4. Модели случайных веб-графов

В этой главе мы поговорим о самых современных моделях случайных графов, которые призваны описывать рост различных сетей — социальных, биологических, транспортных. Но в первую очередь речь пойдет об Интернете. В 90-е годы XX века, когда Интернет только зарождался, исследователи уже задались вопросом, каким законам подчиняется рост Интернета и какова наиболее адекватная модель для описания свойств этой сети. Одними из первых здесь были А.-Л. Барабаши и Р. Альберт. Они нашли ряд важных эмпирических закономерностей в поведении Интернета и на их основе придумали модель, которую впоследствии по-разному формализовывали многие авторы. Мы построим наше изложение следующим образом. В первом параграфе мы расскажем о результатах Барабаши — Альберт. Во втором параграфе мы опишем модель Б. Боллобаша и О. Риордана, которая весьма неплохо ложится на статистики Барабаши — Альберт. В третьем параграфе мы обсудим возможные уточнения модели Боллобаша — Риордана.

4.1. Наблюдения Барабаши — Альберт

В своих работах [66–68] Барабаши и Альберт, а также Х. Джеонг описали те статистики Интернета, которые легли в основу науки о росте этой сети — науки, имеющей глубокие приложения как в собственно интернетовской проблематике, так и в многочисленных близких дисциплинах. В действительности, большинство реальных сетей (социальные, биологические, транспортные и пр.) имеют похожую «топологию».

Итак, сперва договоримся о том, что мы понимаем под сетью Интернет. Это так называемый *веб-граф*, вершины которого суть какие-либо конкретные структурные единицы в Интернете: речь может идти о страницах, сайтах, хостах, владельцах и пр. Для определенности будем считать, что вершинами веб-графа служат именно сайты. Ребрами же мы будем соединять те вершины, между которыми имеются ссылки. При этом разумно проводить столько ребер между двумя вершинами, сколько есть ссылок между соответствующими сайтами. Более того, ребра естественно считать направленными. Таким образом, веб-граф ориентирован и он может иметь кратные ребра, петли и даже кратные петли (ссылки вполне могут идти с одной страницы

данного сайта на другую его страницу). Это такой «псевдомультиорграф». Сразу понятно, что для подобного «зверя» модель Эрдёша — Реньи вряд ли подходит.

Теперь мы готовы перечислить самые основные моменты исследования Барабаши — Альберт. По существу, этих моментов всего три. Во-первых, веб-граф — это весьма разреженный граф. У него на t вершинах примерно kt ребер, где $k \geqslant 1$ — некоторая константа. Для сравнения, у полного графа на t вершинах $C_t^2 = \Theta(t^2)$ ребер (см. значок «Θ» в приложении). Однако — и это во-вторых, — диаметр веб-графа исключительно скромен. (Напомним, что *расстояние* между двумя вершинами графа — это количество ребер в кратчайшем реберном пути между ними, а *диаметр* графа — это максимум попарных расстояний между его вершинами.) В 1999 году диаметр имел величину (см. [68]) 5—7. Это хорошо всем известное свойство любой социальной сети, которое принято в обыденной речи характеризовать выражением «мир тесен». Например, говорят о том, что любые два человека в мире «знакомы через 5—6 рукопожатий». Точно так же и сайты: «кликая» по ссылкам, можно с любого сайта на любой другой перейти за 5—7 нажатий клавиши компьютерной мыши. Конечно, тут есть важная оговорка. Некоторые едва появившиеся сайты могут не быть связаны с внешним по отношению к ним миром. Несколько правильнее сказать, что в веб-графе есть гигантская компонента, и уже ее диаметр невелик. Таким образом, веб-граф очень специфичен: будучи разреженным, он, тем не менее, в известном смысле тесен.

В-третьих, у веб-графа весьма характерное распределение степеней вершин. Эмпирическая вероятность того, что вершина веб-графа имеет степень d (т. е. просто доля вершин степени d), оценивается как c/d^{λ}, где $\lambda \approx 2{,}1$, а c — нормирующий множитель, вычисляемый из условия «сумма вероятностей равна 1». Этот любопытный факт роднит Интернет с очень многими реальными сетями — биологическими, социальными, транспортными. Все они подчиняются «степенному закону» (т. е. закону, в котором вероятности имеют вид c/d^{λ}), только у каждой из них свой показатель λ.

Ввиду перечисленных наблюдений не остается никаких сомнений в том, что модель Эрдёша — Реньи не применима для описания роста Интернета и подобных сетей. Если подбором вероятности p еще можно добиться разреженности и тесноты (хотя и не с теми параметрами), то степенной закон совсем уж не имеет отношения к схеме Бернулли, в рамках которой появляются ребра обычного случайного графа. В модели $G(n, p)$ степень каждой вершины случайного графа биномиальна с параметрами $n - 1$ и p (см. п. 2.11.4), и при тех p, кото-

рые мало-мальски гарантируют разреженность (т. е. при $p = \Theta(1/n)$), указанное биномиальное распределение аппроксимируется пуассоновским, а вовсе не степенным.

