Математические модели интернета

А.РАЙГОРОДСКИЙ

Что такое интернет?

Еще в середине 90-х годов XX века, каких-то 15–20 лет назад, про интернет практически никто не знал. А если и знал про его существование, то вряд ли имел к нему доступ. И потому вопрос, поставленный в заголовке этого раздела, тогдашнему читателю показался бы вполне уместным. Однако сейчас, когда интернет прочно вошел в нашу жизнь, этот вопрос должен вызвать недоумение: «Ну, как «что такое»? Ясное дело. Это источник информации, удобная площадка для общения. Там есть сайты, на них страницы, есть блоги, социальные сети, спам и пр.». Все это верно. И тем не менее, так ли уж много мы знаем об устройстве «всемирной паутины»? Прежде всего, понимаем ли мы, какие законы правят ее формированием? А может, и вовсе нет никаких законов? Ведь, казалось бы, интернет – это совершенно случайная, никем не контролируемая среда, и, стало быть, ничто не мешает ее *непредсказуемому* развитию. Что ж, давайте обсудим.

Будем представлять интернет в виде графа. Вершинами этого графа будут сайты, и между двумя вершинами A, B мы проведем столько ребер, направленных от A к B, сколько есть ссылок с сайта A на сайт B, и столько ребер, направленных от B к A, сколько есть ссылок с сайта B на сайт A. Нас будет интересовать устройство этого графа, который по понятным причинам принято называть *веб-графом*. Оказывается, вопреки сделанному выше предположению о полной непредсказуемости в поведении всемирной паутины, веб-граф обладает рядом устойчивых свойств – свойств, которые остаются неизменными на протяжении всей истории исследований интернета. Не претендуя на полноту, опишем несколько таких свойств. Их уже будет достаточно для понимания того, как сильно подчас реальная картина мира противоречит нашей интуиции.

Первое свойство веб-графа многим хорошо известно, хотя обычно речь идет о другом. Говорят о законе «шести рукопожатий». А именно, у каждого человека

