На правах рукописи

САМОСВАТ Егор Александрович

МОДЕЛИРОВАНИЕ ИНТЕРНЕТА С ПОМОЩЬЮ СЛУЧАЙНЫХ ГРАФОВ

05.13.18 — математическое моделирование, численные методы и комплексы программ

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2014

Работа выполнена на кафедре дискретной математики Федерального государственного автономного образовательного учреждения высшего профессионального образования “Московский физико-технический институт (государственный университет)”.

Научный руководитель: доктор физико-математических наук, профессор Райгородский Андрей Михайлович. Место работы: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования “Московский государственный университет имени М.В. Ломоносова”.

Официальные оппоненты:

- доктор физико-математических наук, Кабатянский Григорий Анатольевич. Место работы: Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).

- кандидат физико-математических наук, доцент, Замараев Виктор Андреевич. Место работы: Национальный исследовательский университет «Высшая школа экономики».

Ведущая организация: Хабаровское отделение Федерального государственного бюджетного учреждения науки Института прикладной математики Дальневосточного отделения Российской академии наук.

Защита состоится 19 июня 2014 года в 14:30 на заседании диссертационного совета Д 002.017.04 при Федеральном государственном бюджетном учреждении науки «Вычислительный центр имени А.А. Дородницына Российской академии наук» по адресу 119991, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН и на сайте http://www.ccas.ru/.

Автореферат разослан ___ _____ 2014 года.

Ученый секретарь диссертационного совета
доктор физико-математических наук,
профессор Н. М. Новикова

Общая характеристика работы

Актуальность проблемы. 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. Это свидетельствует о широком интересе научного сообщества к данной проблематике и необходимости междисциплинарного подхода к изучению сложных сетей.

В последние 10-15 лет было предложено множество моделей сетей. 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 экономический сетей

[1,2,3,4,5,6,7]. Один из естественных способов состоит в том, чтобы рассмотреть сеть как результат некоторого случайного процесса, определенного в терминах простых естественных правил, которые гарантируют желаемые свойства, наблюдаемые в реальных сетях. Видимо, наиболее широко изученной реализацией этого подхода является предпочтительное присоединение.

Механизм предпочтительного присоединения (preferential attachment) был положен в основу модели развития Интернета в 1999 году Барабаши и Альберт [8]. 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, например, в LCD [9] и RAN [10] модели. К актуальным задачам анализа моделей предпочтительного присоединения можно отнести: иссле-

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74:47–97, 2002.

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

[3] 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.

[4] M. O. Jackson. *Social and economic networks*. Princeton University Press, 2010.

[5] M. O. Jackson and A. Watts. The evolution of social and economic networks. *Journal of Economic Theory*, 106(2):265–295, 2002.

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

[7] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

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

[9] B. Bollobás. Mathematical results on scale-free random graphs. *Handbook of Graphs and Networks*, pages 1–34, 2003.

[10] T. Zhou, G. Yan, and B.-H. Wang. Maximal planar networks with large clustering

дование поведения различных характеристик в моделях и их сравнение с наблюдаемыми в реальных сетях, обобщение результатов, полученных для разных моделей, выяснение границ применимости этих моделей.

Ясно, что для некоторых частей Интернета модели предпочтительного присоединения принципиального не подходят. Например, они плохо описывают эволюцию медиа-веба, т.е. высокодинамической части веба, где ежедневно появляется множество новых страниц, связанных с медиаконтентом: новостями, постами в блогах и форумах. Действительно, в новостях и блогах редко цитируют сюжеты, потерявшие свою актуальность, какими бы популярными они ни были до этого. Создание моделей для высокодинамических частей Интернета является интересной проблемой.