Сами Барабаши и Альберт предложили очень разумный взгляд на процесс формирования Интернета. Давайте считать, сказали они, что в каждый момент времени появляется новый сайт, и этот сайт ставит фиксированное количество ссылок на своих предшественников. На кого он предпочтет сослаться? Наверное, на тех, кто и так уже популярен. Можно допустить, что вероятность, с которой новый сайт поставит ссылку на один из прежних сайтов, пропорциональна числу уже имевшихся на тот сайт ссылок.

Модели случайных графов, основанные на описанной идее, называются моделями *предпочтительного присоединения*. В своих работах Барабаши и Альберт никак не конкретизировали, какую именно из этих моделей они предлагают рассматривать. А эти модели исключительно разнородны по своим свойствам. Ведь можно ставить ссылки независимо друг от друга, а можно еще и зависимости между разными ссылками с одного сайта учитывать. В итоге удается доказать даже такой забавный факт.

Теорема 42. *Пусть $f(n)$, $n \geqslant 2$, — произвольная целочисленная функция, такая что $f(2) = 0$, $f(n) \leqslant f(n+1) \leqslant f(n) + 1$ для всех $n \geqslant 2$ и $f(n) \to \infty$ при $n \to \infty$. Тогда существует такая модель типа Барабаши — Альберт, что в ней с вероятностью, стремящейся к единице при $n \to \infty$, случайный граф содержит в точности $f(n)$ треугольников.*

Одну из наиболее правильных спецификаций модели Барабаши — Альберт предложили в начале двухтысячных годов Б. Боллобаш и О. Риордан. В следующем параграфе мы ее обсудим.

4.2. Модель Боллобаша — Риордана

Наиболее полно эта модель описана в книге [1] и обзоре [69]. Также имеется малодоступная книга [70]. Мы представим здесь две основных и, по сути, совпадающих модификации этой модели. В одной дается динамическое, а в другой статическое описание случайности. Интуитивно более понятна динамическая модификация, с нее и начнем.

4.2.1. Динамическая модификация

Сперва построим последовательность (случайных) графов $\{G_1^n\}$, в которой у графа с номером n число вершин и ребер равно n. Затем

сделаем из нее последовательность $\{G_k^n\}$, в которой у графа с номером n число вершин равно n, а число ребер равно kn, $k \in \mathbb{N}$.

Итак, пусть $G_1^1 = (\{1\}, \{(1,1)\})$, т. е. в начальный момент времени есть одна вершина и одна петля. Пусть теперь граф G_1^{n-1} уже построен. У него вершины образуют множество $\{1, \ldots, n-1\}$, а ребер у него тоже $n-1$ штука. Добавим вершину n и ребро (n, i), у которого $i \in \{1, \ldots, n\}$. Ребро (n, n) будет появляться с вероятностью $\dfrac{1}{2n-1}$; ребро (n, i) возникнет с вероятностью $\dfrac{\deg i}{2n-1}$, где $\deg i$ — степень вершины i в графе G_1^{n-1}. Очевидно, что распределение вероятностей задано корректно, поскольку

$$\sum_{i=1}^{n-1} \frac{\deg i}{2n-1} + \frac{1}{2n-1} = \frac{2n-2}{2n-1} + \frac{1}{2n-1} = 1.$$

Случайный граф G_1^n построен, и он удовлетворяет принципу предпочтительного присоединения.

Осталось перейти к G_k^n. Берем G_1^{kn}. Это граф с kn вершинами и kn ребрами. Делим множество его вершин на последовательные куски размера k:

$$\{1, \ldots, k\}, \quad \{k+1, \ldots, 2k\}, \quad \ldots, \quad \{k(n-1)+1, \ldots, kn\}.$$

Объявляем каждый кусок «вершиной», а ребра сохраняем, т. е. если были ребра внутри куска, то будут кратные петли, а если были ребра между двумя различными кусками — будут кратные ребра. Внешне — вполне себе Интернет, как мы его и представляли. Вершин стало n, а ребер — по-прежнему kn. Цель реализована.

4.2.2. Статическая модификация, или LCD-модель

Введем такой объект, который называется *линейной хордовой диаграммой*. Вообще-то, он возник в топологии и теории узлов (см., например, [71]), но его комбинаторика оказывается напрямую связана с формированием веб-графа.

