим делать это эффективно, то нам необходимо изменить расписание переобхода. Как бы то ни было, сами по себе RSS фиды можно использовать для нашей задачи в качестве источников контента.

**Политики обнаружения** Основная идея *политик обнаружения* состоит в том, чтобы сфокусировать поиск в ширину на высококачественном контенте, т.е. приоритизировать найденный, но еще не обойденный, контент (граница обхода) в соответствии с некоторой мерой качества. Некоторые подходы используют ссылочную структуру веба, например, в [33] вначале обходились страницы с самым большим числом входящих ссылок, тогда как страницы с большим PageRank имели больший приоритет в [2, 20]. В [27] эти подходы сравнивались с точки зрения влияния на эффективность поисковой системы. Однако в [42] обсуждается, что корреляция между мерами качества, основанными на ссылках, и реальным пользовательским интересом мала. В этой

связи они предлагают использовать поисковые логи, чтобы направить поисковый робот к страницам, которые потенциально более интересны пользователям. Мы, в свою очередь, следуя последним веяниям, также не полагаемся на одну ссылочную структуру, а используем клики пользователей из поисковых логов.

Наш поисковый робот находит и обходит новые страницы, но между нашим подходом и предыдущими есть принципиальная разница. Предыдущие подходы основаны на предположении, что веб достаточно глубок, и поэтому начиная с некоторого сида поисковый робот должен двигаться все глубже и глубже по возможности в направлении качественных новых страниц. Мы, напротив, утверждаем, что большинство новых страниц могут быть найдены непосредственно на небольшом множестве источников контента, однако ссылки с источников контента короткоживущие, поэтому эти источники надо часто переобходить с целью не пропустить ссылки. Это наблюдение с одной стороны упрощает задачу, но с другой стороны приводит к новому вызову — придумать, как найти правильный баланс между обходом новых страниц и переобходом источников контента.

**Единое упорядочивание обхода** Обычно статьи об обходе веба фокусируются либо на обнаружении новых страниц, либо на обновлении известных, но важность вопроса о том, как разделить ограниченные ресурсы между политиками обновления и обнаружения, обычно недооценивается. Проблема единого упорядочивания обхода, т.е. объединения разных политик в одну единую, была предложена в [40] в качестве важного будущего направления. Некоторые авторы предлагали использовать фиксированные квоты для каждой из политик [33, 44]. Однако, как это следует, например, из нашего анализа (см. параграф 3.3.1), такие фиксированные квоты могут быть далеки от оптимальных. Наш оптимизационный подход одновременно работает с политиками обхода и переобхода и поэтому может найти лучший способ разделения ресурсов. Мы, таким образом, делаем шаг в направлении решения проблемы единого упорядочивания обхода.

# Заключение

Первая глава этой работы посвящена моделям предпочтительного присоединения. В разделе 1.2 мы исследуем распределение подграфов в случайных графах в модификации LCD-модели  $G_m^n$ . Интересный результат состоит в том, что асимптотическое поведение с ростом графа математического ожидания числа копий фиксированного подграфа определяется лишь количеством вершин степени ноль, один и два в этом подграфе. Это означает, что LCD-модель не очень реалистично описывает реальные сети в этом аспекте. Одним из направлений будущей работы является обобщение полученных результатов на другие модели предпочтительного присоединения, например, модели Мори и Холм–Кима.

Далее в разделе 1.3 мы предлагаем новый подход к моделям предпочтительного присоединения, который позволяет получать результаты сразу для множества моделей и не повторять доказательства заново для каждой из них. На основе этого подхода предложена модель с качественно более реалистичным поведением глобального кластерного коэффициента, чем в предыдущих моделях предпочтительного присоединения. Однако, построение модели предпочтительного присоединения с асимптотически постоянным глобальным кластерным коэффициентом для показателя распределения степеней вершин меньше трех остается открытой проблемой. Мы также показываем, что в моделях предпочтительного присоединения поведение глобального кластерного коэффициента и среднего локального кластерного коэффициента разительно отличаются.