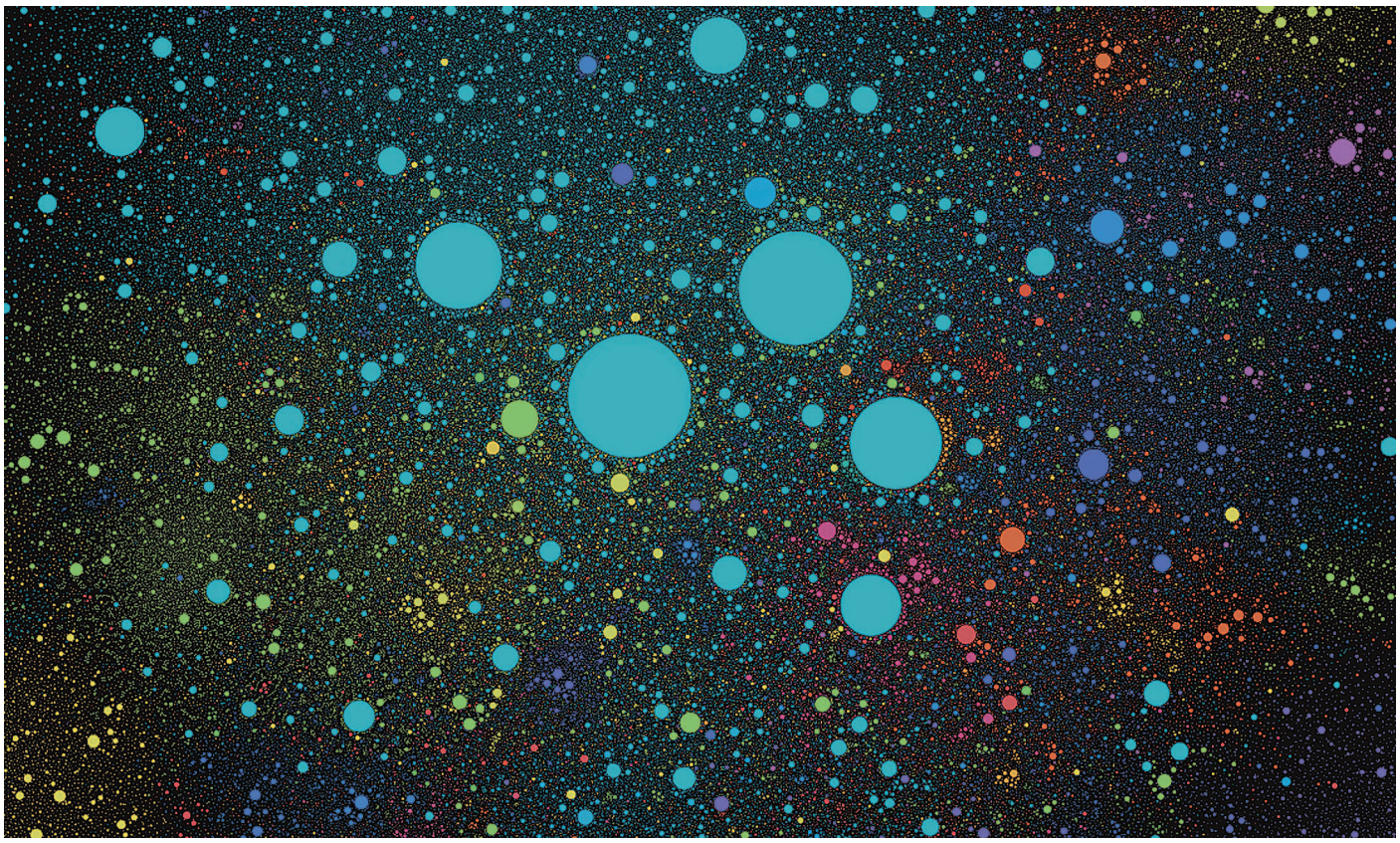

Схема интернета с сайта internet-map.net

есть знакомые, у этих людей также есть свои знакомые и т.д. Наблюдение состоит в том, что от любого человека до любого другого человека на Земле можно «пройти» по такой цепочке взаимных знакомств и что количество «звеньев» в ней не превзойдет шести. Иными словами, я пожму руку своему другу, он пожмет руку одному из своих приятелей, и через не более чем шесть таких рукопожатий (при правильном выборе их последовательности) очередным знакомым окажется президент страны или какой-нибудь разносчик пиццы из Огайо, – вообще, любой наперед заданный человек. Ровно та же история с интернетом. Только здесь рукопожатия заменяются «кликами» (компьютерной мышью по ссылкам) и утверждается, что для перехода с любого сайта на любой другой сайт потребуется не более шести кликов (при правильном выборе их цепочки).

В терминах теории графов речь идет о *диаметре* графа. Дадим его определение. *Расстоянием* между вершинами графа называется количество ребер в кратчайшей реберной цепочке, соединяющей эти вершины. Если граф ориентированный (как, например, веб-граф), то все ребра в рассматриваемых цепочках должны следовать друг за другом в одном и том же направлении. Диаметр – это самое большое расстояние между вершинами в графе. Разумеется, бывают несвязные графы. У каждого из них диаметр считается равным бесконечности. Закон шести кликов – это факт, состоящий в том, что диаметр веб-графа равен шести. Для обозначения диаметра графа G используют запись $\mathrm{diam}\ G$.

Описанное свойство отлично характеризуется выражением «мир тесен». Казалось бы, это должно означать, что в веб-графе довольно много ребер. Как бы не так! И тут интуиция нас подводит. Второе свойство веб-графа состоит в его исключительной «разреженности». Грубо говоря, если вершин у веб-графа n, то ребер у него не более mn с некоторым постоянным $m \geq 1$. Давайте поймем, почему это мало. В самом деле, даже если пренебречь тем, что в веб-графе бывают кратные ребра и кратные петли (с одних страниц данного сайта вполне могут идти ссылки на другие его же страницы), ему ничто не мешает иметь $C_n^2 = \frac{n(n-1)}{2}$ ребер. Но последняя величина растет квадратично по n, тогда как реальное количество ребер значительно меньше: их, как максимум, mn. В некотором смысле особенно меньше и быть-то не может: если у графа на n вершинах меньше чем $n - 1$ ребер, то этот граф заведомо не связен.