Итак, зафиксируем на оси абсцисс на плоскости $2n$ точек: 1, 2, 3, … …, $2n$. Разобьем эти точки на пары, и элементы каждой пары соединим дугой, лежащей в верхней полуплоскости. Полученный объект назовем *линейной хордовой диаграммой* (*linearized chord diagram* или, короче, LCD по-английски). Дуги в нем могут пересекаться, лежать друг под дружкой, но не могут иметь общих вершин. Количество различных LCD легко считается. Оно равно

$$l_n = \frac{(2n)!}{2^n n!}.$$

По каждой LCD построим граф с n вершинами и n ребрами. Действуем так. Идем слева направо по оси абсцисс, пока не встретим впервые правый конец какой-либо дуги. Пусть этот конец имеет номер i_1. Объявляем набор $\{1, \ldots, i_1\}$ первой вершиной будущего графа. Снова идем от $i_1 + 1$ направо до первого правого конца i_2 какой-либо дуги. Объявляем второй вершиной графа набор $\{i_1 + 1, \ldots, i_2\}$. И так далее. Поскольку правых концов у дуг в данной диаграмме n штук, получаем всего n вершин. А ребра порождаем дугами. Иными словами, две вершины соединяем ребром, коль скоро между соответствующими наборами есть дуга. Ребра ориентируем справа налево. Аналогично возникают петли. Дуг n, и ребер n (см. пример на рис. 13).

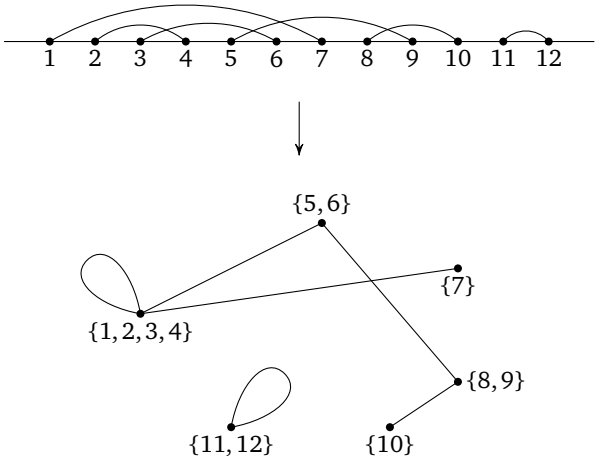

Рис. 13

Теперь считаем LCD случайной, т. е. полагаем вероятность каждой диаграммы равной $1/l_n$. Возникают случайные графы. Можно показать, что такие графы по своим вероятностным характеристикам практически неотличимы от графов G_1^n.

Графы с n вершинами и kn ребрами получаем тем же способом, что и в предыдущем пункте.

4.2.3. Некоторые результаты

Замечательна модель Боллобаша — Риордана не только тем, что с ее помощью наводится порядок в «каше», которую «заварили» Барабаши и Альберт, но еще и тем, что она полностью адекватна эмпирическим наблюдениям. Прежде всего справедлива

Теорема 43. *Для любого $k \geqslant 2$ и любого $\varepsilon > 0$*

$$P\left((1-\varepsilon)\frac{\ln n}{\ln\ln n} \leqslant diam\, G_k^n \leqslant (1+\varepsilon)\frac{\ln n}{\ln\ln n}\right) \to 1, \quad n \to \infty.$$

На первый взгляд, утверждение кажется непонятным. Ну, хорошо: диаметр плотно сконцентрирован (по вероятности) около величины $\ln n / \ln\ln n$. А у нас ведь какие-то 5—7 были? Так ничего странного. Вершин в Интернете образца 1999 года около 10^7. Значит,

$$\frac{\ln 10^7}{\ln\ln 10^7} = \frac{7\ln 10}{\ln 7 + \ln\ln 10} \approx 6.$$

Фантастическое попадание. Отметим, что при недавней проверке с другими цифрами эмпирика снова подтвердилась.

Теорема 43 доказана в работе [72] авторами модели. А в работе [73] была внесена ясность и в вопрос о распределении степеней вершин.

Теорема 44. *Для любого $k \geqslant 1$ и любого $d \leqslant n^{1/15}$*

$$M\left(\frac{\left|\left\{i = 1, \ldots, n : \mathrm{indeg}_{G_k^n} i = d\right\}\right|}{n}\right) \sim \frac{2k(k+1)}{(d+k)(d+k+1)(d+k+2)}, \qquad (2)$$

где $\mathrm{indeg}_{G_k^n} i$ *— количество ребер, имеющих вершину i своим левым концом в графе G_k^n.*

Поскольку k — константа, выражение в правой части (2) имеет вид const/d^3. Да это же в точности степенной закон! Правда, в формулировке теоремы написано математическое ожидание, а не вероятность, но одно из другого получается за счет мартингальных неравенств и соответствующих теорем о плотной концентрации меры около среднего (см. гл. 1 и [73]).

У теоремы 44 есть все же два неприятных момента. Первый состоит в том, что степень d в степенном законе, который в ней устанавливается, равна не 2,1, а 3. Второй — это ограничение $d \leqslant n^{1/15}$, которое ставит крест на практической применимости теоремы. Даже при $n \approx 10^{12}$, чего в природе (пока) не бывает, мы имеем лишь $d \leqslant 10^{4/5}$, и это нелепо.

Последний недостаток недавно устранил Е. А. Гречников — исследователь-разработчик в Яндексе, который получил более точный результат практически без ограничений на d. Грубо говоря, асимптотика (2) имеет место при всех $d = o(n)$. Статья Гречникова еще не опубликована.

Первым же недостатком занимались много и, в частности, предлагали различные альтернативные модели. Одну из таких моделей

мы обсудим в параграфе 4.3. Но прежде скажем еще несколько слов о свойствах LCD-модели.