Другой областью, где анализ сложных сетей находит применение, является информационный поиск. Например, актуальной задачей в информационном поиске является эффективный обход Интернета поисковым роботом. Поисковый робот традиционно выполняет две задачи: обнаружение неизвестных качественных страниц и обновление обнаруженных ранее. Обе эти проблемы активно исследовались в течение последнего десятилетия [11]. 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. Эта задача особенно актуальна для страниц, к которым интерес пользователей быстро пропадает (эфемерных страниц). Создание адекватных моделей эволюции Интернета может помочь в решении этой задачи.

Цель работы:

1. Исследование распределения числа подграфов, изоморфных фиксированному графу, в моделях предпочтительного присоединения.

2. Обобщение результатов о распределении степеней вершин и поведении кластерных коэффициентов в моделях предпочтительного при-

---

coefficient and power-law degree distribution. *Phys. Rev. E*, 71(4), 2005.

[11] C. Olston and M. Najork. Web crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010.

соединения.

3. Построение адекватной модели эволюции медиа-веба и ее валидация.

4. Разработка алгоритма эффективного обхода эфемерных страниц поисковым роботом.

Научная новизна. В диссертации получены следующие результаты, характеризующиеся научной новизной:

1. Для модифицированной LCD-модели доказаны теоремы, позволившие оценить распределение числа подграфов, изоморфных фиксированному графу, в этой модели.

2. Предложен подход к обобщению результатов для моделей предпочтительного присоединения, а именно, введен класс моделей, включающий в себя множество моделей этого типа. Для всего класса получены результаты о распределении степеней вершин и поведении кластерных коэффициентов. На основе подхода предложена модель, обладающая более реалистичным поведением глобального кластерного коэффициента и рядом других интересных свойств.

3. Построена адекватная модель эволюции медиа-веба, проведена ее теоретическая и эмпирическая (с помощью метода максимального правдоподобия) валидация.

4. На основе модели эволюции медиа-веба разработан алгоритм эффективного обхода эфемерных страниц поисковым роботом.

Теоретическая и практическая ценность работы. Теоретическая значимость результатов исследования состоит в анализе распределения числа копий фиксированного подграфа в моделях предпочтительного присоединения, в обобщении результатов о распределении степеней вершин и поведении кластерных коэффициентов для моделей предпочтительного присоединения, в построении модели предпочтительного присоединения, обладающей более реалистичным поведением глобального кластерного коэффициента.

Практическая значимость работы заключается в анализе свойств медиа-веба, в построении адекватной модели медиа-веба и ее валидации, в разработке на основе модели эволюции медиа-веба алгоритма эффективного обхода эфемерных страниц поисковым роботом.

Публикации и апробация работы. По материалам диссертационной работы опубликовано 5 печатных работ из списка, рекомендованного ВАК.

Результаты настоящего исследования были представлены на следующих научных конференциях:

- 8th French Combinatorial Conference, г. Париж, июнь 2010 года. Тема доклада: “On estimating the expected number of sub-graphs in Barabási and Albert model of random graph”.

- 53-я научная конференция МФТИ, г. Долгопрудный, ноябрь 2010 года. Тема доклада: “О числе подграфов в случайном графе Барабаши–Альберт”.

- 54-я научная конференция МФТИ, г. Долгопрудный, ноябрь 2011 года. Тема доклада: “О новой модели случайного веб-графа”.

- 55-я научная конференция МФТИ, г. Долгопрудный, ноябрь 2012 года. Тема доклада: “Модели предпочтительного присоединения”.

- Workshop on Random Graphs and their Applications, г. Москва, октябрь 2013 года. Тема доклада: “Evolution of the Media Web”.

- 22nd ACM International Conference on Information and Knowledge Management, г. Сан-Франциско, октябрь 2013 года. Тема доклада: “Timely crawling of high-quality ephemeral new content”.

- 56-я научная конференция МФТИ, г. Долгопрудный, ноябрь 2013 года. Тема доклада: “Об эволюции медиа-веба”.

- 10th Workshop on Algorithms and Models for the Web Graph, Harvard University, г. Кембридж, декабрь 2013 года. Темы докладов: “Evolution of the Media Web” и “Generalized preferential attachment: tunable power-law degree distribution and clustering coefficient”.