И еще одно, третье, свойство. Давайте смотреть на *степени* вершин веб-графа. Тут, конечно, есть тонкость, связанная с тем, что у ориентированного графа бывают как *входящие степени* $\mathrm{indeg}\ v$ (число ребер, правым концом которых служит данная вершина v), так и *исходящие степени* $\mathrm{outdeg}\ v$. Если не оговорено противное, мы будем понимать под степенью вершины сумму ее входящей и исходящей степеней, т.е. число всех ребер, концом которых она является: $\deg v = \mathrm{indeg}\ v + \mathrm{outdeg}\ v$. Нас интересует доля вершин веб-графа, имеющих данную степень. Иными словами, пусть n – количество вершин веб-графа, а d – некоторое фиксированное число. Обозначим через $\#(n,d)$ величину

$$\frac{\left|\{v : \deg v = d\}\right|}{n},$$

т.е. мы делим количество вершин степени d (модулем обозначена мощность множества, заключенного в фигурных скобках) на общее количество вершин и получаем искомую долю. Оказывается, что всегда

$$\#(n,d) \approx \frac{c}{d^{2,3}}.$$

Здесь $d \neq 0$, поскольку веб-граф связен, а c – константа, которую легко найти, ведь мы знаем, что сумма всех величин $\left|\{v : \deg v = d\}\right|$ равна n, откуда $\sum_d \#(n,d) = 1$. В сущности, $\#(n,d)$ – это *вероятность* того, что вершина графа имеет степень d, а сумма всех вероятностей должна равняться единице. Гораздо удивительнее здесь константа 2,3, которая не меняется с течением времени! Описанное свойство называется *степенным законом распределения* степеней вершин веб-графа.

Итак, ситуация весьма любопытная. Несмотря на кажущуюся хаотичность в процессе образования интернета, есть весьма жесткие статистические ограничения, которым он годами подчиняется. Почему это так? Что стоит за всеми свойствами интернета? Каковы законы, управляющие формированием сети? Мало того, что все эти вопросы крайне важны для понимания устройства мира, – ответы на них не могут не принести и серьезную практическую пользу: имея правильную *модель* интернета, можно пытаться лучше выявлять некоторые виды спама («неестественные ссылочные структуры», называемые *линковыми кольцами*), тестировать алгоритмы обхода интернета поисковым роботом и др. В следующих разделах мы обсудим все это – и модели, и приложения.

Идея предпочтительного присоединения

В 1999 году двое исследователей – А.Л.Барабаши и Р.Альберт – предложили крайне простую идею, которая, однако ж, оказалась весьма продуктивной. Идея заключалась в том, что когда новый сайт появляется на свет, он, скорее всего, «предпочитает» сослаться на те сайты, которые и без того уже многими цитированы. Более точно, вероятность, с которой новый сайт ставит ссылку на сайт-предшественник, пропорциональна (входящей) степени вершины веб-графа, отвечающей этому сайту. Один из наиболее удобных и математически строгих вариантов реализации идеи Барабаши–Альберта (идеи о *предпочтительном присоединении*) сформулировали в 2000 году математики Б.Боллобаш и О.Риордан.

Построение модели Боллобаша–Риордана состоит из двух этапов. Сперва строится последовательность графов G_1^n, $n = 1, 2, 3, \ldots$ У этих графов будет по n вершин и n ребер. Затем эта последовательность с помощью несложного трюка преобразуется в последовательность

G_m^n, где m – натуральное число, а количество ребер графа G_m^n, имеющего n вершин, равно mn. Таким образом, графы G_m^n автоматически оказываются обладающими вторым свойством веб-графа.