Пусть H — фиксированный граф. Обозначим через $\sharp(H, G_k^n)$ случайную величину, равную количеству подграфов графа G_k^n, изоморфных графу H. Как распределена эта величина? Изучали ее математическое ожидание в разных специальных случаях. Например, в работе [69] приводится громоздкая общая формула и пара ее симпатичных следствий, которые мы выпишем и здесь.

Теорема 45. *Пусть $k \geqslant 2$. Пусть также K_3 — полный граф на трех вершинах. Тогда*

$$M\left(\sharp(K_3, G_k^n)\right) = (1+o(1)) \cdot \frac{k\cdot(k-1)\cdot(k+1)}{48} \cdot (\ln n)^3$$

при $n \to \infty$.

Теорема 46. *Пусть фиксированы $k \geqslant 2$ и $l \geqslant 3$. Пусть также C_l — цикл на l вершинах. Тогда*

$$M\left(\sharp(C_l, G_k^n)\right) = (1+o(1)) \cdot c_{k,l} \cdot (\ln n)^l$$

при $n \to \infty$, где $c_{k,l}$ — это положительная константа. Более того, при $k \to \infty$ имеем $c_{k,l} = \Theta(k^l)$.

Студенты МФТИ А. Рябченко и Е. Самосват недавно (в несколько иной, но очень близкой модели) установили следующий общий факт.

Теорема 47. *Пусть задан граф H, степени вершин которого равны $d_1, \ldots, d_s$. Обозначим через $\sharp(d_i = m)$ число вершин в H, степень каждой из которых равна m. Тогда*

$$M\left(\sharp(H, G_k^n)\right) = \Theta\left(n^{\sharp(d_i=0)} \cdot (\sqrt{n})^{\sharp(d_i=1)} \cdot (\ln n)^{\sharp(d_i=2)}\right).$$

Зависимость от k занесена в константу Θ.

Надо полагать, что нечто подобное было известно и авторам статьи [69], но мы ничего похожего в литературе не встречали. А такая запись результата очень удобна. Скажем, в теореме 45 речь идет про K_3. Ясно, что для K_3 выполнено

$$\sharp(d_i = 0) = \sharp(d_i = 1) = 0, \quad \sharp(d_i = 2) = 3.$$

По теореме 47

$$M\left(\sharp(K_3, G_k^n)\right) = \Theta((\ln n)^3),$$

и это прекрасно согласуется с теоремой 45. Аналогично можно разобраться и с циклами (теорема 46). А если взять K_4 — полный граф на четырех вершинах, — то теорема 47 скажет, что средняя его встречаемость в веб-графе постоянна. Иными словами, «тетраэдров» в веб-графах почти не бывает.

Отметим, что в реальном вебе случаются не только тетраэдры, но и клики куда большей мощности. Это связано с деятельностью спамеров, которые искусственно расставляют ссылки, желая повысить рейтинги сайтов, заплативших за раскрутку. Спам в модели Боллобаша — Риордана не учтен, и это тоже минус.

В следующем пункте мы приведем доказательство теоремы 44 в случае $k=1$. Мы будем работать в терминах LCD-модели.

4.2.4. Доказательство теоремы 44 при $k=1$

Доказательство использует очень красивую «школьную» комбинаторику и идеи результатов о плотной концентрации. Договоримся сперва об обозначениях. Пусть d_i — это полная степень вершины i в графе G_1^n, т. е. в нашем случае d_i и $\mathrm{indeg}_{G_1^n} i$ суть величины, отличающиеся ровно на единицу. Положим $D_m = d_1 + \ldots + d_m$.

Рассмотрим событие $\{D_m - 2m = s\}$, где $0 \leqslant s \leqslant n-m$. Очевидно, что это событие состоит в том, что последние $n-m$ вершин графа G_1^n делают s ссылок-ребер на первые m его вершин. Действительно, сумма степеней вершин графа G_1^m равна $2m$, и, значит, оставшиеся s ребер в сумме степеней D_m идут извне графа G_1^m. В терминах диаграмм это, в свою очередь, эквивалентно тому, что m-й по счету правый конец дуги совпадает с точкой $2m+s$. Попробуем вычислить вероятность описанного события.

Искомая вероятность «классическая», т. е. нам просто нужно разделить на l_n (см. п. 4.2.2) количество тех диаграмм, которые благоприятствуют нашему событию: у каждой из этих диаграмм

1) точка $2m+s$ служит правым концом некоторой дуги;
2) есть какие-то s точек среди $\{1, \ldots, 2m+s-1\}$ и какие-то s точек среди $\{2m+s+1, \ldots, 2n\}$, которые соединены s дугами (с одним концом в множестве $\{1, \ldots, 2m+s-1\}$ и другим — в множестве $\{2m+s+1, \ldots, 2n\}$) в том или ином порядке (всего возможных порядков $s!$);
3) оставшиеся точки из множества $\{1, \ldots, 2m+s-1\}$ (их $2m-2$) как-то разбиты на пары (дуги), и то же самое верно относительно оставшихся точек из множества $\{2m+s+1, \ldots, 2n\}$ (их $2n-2m-2s$).