- Научные семинары Вычислительного центра имени А. А. Дородницына Российской академии наук, Московского физико-технического института, Московского государственного университета имени М. В. Ломоносова.

Структура диссертации и объем работы. Диссертационная работа состоит из введения, трех глав, заключения и списка использованной литературы. Основная часть работы изложена на 98 страницах машинописного текста. Список использованной литературы включает 55 наименований.

Результаты работы и их обсуждение

Во введении обосновывается актуальность темы диссертационной работы, формулируется цель и научная новизна полученных результатов, дается краткое содержание работы по главам.

В первой главе исследуются модели предпочтительного присоединения. Эта глава соответствует математическому подходу к анализу сложных сетей и основана на статьях [1, 2, 3].

В разделе 1.1 дается краткий обзор результатов касательно моделей предпочтительного присоединения, определяется LCD-модель $G_m^{(n)}$, а также следующая ее модификация G_m^n.

Рассмотрим последовательность вершин $\{1, 2, \ldots\}$. Мы индуктивно определим процесс $(G_m^t)_{t\geq 1}$ так, что G_m^t является графом на множестве вершин $\{1, \ldots, t\}$ 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 $i \in \{1, \ldots, t\}$ выбирается случайно, причем

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

Здесь, 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. Мы используем это отличие для указания того, о какой модели идет речь. Главное отличие этой 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 [9].

В разделе 1.2 для модели G_m^n сначала доказывается новая теорема о распределении числа треугольников, которая затем обобщается на случай произвольного подграфа. В параграфе 1.2.1 доказывается теорема о распределении числа треугольников.

---

[9] B. Bollobás. Mathematical results on scale-free random graphs. *Handbook of Graphs and Networks*, pages 1–34, 2003.

Обозначим через $\#\left(G_0, G_m^n\right)$ число подграфов, изоморфных графу G_0, в графе G_m^n.

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

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

В параграфе 1.2.2 техника, применяемая при подсчете математического ожидания числа треугольников, обобщается, а именно доказываются 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 ξ_k – это произвольная функция от G_m^k, в частности, – любой многочлен от произвольного числа величин d_i^l, e_{pq}, где $i \le l \le k$, $p \le q \le k$.

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

$$\mathsf{E}\left(\xi_{n-1} \cdot \left(d_{\alpha_1}^n\right)^{a_1} \cdot \ldots \cdot \left(d_{\alpha_k}^n\right)^{a_k} \cdot e_{\beta_1 n} \cdot \ldots \cdot e_{\beta_r n}\right) =$$
$$= \mathsf{E}\left(\xi_{n-1} \cdot \left(d_{\alpha_1}^{n-1}\right)^{a_1} \cdot \ldots \cdot \left(d_{\alpha_k}^{n-1}\right)^{a_k} \cdot d_{\beta_1}^{n-1} \cdot \ldots \cdot d_{\beta_r}^{n-1}\right) \left(\frac{1}{n}\right)^r \cdot \Theta(1)$$

при $r \le m$ и $\beta_i \ne \beta_j$. 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*

$$\mathsf{E}\left(\xi_l \cdot \left(d_{\alpha_1}^n\right)^{a_1} \cdot \ldots \cdot \left(d_{\alpha_k}^n\right)^{a_k}\right) =$$
$$= \mathsf{E}\left(\xi_l \cdot \left(d_{\alpha_1}^l\right)^{a_1} \cdot \ldots \cdot \left(d_{\alpha_k}^l\right)^{a_k}\right) \left(\frac{n}{l}\right)^{\frac{\sum a_i}{2}} \cdot \Theta(1).$$

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.

Из теорем о длинном и коротком спуске вытекает следующая теорема о произвольном подграфе.