Итак, начнем с G_1^n. Будем строить эти графы по индукции. Пусть G_1^1 – это граф с одной вершиной (обозначим ее просто 1) и одной петлей $(1, 1)$. [1] Предположим, граф G_1^{n-1} с $n \geq 2$ уже построен. Обозначим его вершины $1, \ldots, n-1$ и будем помнить, что ребер у него, как и вершин, $n-1$. Граф G_1^n мы получим путем добавления к графу G_1^{n-1} одной вершины (одного сайта) с «именем» n и одного ребра (ссылки, которую делает новый сайт). Это ребро будет направлено либо из n в n (если снова шутить, то в данном случае надо говорить уже не об одиночестве, а о самолюбовании), либо из n в какую-то вершину $v \in \{1, \ldots, n-1\}$. Само направление выбирается случайно: с вероятностью $\frac{1}{2n-1}$ ссылка из n пойдет на само n («самолюбование»); с вероятностью $\frac{\deg v}{2n-1}$ сайт n процитирует сайт $v \in \{1, \ldots, n-1\}$. Здесь $\deg v$ – степень вершины v в графе G_1^{n-1}, т.е. в чистом виде реализуется идея предпочтительного присоединения. В делении на $2n-1$ также нет ничего загадочного. Просто сумма вероятностей должна равняться единице, но $\sum_{v=1}^{n-1} \deg v$ – это удвоенное число ребер графа G_1^{n-1}, т.е. $2n-2$.

Подчеркнем, что построенная последовательность графов *случайная*. В принципе, она может принимать весьма разнообразные формы в зависимости от того, что произойдет на очередном этапе ее построения. Но это и хорошо: ведь с самого начала интуиция говорила нам, что интернет «случаен». Важно лишь, какие законы управляют этой случайностью, и сейчас мы допускаем, что один из этих законов – по сути, психологический – это закон предпочтительного присоединения.

Перейдем ко второму этапу. Зафиксируем натуральное $m \geq 2$. Рассмотрим найденный на первом этапе граф G_1^{mn}. У него mn вершин и mn ребер. Обозначим v_1 группу из первых m его вершин, т.е. множество $\{1, \ldots, m\}$. Обозначим v_2 следующую группу его вершин $\{m+1, \ldots, 2m\}$. И так далее. Таким образом, у нас возникнут группы $v_1, \ldots, v_n$. Будем считать их вершинами нового графа G_m^n. Для каждого i образуем в вершине v_i столько петель, сколько есть ребер в графе G_1^{mn} между теми его вершинами, множество которых мы обозначили v_i. Для любых i, j с условием $1 \leq i < j \leq n$ проведем столько ребер из v_j в v_i, сколько есть в графе G_1^{mn} ребер, правый конец которых расположен в множестве, отвечающем v_j, а левый – в множестве, отвечающем v_i. В некотором смысле, мы схлопываем вершины из G_1^{mn} в своего рода «метасайты», а все прежние ссылки сохраняем. На выходе имеем граф с n вершинами и mn ребрами. Пример перехода от графа G_1^{12} к графу G_2^6 показан на рисунке 1.