В итоге имеем

$$P(D_m - 2m = s) = \frac{s!(2m+s-1)C^s_{2m+s-2} l_{m-1} C^s_{2n-2m-s} l_{n-m-s}}{l_n} = \\ = \frac{2^{s+1}(2m+s-1)!(2n-2m-s)!n!}{s!(m-1)!(n-m-s)!(2n)!}.$$

Сейчас мы найдем «наиболее вероятное» значение случайной величины $D_m - 2m$, т. е. величину s, при которой максимальна вероятность $p_s = P(D_m - 2m = s)$. Для этого сперва посмотрим на величину $r_s = p_{s+1}/p_s$:

$$r_s = \frac{p_{s+1}}{p_s} = 2\frac{(2m+s)(n-m-s)}{(s+1)(2n-2m-s)}.$$

Допустим, мы доказали, что функция r_s убывающая. Что это значит? А это значит, что функция p_s либо возрастает, либо убывает, либо сначала возрастает, а затем убывает. В любом случае ее максимум соответствует моменту, когда $r_s \approx 1$. Мы не пишем $r_s = 1$, так как этого может вовсе не случиться при натуральных s. По сути, едва мы докажем убывание r_s, нам понадобится найти *вещественный* корень уравнения $r_s = 1$ и взять от него верхнюю целую часть.

Итак,

$$\frac{r_{s+1}}{r_s} = \left(1 - \frac{2m-1}{(s+2)(2m+s)}\right)\left(1 - \frac{n-m}{(2n-2m-s-1)(n-m-s)}\right) \leqslant$$
$$\leqslant \left(1 - \frac{2m-1}{2n^2}\right)\left(1 - \frac{n-m}{2n^2}\right) \leqslant e^{-\frac{2m-1}{2n^2}} \cdot e^{-\frac{n-m}{2n^2}} \leqslant e^{-\frac{1}{2n}}.$$

Убывание r_s доказано. Ищем, как и обещали, корень уравнения $r_s = 1$, благо это уравнение квадратное. Получаем точку

$$s_0 = \left\lceil -2m + \sqrt{4mn - 2n + 1/4} + 1/2 \right\rceil,$$

в которой достигается максимум вероятности p_s. В окрестности этой точки величина $D_m - 2m$ плотно сконцентрирована. Оценим эту концентрацию.

Мы знаем, что $r_{s_0} \leqslant 1$. Следовательно, при $x > 0$

$$r_{s_0+x} \leqslant r_{s_0+x-1}e^{-\frac{1}{2n}} \leqslant r_{s_0+x-2}e^{-\frac{2}{2n}} \leqslant \ldots \leqslant r_{s_0}e^{-\frac{x}{2n}} \leqslant e^{-\frac{x}{2n}}.$$

Имеем в итоге

$$p_{s_0+x} \leqslant p_{s_0+x-1}e^{-\frac{x}{2n}} \leqslant p_{s_0+x-2}e^{-\frac{x}{2n}}e^{-\frac{x-1}{2n}} \leqslant \ldots \leqslant p_{s_0}e^{-\frac{x(x-1)}{4n}} \leqslant e^{-\frac{x(x-1)}{4n}}.$$

Аналогичную оценку можно доказать и для p_{s_0-x}. Значит, при достаточно больших n

$$P\left(|D_m - (2m+s_0)| \geqslant 3\sqrt{n\ln n}\right) \leqslant 2e^{-\frac{8n\ln n}{4n}} = o\left(\frac{1}{n}\right).$$

Поскольку для каждого m выполнено неравенство

$$\left|s_0 - \left(2\sqrt{mn} - 2m\right)\right| \leqslant 2\sqrt{n},$$

имеем

$$P\left(\left|D_m - 2\sqrt{mn}\right| \geqslant 4\sqrt{n\ln n}\right) = o\left(\frac{1}{n}\right),$$

т. е. величина D_m при не слишком больших m плотно сконцентрирована около $2\sqrt{mn}$.

Теперь мы готовы найти условную вероятность

$$P\left(\operatorname{indeg}_{G_1^n}(m+1) = d \,|\, D_m - 2m = s\right) = P\left(d_{m+1} = d+1 \,|\, D_m - 2m = s\right).$$

Коль скоро условие $D_m - 2m = s$ выполнено, часть LCD уже зафиксирована. А именно, среди точек $1, \ldots, 2m+s-1$ есть множество точек A мощности s, состоящее из левых концов дуг, правые концы которых находятся где-то среди точек $2m+s+1, \ldots, 2n$; в множестве $\{1, \ldots, 2m+s-1\} \setminus A$ есть некоторая точка, служащая левым концом дуги, правый конец которой совпадает с $2m+s$; остальные $2m-2$ точки как-то разбиты на пары-дуги. При любом описанном способе фиксации «левой» части LCD есть $s! C^s_{2n-2m-s} l_{n-m-s}$ способов достроить ее до полноценной диаграммы. Указанное количество и будет знаменателем искомой вероятности. В числителе же будет количество t способов достроить левую часть LCD при дополнительном ограничении $d_{m+1} = d+1$, которое означает следующее:

1) точка $2m+s+d+1$ служит правым концом некоторой дуги, а все точки из множества $B = \{2m+s+1, \ldots, 2m+s+d\}$ являются левыми концами дуг;
2) среди точек из $A \cup B$ (их $s+d$) есть левый конец x дуги с правым концом в точке $2m+s+d+1$;
3) в множестве $\{2m+s+d+2, \ldots, 2n\}$ (его мощность равна $2n - 2m - s - d - 1$) есть подмножество C, состоящее из $s+d-1$ правых концов дуг, левые концы которых принадлежат множеству $A \cup B \setminus \{x\}$;
4) множество $\{2m+s+d+2, \ldots, 2n\} \setminus C$ (его мощность равна $2n - 2m - 2s - 2d$) разбито на дуги.

Ясно, что

$$t = (s+d) C^{s+d-1}_{2n-2m-s-d-1} (s+d-1)! l_{n-m-s-d}.$$

Здесь $s+d$ — это число способов выбрать x, биномиальный коэффициент — это число способов выбрать C, величина $(s+d-1)!$ выражает количество способов соединить дугами точки из множеств $A \cup B \setminus \{x\}$ и C, а множитель $l_{n-m-s-d}$ возникает за счет свойства 4.

Итак,

$$P\big(d_{m+1}=d+1|D_m-2m=s\big)=\frac{(s+d)C^{s+d-1}_{2n-2m-s-d-1}(s+d-1)!l_{n-m-s-d}}{s!C^{s}_{2n-2m-s}l_{n-m-s}}=$$
$$=2^d(s+d)\frac{(n-m-s)_d}{(2n-2m-s)_{d+1}},\quad (a)_b=\frac{a!}{(a-b)!}.$$

Разобьем область значений величин m, d и s на две части. Сперва опишем первую часть $\mathscr{L}_1$ (тогда ко второй части $\mathscr{L}_2$ отойдет все остальное). В самом деле, пусть $M=[n^{4/5}/\ln n]$. Возьмем в качестве m любое число в пределах $M\leqslant m\leqslant n-M$. В качестве d возьмем любое число из промежутка $0\leqslant d\leqslant n^{1/15}$. А s выберем из соотношения $2m+s=D$, где D — любое число, удовлетворяющее неравенству $|D-2\sqrt{mn}|\leqslant 4\sqrt{n\ln n}$. Иными словами, D — это одно из наиболее вероятных значений величины D_m.

Понятно, что внутри $\mathscr{L}_1$ мы имеем

$$s+d=D-2m+d=2\sqrt{mn}-2m+O\big(\sqrt{n\ln n}\big)+O\big(n^{1/15}\big)=$$
$$=2\sqrt{mn}-2m+O\big(\sqrt{n\ln n}\big),$$
$$n-m-s=n+m-2\sqrt{mn}+O\big(\sqrt{n\ln n}\big),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$n-m-s-d=n+m-2\sqrt{mn}+O\big(\sqrt{n\ln n}\big),$$

причем константы во всех о-больших можно сделать одинаковыми. Аналогично

$$2n-2m-s=2n-2\sqrt{mn}+O\big(\sqrt{n\ln n}\big),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$2n-2m-s-d=2n-2\sqrt{mn}+O\Big(\sqrt{n\ln n}\Big)$$

с одними и теми же константами в о-большом. Значит,

$$P\big(d_{m+1}=d+1|D_m-2m=s\big)=$$
$$=2^d\big(2\sqrt{mn}-2m+O\big(\sqrt{n\ln n}\big)\big)\frac{\big(n+m-2\sqrt{mn}+O\big(\sqrt{n\ln n}\big)\big)^d}{\big(2n-2\sqrt{mn}+O\big(\sqrt{n\ln n}\big)\big)^{d+1}}.$$

Нетрудно видеть, далее, что

$$2\sqrt{mn}-2m+O\big(\sqrt{n\ln n}\big)\sim 2\sqrt{mn}-2m.$$

Более того, мы покажем, что

$$\left(n+m-2\sqrt{mn}+O\left(\sqrt{n\ln n}\right)\right)^d \sim \left(n+m-2\sqrt{mn}\right)^d,$$
$$\left(2n-2\sqrt{mn}+O\left(\sqrt{n\ln n}\right)\right)^{d+1} \sim \left(2n-2\sqrt{mn}\right)^{d+1}.$$

Вернее, мы проверим только первую асимптотику, так как вторая получается совершенно аналогично. Итак,

$$\left(n+m-2\sqrt{mn}+O\left(\sqrt{n\ln n}\right)\right)^d = \\ = \left(n+m-2\sqrt{mn}\right)^d\left(1+O\left(\frac{\sqrt{n\ln n}}{n+m-2\sqrt{mn}}\right)\right)^d.$$