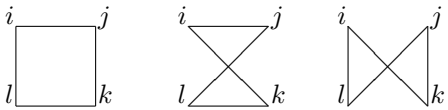
Теорема 9 (о произвольном подграфе). *Пусть задан граф G_0, степени вершин которого равны $d_1, \ldots, d_s$. Обозначим через $\#(d_i = k)$ число вершин в G_0, степень каждой из которых равна k. Тогда*

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

Действительно, заметим, что количество копий фиксированного подграфа может быть записано с помощью величины e_{pq}. Рассмотрим, например, случай $G_0 = K_{2,2}$ (полный двудольный граф с долями размеров два и два), который мало отличается от случая произвольного G_0.

В этом случае

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



Далее требуется аккуратное применение теорем о длинном и коротком спуске. Доказательства теорем 7, 8, 9 приведены в параграфах 1.2.3–1.2.5.

В разделе 1.3 предложен подход к обобщению результатов для моделей предпочтительного присоединения, а именно, введен класс моделей, включающий в себя множество моделей этого типа.

В параграфе 1.3.1 вводится PA-класс моделей (PA от preferential attachment). Пусть G_m^n $(n \ge n_0)$ — это граф с n вершинами $\{1, \ldots, n\}$ 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 $(n \ge n_0)$ 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 $\{1, \ldots, n, n+1\}$. Обозначим через d_v^n степень вершины v в графе G_m^n. Если для некоторых констант A и B выполняются следующие условия

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

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

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

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

то мы говорим, что случайный граф G_m^n — это модель из PA-класса. Условие (4) 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.

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 (1–4). Например, LCD-модель, модель Холм–Кима [12] и RAN-модель принадлежат PA-классу с параметрами $A = 1/2$ и $B = 0$. Модели Бакли–Остуса [13] и Мори [14] также принадлежат PA-классу с параметрами $A = \frac{1}{2+\beta}$ и $B = \frac{m\beta}{2+\beta}$. Заметим, что наш класс моделей шире класса Барабаши–Албert, 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 получаем результаты об общих свойствах всего PA-класса, заданного в терминах соотношений на вероятности событий, оставляя тем самым большую свободу в точном определении модели.

---

[12] P. Holme and B. Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(2), 2002.

[13] P. G. Buckley and D. Osthus. Popularity based random graph models leading to a scale-free degree sequence. *Discrete Mathematics*, 282:53–63, 2004.

[14] T. F. Móri. The maximum degree of the barabási-albert random tree, combinatorics. *Probability and Computing*, 14:339–348, 2005.

Для всего класса получены результаты о распределении степеней вершин и поведении кластерных коэффициентов. Мы используем следующие обозначения. Под **whp** ("with high probability") мы имеем в виду, что для некоторой последовательности событий A_n выполнено $\mathsf{P}(A_n) \to 1$, когда $n \to \infty$. Мы говорим, что $a_n \sim b_n$, если $a_n = (1 + o(1))b_n$, и $a_n \approx b_n$, если $C_0 b_n \le a_n \le C_1 b_n$ для некоторых констант $C_0, C_1 > 0$. Мы говорим, что **whp** $a_n \sim b_n$, если $\exists \phi : \phi = o(1)$ и $\mathsf{P}(a_n = (1 + \phi(n))b_n) \to 1$, когда $n \to \infty$.

В параграфе 1.3.2 сначала оценивается $N_n(d)$ — количество вершин степени d в графе G_m^n, а именно доказывается следующий результат о математическом ожидании $\mathsf{E}N_n(d)$ случайной величины $N_n(d)$.

Теорема 10. *Пусть $d \ge m$. Тогда*

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

где

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

а $\Gamma(x)$ — это гамма-функция.

Затем доказывается, что количество вершин степени d плотно сконцентрировано около своего математического ожидания.

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

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

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

$$|N_n(d) - \mathsf{E}N_n(d)| \le \varphi(n)\,\mathsf{E}N_n(d)\ .$$

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

В параграфе 1.3.3 исследуется поведение кластерного коэффициента в моделях 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.