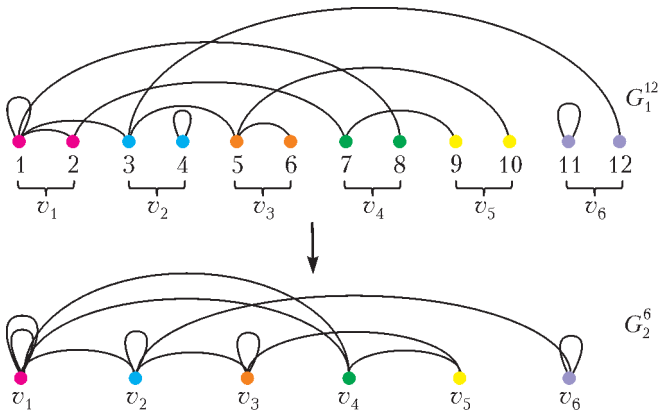

Рис. 1

Конструкция кажется довольно искусственной и, уж как минимум, чрезвычайно упрощенной. Тем не менее, она на удивление хорошо отражает ожидаемые свойства веб-графа. Про второе свойство говорить не нужно, так как оно заложено прямо в построение. Обсудим, стало быть, первое свойство. Авторы модели – Боллобаш и Риордан – доказали, что при $m \geq 2$ и при любом $\varepsilon > 0$ с увеличением числа вершин графа G_m^n все ближе к единице становится вероятность того, что диаметр графа G_m^n заключен в пределах от $(1-\varepsilon)\frac{\ln n}{\ln \ln n}$ до $(1+\varepsilon)\frac{\ln n}{\ln \ln n}$. Иными словами, хотя граф и случаен, но «почти наверняка» его диаметр практически не отличается от дроби $\frac{\ln n}{\ln \ln n}$. Почему это хорошо? А дело в том, что у веб-графа порядка $10^7 - 10^8$ вершин. Подставляя оба числа вместо n в отношение логарифмов, получаем 5,8–6,2, т.е. 6, ведь диаметр – целое число. На самом деле, даже при $n = 10^9$ отношение логарифмов не выходит за пределы семерки. Это очень медленно растущая функция, и это может объяснять, почему закон шести рукопожатий столь незыблем. Не правда ли, замечательное попадание?

Но и это не все. Есть еще третье свойство. Оно также наблюдается в модели Боллобаша–Риордана. Авторы модели доказали это при определенных ограничениях на величину d, фигурирующую в свойстве. Недавно Е.Гречников – выпускник механико-математического факультета МГУ и Независимого университета, работающий сейчас в отделе теоретических и прикладных исследований компании Яндекс, – доказал, что для всех m и d с ростом n все ближе к единице становится вероятность того, что величина $\#(n,d)$, определенная на графе G_m^n, практически не отличается от величины $\frac{c}{d^3}$, где c зависит лишь от m.

В последнем результате все чуть менее радужно, нежели в результате о диаметре: все-таки 3 – это не 2,3. Да, это тоже степенной закон, и это замечательно, но

[1] Однажды на лекции я пошутил: дескать, в начале всех времен был один сайт, было ему очень одиноко, и решил он поставить ссылку сам на себя… Голос из аудитории: «А что это был за сайт?» Разумеется, речь идет о модели, а не о реальности.

степень в нем немного другая. Что ж: никто и не обещал, что тривиальная модель сразу решит все проблемы.

Уточнение модели Боллобаша–Риордана

В конце предыдущего раздела мы поняли, что модель Боллобаша–Риордана не совсем адекватно отражает даже те три свойства веб-графа, которые мы выделили с самого начала. А именно, есть небольшие проблемы со степенным законом. Эти проблемы несложно устранить, и многие исследователи в разное время приходили к одному и тому же простому решению. Мы выделим здесь П. Бакли и Д. Остгуса, которые первыми дали строгое обоснование этого подхода. Вслед за Бакли и Остгусом возьмем произвольное число $a > 0$. Будем строить графы $H^n_{a,m}$ по практически той же схеме, по какой строились графы G^n_m. Лишь слегка изменим вероятности в определении G^n_1. Если там они равнялись $\frac{1}{2n-1}$ и $\frac{\deg v}{2n-1}$, то тут мы положим их равными $\frac{a}{(a+1)n-1}$, $\frac{\deg v + a - 1}{(a+1)n-1}$. С суммой все вновь в порядке:

$$\sum_{v=1}^{n-1} \frac{\deg v + a - 1}{(a+1)n-1} + \frac{a}{(a+1)n-1} = \frac{2n-2+(n-1)(a-1)+a}{(a+1)n-1} = 1.$$

Более того, при $a = 1$ имеем модель Боллобаша–Риордана. Число a называется *начальной притягательностью вершины*. Смысл в том, что независимо от степени вершины оно дает дополнительный вклад в вероятность присоединения.

Замечательно то, что этого хватает! Утверждение про диаметр остается неизменным, а величина $\#(n,d)$, определенная на графе $H^n_{a,m}$, становится почти наверняка приближенно равной $\frac{c}{d^{2+a}}$. Таким образом, при $a = 0{,}3$ получаем полное соответствие модели той части реальности, которая отражена в выделенных нами трех свойствах интернета.

К сожалению, доказательства результатов и этого, и предыдущего разделов крайне трудны и техничны. Поэтому они выходят за рамки этой статьи. Заинтересованного читателя мы отсылаем к книгам [1], [2] и к статье [3].