Далее,

$$\left(1+O\left(\frac{\sqrt{n\ln n}}{n+m-2\sqrt{mn}}\right)\right)^d \leqslant e^{O\left(\frac{d\sqrt{n\ln n}}{n+m-2\sqrt{mn}}\right)},$$

и нам остается показать, что

$$\frac{d\sqrt{n\ln n}}{n+m-2\sqrt{mn}} \to 0.$$

Поскольку $m \leqslant n-M$, можно написать

$$\frac{d\sqrt{n\ln n}}{n+m-2\sqrt{mn}} = \frac{d\sqrt{n\ln n}}{\left(\sqrt{n}-\sqrt{m}\right)^2} = \frac{d\sqrt{n\ln n}}{n\left(1-\sqrt{\frac{m}{n}}\right)^2} \leqslant \frac{d\sqrt{n\ln n}}{n\left(1-\sqrt{\frac{n-M}{n}}\right)^2} = \\ = \frac{d\sqrt{n\ln n}}{n\left(1-\sqrt{1-O\left(\frac{1}{n^{1/5}\ln n}\right)}\right)^2} = \frac{d\sqrt{n\ln n}}{nO\left(\frac{1}{n^{2/5}\ln^2 n}\right)} = \\ = O\left(\frac{n^{1/15}\sqrt{n\ln n}\ln^2 n}{n^{3/5}}\right) = O\left(\frac{(\ln n)^{5/2}}{n^{1/30}}\right) = o(1).$$

Мы доказали обещанные асимптотики. Значит,

$$P(d_{m+1} = d+1 | D_m = D) \sim \\ \sim \frac{2\sqrt{mn}-2m}{2m-2\sqrt{mn}}\left(\frac{2\left(\sqrt{n}-\sqrt{m}\right)^2}{2\left(n-\sqrt{mn}\right)}\right)^d \sim \sqrt{\frac{m}{n}}\left(1-\sqrt{\frac{m}{n}}\right)^d,$$

причем асимптотика равномерна по всем D из нашей области $\mathscr{L}_1$, т. е. по всем D, таким что $\left|D-2\sqrt{mn}\right| \leqslant 4\sqrt{n\ln n}$. Имеем, стало быть, по

формуле полной вероятности (см. § 1.5)

$$P(d_{m+1}=d+1)=\sum_{D:\,\left|D-2\sqrt{mn}\right|\leqslant 4\sqrt{n\ln n}} P(d_{m+1}=d+1|D_m=D)P(D_m=D)+$$

$$+\sum_{D:\,\left|D-2\sqrt{mn}\right|>4\sqrt{n\ln n}} P(d_{m+1}=d+1|D_m=D)P(D_m=D)=$$

$$=(1+o(1))\sqrt{\frac{m}{n}}\left(1-\sqrt{\frac{m}{n}}\right)^d \sum_{D:\,\left|D-2\sqrt{mn}\right|\leqslant 4\sqrt{n\ln n}} P(D_m=D)+$$

$$+O\left(\sum_{D:\,\left|D-2\sqrt{mn}\right|>4\sqrt{n\ln n}} P(D_m=D)\right)=$$

$$=(1+o(1))\sqrt{\frac{m}{n}}\left(1-\sqrt{\frac{m}{n}}\right)^d+o(1/n),$$

ведь

$$1\geqslant \sum_{D:\,|D-2\sqrt{mn}|\leqslant 4\sqrt{n\ln n}} P(D_m=D)=P\left(|D_m-2\sqrt{mn}|\leqslant 4\sqrt{n\ln n}\right)\geqslant$$

$$\geqslant (1-o(1/n))\sim 1,$$

$$\sum_{D:\,|D-2\sqrt{mn}|>4\sqrt{n\ln n}} P(D_m=D)=P\left(|D_m-2\sqrt{mn}|>4\sqrt{n\ln n}\right)=o(1/n).$$

Итак, для всех наборов m, d, s из области $\mathscr{L}_1$ имеем равномерную по этой области асимптотику

$$P(d_{m+1}=d+1)=(1+o(1))\sqrt{\frac{m}{n}}\left(1-\sqrt{\frac{m}{n}}\right)^d+o(1/n).$$

Обозначим через ξ количество вершин вида $m+1$ с $m\in[M, n-M]$ и со свойством $d_{m+1}=d+1$ (мы по-прежнему живем в области $\mathscr{L}_1$). Тогда ввиду линейности математического ожидания

$$M\xi=\sum_{m=M}^{n-M} P(d_{m+1}=d+1)=o(1)+\sum_{m=M}^{n-M}(1+o(1))\sqrt{\frac{m}{n}}\left(1-\sqrt{\frac{m}{n}}\right)^d.$$

Искомое же математическое ожидание отличается от уже найденного на величину порядка M:

$$M\left|\{i=1,\ldots,n\colon \operatorname{indeg}_{G_1^n} i=d\}\right|=$$

$$=O(M)+o(1)+(1+o(1))\sum_{m=M}^{n-M}\sqrt{\frac{m}{n}}\left(1-\sqrt{\frac{m}{n}}\right)^d.$$