Во второй главе мы изучаем медиа-веб и показываем, что модели предпочтительного присоединения не подходят для объяснения его свойств, так как не отражают свойство устаревания медиа-страниц. Поэтому в разделе 2.3 мы предлагаем новый класс моделей эволюции сетей, в котором возможны разные функции привлекательности страниц, включая взятые из моделей предпочтительного присоединения и модели приспособления, но также и новые — с устареванием, отвечающие за особенности медиа-веба. В разделе 2.4 мы анализируем модели с устареванием теоретически и показываем, какие

из них реалистично предсказывают одновременно и распределение степеней вершин, и свойство устаревания медиа-веба. Наконец, в 2.5 мы сравниваем эти модели путем оценивания для каждой модели правдоподобия реальных данных при условии, что данные появились в соответствии с этой моделью. Один из самых удивительных выводов состоит в том, что в медиа-вебе вероятность процитировать страницу определяется скорее качеством страницы, чем ее текущей популярностью. Представляется, что полученная модель может найти множество практических применений. Например, имея ссылочных граф и предполагая, что он появился в соответствии с моделью, мы можем предсказывать качества и времена создания страниц, т.е. информацию, которая полезна в приложениях, но далеко не всегда известна. Еще одно приложение рассмотрено в третьей главе.

В третьей главе мы рассматриваем приложение моделей с устареванием для задачи обхода эфемерных страниц поисковым роботом. В разделе 3.1 мы формализуем задачу обхода эфемерных страниц путем введения подходящей метрики. Затем в разделе 3.2 мы проверяем, что большинство интересных пользователям эфемерных страниц могут быть на небольшом множестве источников контента. Наконец в разделе 3.3 мы предлагаем алгоритм, который динамически оценивает для каждого источника контента скорость появления новых ссылок (она зависит от времени суток и дня недели), а также их “качество” (учитывая текущие потребности пользователей). Затем в результате решения оптимизационной задачи алгоритм подбирает оптимальное расписание переобхода источников контента и скачивания новых страниц. Наш оптимизационный подход одновременно работает с политиками обхода и переобхода и поэтому может найти лучший способ разделения ресурсов. Мы, таким образом, делаем шаг в направлении решения проблемы единого упорядочивания обхода.

## Список литературы

- [1] <http://www.memetracker.org/data.html>.
- [2] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. WWW Conference, 2003.
- [3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74:47–97, 2002.
- [4] S. Bansal, S. Khandelwal, and L. A. Meyers. Exploring biological network structure with clustered random networks. *BMC Bioinformatics*, 10(405), 2009.
- [5] A.-L. Barabási and R. Albert. Emergence of scaling in random network. *Science*, 286(5439):509–512, 1999.
- [6] I. Bezáková, A. Kalai, and R. Santhanam. Graph model selection using maximum likelihood. pages 105–112. ICML Conference, 2006.
- [7] G. Bianconi and A.-L. Barabási. Bose–Einstein condensation in complex networks. *Physical Review Letters*, 86(24):5632–5635, 2001.
- [8] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308, 2006.
- [9] B. Bollobás. Mathematical results on scale-free random graphs. *Handbook of Graphs and Networks*, pages 1–34, 2003.
- [10] B. Bollobás, O. M. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, 18(3):279–290, 2001.
- [11] C. Borgs, J. Chayes, C. Daskalakis, and S. Roch. First to market is not everything: an analysis of preferential attachment with fitness. pages 135–144. ACM Symposium on Theory of Computing, 2007.

- [12] B. E. Brewington and G. Cybenko. How dynamic is the web? *Computer Networks*, 33(1):257–276, 2000.
- [13] B. E. Brewington and G. Cybenko. Keeping up with the changing web. *Computer Networks*, 33(5):52–58, 2000.
- [14] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(16):309–320, 2000.
- [15] P. G. Buckley and D. Osthus. Popularity based random graph models leading to a scale-free degree sequence. *Discrete Mathematics*, 282:53–63, 2004.
- [16] J. Cho and H. Garcia-Molina. Synchronizing a database to improve freshness. SIGMOD Conference, 2000.
- [17] J. Cho and H. Garcia-Molina. Effective page refresh policies for web crawlers. *ACM Transactions on Database Systems*, 28(4), 2003.
- [18] J. Cho and H. Garcia-Molina. Estimating frequency of change. *ACM TOIT Conference*, 3(3), 2003.
- [19] J. Cho and A. Ntoulas. Effective change detection using sampling. VLDB Conference, 2002.
- [20] J. Cho and U. Schonfeld. Rankmass crawler: a crawler with high personalized pagerank coverage guarantee. VLDB Conference, 2007.
- [21] C. Cooper. Distribution of vertex degree in web-graphs. *Combinatorics, Probability and Computing*, 15:637–661, 2006.
- [22] C. Cooper and A. Frieze. A general model of web graphs. *Random Structures and Algorithms*, 22(3):311–335, 2003.
- [23] C. Cooper and P. Prałat. Scale-free graphs of increasing degree. *Random Structures and Algorithms*, 38(4):396–421, 2011.
- [24] M. Deijfen, H. van den Esker, R. van der Hofstad, and G. Hooghiemstra. A preferential attachment model with random initial degrees. *Ark. Mat.*, 47:41–72, 2009.
- [25] N. Eggemann and S. D. Noble. The clustering coefficient of a scale-free random graph. *Discrete Applied Mathematics*, 159(10):953–965, 2011.