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

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

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.

В некоторых предположениях 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 (5). Тогда*

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

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

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

Мы приводим не строго математическое доказательство теоремы 14. Однако используемые нами предположения точнее метода среднего поля, часто применяемого в физических статьях. Теорема 14 показывает, что в некоторых случаях ($2A \ge 1$) глобальный кластерный коэффициент $C_1(n)$ стремится к нулю с ростом размера графа, даже если выполнено условие (5). 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**

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

В параграфе 1.3.4 предложена *полиномиальная модель*, обладающая более реалистичным поведением глобального кластерного коэффициента и рядом других интересных свойств. Как обычно, 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 $n+1$ в вершины *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 (1–4) 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

$$\mathsf{P}\,(\text{ребра } e_1,\dots,e_m \text{ идут в вершины } i_1,\dots,i_m, \text{ соответственно}) = \\ = \sum_{k=0}^{m/2} \sum_{l=0}^{m-2k} \alpha_{k,l} M_{k,l}^{n,m}(i_1,\dots,i_m)\,. \quad (6)$$

Приведенная таблица резюмирует сравнение полиномиальной модели с остальными упомянутыми моделями предпочтительного присоединения:

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

В параграфах 1.3.5 и 1.3.6 наши теоретические результаты демонстрируются с помощью компьютерного моделирования. В параграфе 1.3.7 обсуждаются полученные результаты, а параграф 1.3.8 посвящен доказательству теорем.

Во второй главе построена адекватная модель эволюции медиа-веба, проведена ее теоретическая и эмпирическая (с помощью метода максимального правдоподобия) валидация. Эта глава соответствует физическому подходу к анализу сложных сетей и основана на статье [4].

В разделе 2.1 мы описываем базовые модели, от которых мы отталкивались при построении модели медиа-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. Обозначим через $e(T)$ долю ребер, соединяющих страницы с разницей возрастов больше T. Мы проанализировали

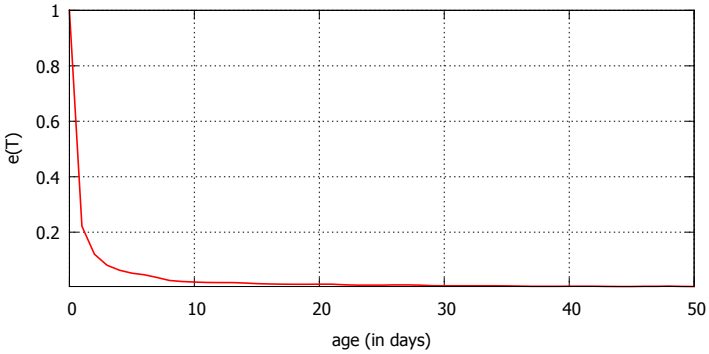


Рис. 1: Свойство устаревания

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. 1) и заметили, что $e(T)$ убывает экспоненциально быстро, что не соответствует моделям предпочтительного присоединения. Это наблюдение натолкнуло нас на мысль модифицировать предпочтительное присоединение, понижая вероятность ссылки на неактуальные страницы.

В разделе 2.3 мы предлагаем новый класс моделей эволюции сетей, которые учитывают свойство устаревания страниц. Модели, обладающие таким свойством, мы называем моделями *с устареванием*.

Предположим, что мы имеем фиксированный набор хостов $H_1, \ldots, H_n$ и каждый хост, т.е. множество страниц, H_i характеризуется скоростью появления на нем новых страниц λ_i. Мы стартуем с пустого графа, т.е. в начале имеется n пустых множеств $H_1, \ldots, H_n$. Далее мы предполагаем, что новые страницы на хосте H_i появляются согласно пуассоновскому процессу с параметром λ_i. 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 ρ_{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:

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

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

Например, $f(d,q,a) = d$ (т.е. $\overrightarrow{\alpha} = (0,1,0)$) приводит к предпочтительному присоединению, тогда как $f(d,q,a) = q \cdot d$ (т.е. $\overrightarrow{\alpha} = (1,1,0)$) приводит к модели приспособления (fitness model) [15]. Мы изучаем разные варианты и показываем, какие из них лучшим образом отражают поведение медиа-веба в разделах 2.4 и 2.5.