Можно показать (интуитивно это понятно), что

$$\sum_{m=M}^{n-M}\sqrt{\frac{m}{n}}\left(1-\sqrt{\frac{m}{n}}\right)^d \sim$$

$$\sim n\int\limits_{(M+1)/n}^{1-M/n}\sqrt{v}(1-\sqrt{v})^d\,dv \sim n\int\limits_0^1\sqrt{v}(1-\sqrt{v})^d\,dv.$$

Последний интеграл явно вычисляется путем замены $v=(1-u)^2$:

$$\int\limits_0^1\sqrt{v}(1-\sqrt{v})^d\,dv = 2\int\limits_0^1(1-u)^2u^d\,du =$$

$$= 2\int\limits_0^1(u^d-2u^{d+1}+u^{d+2})\,du = \frac{4}{(d+1)(d+2)(d+3)},$$

и мы получаем, наконец,

$$M\left|\{i=1,\ldots,n\colon \operatorname{indeg}_{G_1^n} i=d\}\right| =$$

$$= O(M)+(1+o(1))\frac{4n}{(d+1)(d+2)(d+3)} \sim \frac{4n}{(d+1)(d+2)(d+3)},$$

что и требовалось доказать.

4.3. Модель копирования

Здесь мы опишем еще одну очень интересную модель, которая также призвана объяснить феномен степенного закона в реальных сетях. Эта модель возникла практически в одно время с моделью Барабаши — Альберт. Она принадлежит Р. Кумару, П. Рагхавану, С. Раджагопалану, Д. Сивакумару, А. Томкинсу и Э. Упфалу (см. [74]).

Фиксируем $\alpha\in(0,1)$ и $d\geqslant 1$, $d\in\mathbb{N}$. Случайный граф будет расти, и это будет похоже на процесс из пункта 4.2.1. Однако здесь процесс будет устроен совсем по-другому.

В качестве начального графа возьмем любой d-регулярный граф (граф, у которого степень каждой вершины равна d). Пусть построен граф с номером t. Обозначим его $G_t=(V_t,E_t)$. Здесь $V_t=\{u_1,\ldots,u_s\}$, где s отличается от t на число вершин начального графа, т. е. на некоторую константу, выражаемую через d. Добавим к G_t одну новую вершину u_{s+1} и d ребер, выходящих из u_{s+1}. Для этого сперва выберем случайную вершину $p\in V_t$ (все вершины в V_t равновероятны). Одно за другим строим ребра из u_{s+1} в V_t. На шаге с номером i, $i\in\{1,\ldots,d\}$,

разыгрываем случайную величину, которая с вероятностью α принимает значение 1 («монетка падает решкой кверху») и с вероятностью $1-\alpha$ принимает значение 0 («монетка падает орлом кверху»). Если вышла единица, то выпускаем ребро из u_{s+1} в случайную вершину из V_t (все вершины в V_t равновероятны). Если вышел ноль, то берем i-го по номеру соседа вершины p. Последнее действие всегда возможно, так как по построению у каждой вершины не менее d соседей.

Объяснить это можно так. Появляется новый сайт. Проставляя очередную ссылку, его владелец с некоторой вероятностью будет ориентироваться на кого-то из своих предшественников. Скажем, сайт посвящен автомобилям. Вероятно, владелец возьмет один из уже существовавших сайтов про автомобили и *скопирует* оттуда ссылку (с точки зрения стороннего наблюдателя, вполне случайную). Это ситуация, когда монетка выпала орлом кверху (p — это сайт, с которого копируются ссылки). Однако при простановке ссылки владелец может и никого не копировать, а случайно (по нашему мнению) цитировать кого-то из предшественников. Это случай выпадения решки. Таким образом, $1-\alpha$ — это вероятность копирования или, если угодно, вероятность выбора, мотивированного тематикой сайта.

Основной результат из [74] — это теорема 48.

Теорема 48. *Пусть $N_{t,r}$ — это математическое ожидание числа вершин степени r в графе G_t. Тогда*

$$\lim_{t\to\infty}\frac{N_{t,r}}{t}=\Theta\left(r^{-\frac{2-\alpha}{1-\alpha}}\right).$$

Пафос теоремы в том, что в ней мы снова приходим к степенному закону. Более того, если вероятность копирования близка к 1 (а величина α — к нулю), то показатель степени может равняться ожидаемой величине 2,1, чего до сих пор у нас не было.

В целом распределение степеней вершин в модели копирования очень похоже на распределение степеней вершин в модели Боллобаша — Риордана. В остальном модели сильно разнятся. Например, в модели Боллобаша — Риордана практически отсутствуют плотные двудольные подграфы (см. теорему 47); в модели копирования таких подграфов полно. Это особенно важно ввиду того, что спамерские структуры, о которых мы вскользь говорили в конце пункта 4.2.3, зачастую образуют именно двудольные графы с плотной перелинковкой.