- [26] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. SIGCOMM Conference, 1999.
- [27] D. Fetterly, N. Craswell, and V. Vinay. The impact of crawl policy on web search effectiveness. In *SIGIR Conference*, pages 580–587, 2009.
- [28] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- [29] E. A. Grechnikov. An estimate for the number of edges between vertices of given degrees in random graphs in the Bollobás–Riordan model. *Moscow Journal of Combinatorics and Number Theory*, 1(2):40–73, 2011.
- [30] P. Holme and B. Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(2), 2002.
- [31] M. O. Jackson. Social and economic networks. Princeton University Press, 2010.
- [32] M. O. Jackson and A. Watts. The evolution of social and economic networks. *Journal of Economic Theory*, 106(2):265–295, 2002.
- [33] R. Kumar, K. Lang, C. Marlow, and A. Tomkins. Efficient discovery of authoritative resources. *Data Engineering*, 2008.
- [34] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. pages 497–506. ACM SIGKDD Conference, 2009.
- [35] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. pages 462–470, 2008.
- [36] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.
- [37] N. Matloff. Estimation of internet file-access/modification rates from indirect data. *ACM TransModel. Comput. Simul*, 15(3):233–253, 2005.
- [38] T. F. Móri. The maximum degree of the Barabási–Albert random tree. *Probability and Computing*, 14:339–348, 2005.



- [39] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.
- [40] C. Olston and M. Najork. Web crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010.
- [41] S. Pandey and C. Olston. *User-centric web crawling*. WWW Conference, 2005.
- [42] S. Pandey and C. Olston. Crawl ordering by search impact. WSDM Conference, 2008.
- [43] K. Radinsky and P. N. Bennett. Predicting content change on the web. WSDM Conference, 2013.
- [44] U. Schonfeld and N. Shivakumar. Sitemaps: above and beyond the crawl of duty. WWW Conference, 2009.
- [45] M. A. Serrano and M. Boguñá. Tuning clustering in random networks with arbitrary degree distributions. *Physical Review E*, 72(3):036133, 2005.
- [46] K. C. Sia and J. Cho. Efficient monitoring algorithm for fast news alert. *In IEEE Transaction on Knowledge and Data Engineering*, 19(7):950–961, 2007.
- [47] K. C. Sia, J. Cho, K. Hino, Y. Chi, S. Zhu, and B. L. Tseng. Monitoring rss feeds based on user browsing pattern. ICWSM Conference, 2007.
- [48] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.
- [49] E. Volz. Random networks with tunable degree distribution and clustering. *Phys. Rev. E*, 70(5), 2004.
- [50] T. Zhou, G. Yan, and B.-H. Wang. Maximal planar networks with large clustering coefficient and power-law degree distribution. *Phys. Rev. E*, 71(4), 2005.

## Публикации автора

- [51] А. Рябченко и Е. Самосват. О числе подграфов в случайном графе Барабаши-Альберт. *Доклады Академии наук*, том 435, стр. 587–590, 2010.

- [52] А. Рябченко и Е. Самосват. О числе подграфов в случайном графе Барабаши-Альберт. *Известия Российской академии наук. Серия математическая*, том 76, стр. 183–202, 2012.
- [53] L. Ostroumova, A. Ryabchenko, and E. Samosvat. Generalized preferential attachment: tunable power-law degree distribution and clustering coefficient. In *Algorithms and Models for the Web Graph*, pp. 185–202. LNCS, vol. 8305, 2013.
- [54] D. Lefortier, L. Ostroumova, and E. Samosvat. Evolution of the media web. In *Algorithms and Models for the Web Graph*, pp. 80–92. LNCS, vol. 8305, 2013.
- [55] 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 Conference on information & knowledge management*, pp. 745–750. ACM, 2013.

В совместных работах Самосвату Е. принадлежат основные результаты, соавторы помогали в редактировании текста, проведении экспериментов и доказательстве некоторых теорем.