В разделе 2.4 в приближении среднего поля мы доказываем теоремы о распределении степеней вершин для моделей с устареванием.

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

Здесь W_k и N_k — это некоторые константы, объяснение смысла которых можно найти в диссертации. Из теоремы 15 следует, что в первом случае, чтобы иметь степенной закон распределения случайной величины d_p, качество q_p должно быть распределено экспоненциально, при этом контролировать параметр степенного закона оказывается очень трудно. Во

[15] G. Bianconi and A.-L. Barabási. Bose–Einstein condensation in complex networks. *Physical Review Letters*, 86(24):5632–5635, 2001.

Таблица 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	**-5.17**	-5.45

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.

Также в этом разделе в приближении среднего поля мы доказываем теорему о свойстве устаревания для моделей с устареванием.

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.

В разделе 2.5 с помощью метода максимального правдоподобия проведена эмпирическая валидация предложенных моделей. А именно, 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, затем мы нормализуем сумму на количество ребер. Результаты приведены в Таблице 1.

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.

Таблица 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	**0.30**	0.16

Таблица 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	-

Согласно разным моделям ребра имеют разные вероятности, та из моделей, которая приписывает ребру наибольшую вероятность, называется *побеждающей* на этом ребре (см. Таблицу 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 3). Из обеих таблиц ясно видно, что фактор устаревания играет очень важную роль.

Мы видим, что наиболее вероятная модель имеет функцию привлекательности $f = qe^{\frac{-a}{\tau}}$. Это означает, что в медиа-вебе вероятность процитировать страницу определяется скорее качеством страницы, чем ее текущей популярностью.

Третья глава посвящена разработке алгоритмов эффективного обхода эфемерных страниц поисковым роботом. В этой главе существенно используются результаты главы 2 и она основана на статье [5].

Как уже обсуждалось, медиа-страницы интересны пользователям лишь несколько дней после своего появления. Чтобы подчеркнуть эту особен-

ность, 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 моментом 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.

В разделе 3.1 мы формализуем проблему обхода эфемерных страниц поисковым роботом посредством введения подходящей метрики.

Предположим, что для каждой страницы i мы знаем убывающую функцию $P_i(\Delta t)$, которая есть "польза" от ее скачивания с задержкой Δt секунд после ее появления t_i (под пользой можно иметь в виду количество "кликов" или "показов", которые страница соберет в поисковой выдаче). Если в итоге каждая страница i была скачана с задержкой Δt_i, мы можем определить *динамическое качество* поискового робота:

$$Q_T(t) = \frac{1}{T} \sum_{i:t_i+\Delta t_i \in [t-T,t]} P_i(\Delta t_i). \tag{7}$$

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:

$$Q = \lim_{T \to \infty} Q_T(t) \tag{8}$$

которое не зависит от t.

Под пользой $P_i(\Delta t)$ от скачивания страницы i в момент времени $t_i + \Delta t$ мы имеем 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.

В разделе 3.2 описаны результаты экспериментов на реальных данных Яндекса, которые подтвердили по гипотезу о том, что большинство нового эфемерного контента может быть найдено на небольшом количестве источников, и предложен метод их нахождения.

В разделе 3.3 мы находим оптимальное расписание "переобхода" (т.е. повторного обхода) источников контента, которое позволит быстро находить появляющиеся новые качественные страницы, и понимаем, как распределить ресурсы между обходом новых страниц и переобходом источников контента.

Для поиска оптимального расписания обхода источников контента и скачивания новых страниц мы используем следующую аппроксимацию $P_i(\Delta t)$ (т.е. количества будущих кликов):

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

где *скорость устаревания* μ_i и *польза* P_i зависят от источника контента и могут быть оценены из исторических данных.

Заметим, что используемая нами аппроксимация функции полезности фактически является лучшим вариантом функции привлекательности, который мы нашли во второй главе, а именно $f(d, q, a) = q \cdot e^{-\frac{a}{\tau}}$.

Предположим, что нам дано множество источников контента $S_1, \ldots, S_n$. Пусть λ_i — это *скорость появления новых ссылок* на источнике S_i, т.е. среднее количество ссылок на новые страницы, появляющихся в секунду.

Рассмотрим алгоритм, который обходит источник S_i каждые I_i секунд, находит ссылки на новые страницы, а затем скачивает все эти новые страницы. Мы хотим найти такое расписание обхода источников, которое максимизировало бы среднее качество Q, 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 обхода:

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

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:

$$\sum_i \frac{1}{I_i} = N - \sum_i \lambda_i. \tag{9}$$

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

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

Мы заменяем $P_j(\Delta t_j)$ приближением $P_i e^{-\mu_i \Delta t_j}$ и получаем:

$$Q = \sum_i \frac{P_i}{I_i} \sum_{j=0}^{\lambda_i I_i - 1} e^{-\mu_i \frac{j}{\lambda_i}} =$$

$$= \sum_i \frac{P_i}{I_i} \frac{1 - e^{-\mu_i I_i}}{1 - e^{-\frac{\mu_i}{\lambda_i}}} = \sum_i p_i x_i \left(1 - e^{-\mu_i / x_i}\right),$$

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 $p_1 \le \ldots \le p_n$. Теперь для максимизации $Q(x_1, \ldots, x_n)$ при условии (9) мы используем метод множителей Лагранжа:

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

где w — это множитель Лагранжа.

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

$$\begin{cases} p_i \left(1 - (1 + \mu_i I_i) e^{-\mu_i I_i}\right) = \omega, \quad i = 1, \ldots, n, \\ \sum_i \frac{1}{I_i} = N - \sum_i \lambda_i. \end{cases} \tag{10}$$

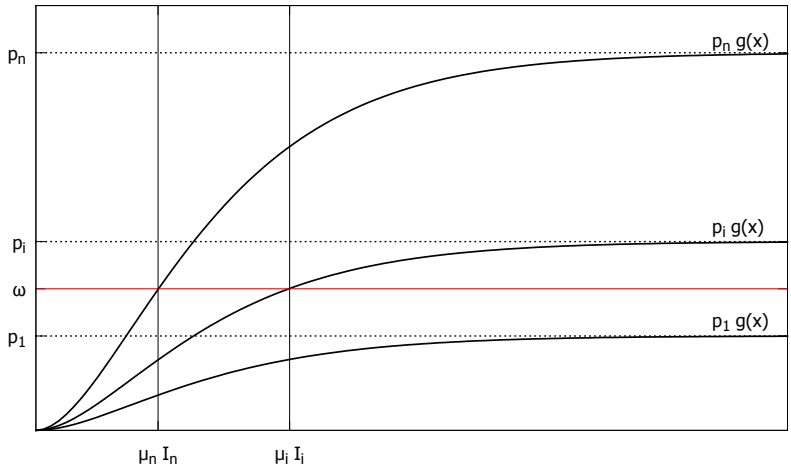


Рис. 2: Оптимизация I_i

Функция $g(x) = (1 - (1 + x)e^{-x})$ монотонно возрастает для $x > 0$, причем $g(0) = 0$ и $g(+\infty) = 1$. Следовательно, для любого ω $(0 < \omega < p_i)$ существует единственное $\mu_i I_i = g^{-1}(\frac{\omega}{p_i})$, как показано на Рис. 2. Так как, $\mu_i I_i$ монотонно возрастающая функция ω, то $\sum_i \frac{1}{I_i}$ является монотонной функцией ω, и мы используем бинарный поиск для того, что удовлетворить условию $\sum_i \frac{1}{I_i} = N - \sum_i \lambda_i$.

Затем мы описываем конкретный практический алгоритм ECHO (от Ephemeral Content Holistic Ordering), основанный на нашем теоретическом анализе.

В разделе 3.4 мы экспериментально сравниваем качество работы ECHO алгоритма с некоторыми другими подходами для трех разных скоростей обхода N. В Таблице 4 показаны полученные средние значения общего качества и их стандартные отклонения, а на Рис. 3 показано динамическое качество для окна размером в 5 часов и $N = 0.1$.

ECHO-newpages (одна из модификаций ECHO алгоритма) демонстрирует наилучшие результаты, которые при этом близки к оценке сверху, хотя скорость обхода гораздо меньше скорости появления новых ссылок. Это значит, что наш алгоритм эффективно расходует ресурсы и вначале

Таблица 4: 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
ECHO-greedy	0.60±0.03	0.68±0.03	0.69±0.03
ECHO-schedule	0.52±0.02	**0.69**±0.03	**0.71**±0.03
ECHO-newpages	**0.62**±0.04	**0.69**±0.03	**0.71**±0.03
Оценка сверху	*0.72*	*0.72*	*0.72*

обходит наиболее качественные страницы и источники.

В разделе 3.5 мы обсуждаем полученные в этой главе результаты и их связь с предыдущими исследованиями других авторов.

В заключении сформулированы основные результаты работы, полученные в диссертации.

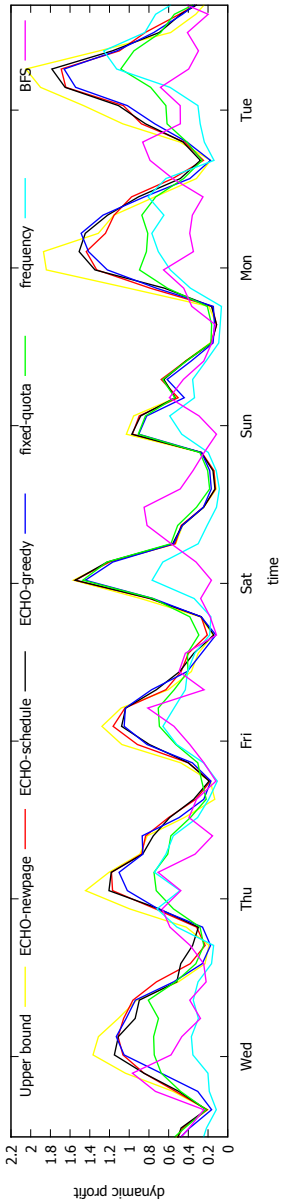


Рис. 3: Динамическое качество для временного окна в 5 часов.

Результаты, выносимые на защиту:

1. Для модифицированной LCD-модели доказаны теоремы, позволившие оценить распределение числа подграфов, изоморфных фиксированному графу.

2. Введен новый класс моделей, PA-модели, обобщающий известные ранее разрозненные модели. Для PA-класса получены как принципиально новые, так и обобщающие известные ранее результаты о распределении степеней вершин и поведении кластерных коэффициентов.

3. Введены и исследованы модели с устареванием. Для этих моделей в приближении среднего поля доказаны теоремы о распределении степеней вершин и свойстве устаревания. Среди моделей с устареванием выбрана модель, наилучшим образом описывающая эволюцию медиа-веба. Показано, что в медиа-вебе вероятность процитировать страницу определяется скорее качеством страницы, чем ее текущей популярностью.

4. Посредством введения подходящей метрики формализована проблема обхода эфемерных страниц поисковым роботом. Разработан алгоритм эффективного обхода эфемерных страниц поисковым роботом. Качество алгоритма экспериментально проанализировано на реальных данных. Алгоритм внедрен в компании Яндекс.

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.

Основные результаты диссертации изложены в следующих публикациях:

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.