А в следующем разделе мы расскажем еще об одной идее построения модели интернета.

Модель копирования

Здесь идея такая: когда появляется новый сайт, он либо цитирует какого-то «случайного» (с точки зрения стороннего наблюдателя) предшественника, либо *копирует* ссылки с некоторого (также случайного) сайта, чья тематика близка его автору. Эта идея призвана объяснить не только степенной закон, но и факт наличия в интернете плотных сообществ, участники которых объединены общими интересами.

Строгое описание простейшего варианта модели копирования следующее. Дано натуральное число m и действительное число $\alpha \in (0,1)$. Как и в случае моделей Боллобаша–Риордана и Бакли–Остгуса, строится случайная последовательность графов $G^n_{m,\alpha}$. И строится она тоже по индукции. В начальный момент времени есть одна вершина 1 и m петель в ней. Пусть $n \geq 2$ и граф $G^{n-1}_{m,\alpha}$ с вершинами 1, …, $n-1$ уже построен. Добавим к нему вершину n и m исходящих из нее ребер. На сей раз «самолюбование» исключается, и ребра идут в вершины 1, …, $n-1$. Опишем, как устроен выбор их левых концов. Прежде всего выбирается случайная вершина $p \in \{1, \ldots, n-1\}$. Имеется в виду, что p принимает тот или иной конкретный вид с вероятностью $\frac{1}{n-1}$. Что ж, выбрали p и зафиксировали. Это будет тот самый сайт, тематика которого интересна автору сайта n. С него он иногда будет копировать ссылки. Теперь укажем первое ребро, исходящее из n. Для этого бросим «кривую» монетку, которая с вероятностью α ложится кверху орлом и с вероятностью $1-\alpha$ ложится кверху решкой. Если выпала решка, то отправляем наше ребро в первую по величине вершину среди тех, на которые ссылается сайт p. Иными словами, с вероятностью $1-\alpha$ мы копируем первую ссылку с сайта p. Если выпал орел, то мы ничего не копируем, а случайно выбираем вершину среди $\{1, \ldots, n-1\}$ и отправляем ребро в нее. Второе ребро, исходящее из n, ищется точно так же. С вероятностью $1-\alpha$ оно идет во вторую по величине вершину из числа тех, на которые ссылается p; с вероятностью α его левый конец выбирается случайно. И так далее. Поскольку на каждом шаге построения очередная вершина испускает m ребер, то у p ровно m соседей, и описанную процедуру мы сможем проделать необходимые m раз.

Можно показать, что с близкой к единице вероятностью величина $\#(n,d)$, определенная на графе $G^n_{m,\alpha}$, ведет себя примерно как $\frac{c}{d^{\frac{2-\alpha}{1-\alpha}}}$, где c – константа, зависящая от m и α. Результат замечательный, так как снова при правильном подборе α мы можем получить любой показатель степени d, больший двойки, и, в частности, показатель 2,3. В последнем случае вероятность копирования довольно близка к единице.

И это все?

Вопрос, поставленный в заголовке этого раздела, вполне может прийти в голову пытливому читателю. Да, конечно, мы рассказали о паре идей, красиво объясняющих закономерности в «жизни» интернета – закономерности, которые поначалу казались столь удивительными. Но ведь ясно, что у описанных моделей есть масса недостатков.

Например, в графах, которые могут возникнуть в рамках моделей, каждая вершина имеет *фиксированную* исходящую степень. Ничего подобного в интернете нет! По сути, получается, что ориентация в смоделированных графах носит весьма условный характер. Вряд

ли структура графов сильно поменяется, если мы снимем с ребер все «стрелки».

Кроме того, ясно, что в моделях у более старых вершин гораздо больше шансов иметь большую входящую степень, нежели у более новых вершин. Это заведомо плохо согласуется с новостными «взрывами», которые ежедневно случаются в интернете: едва появляется страница с важной новостью, как на нее приходят тысячи ссылок. Мы уж не говорим о том, что сайты, страницы и ссылки на них зачастую умирают, и это тоже никак не отражено в моделях.

Разумеется, люди, которые занимаются исследованиями сети, прекрасно все это понимают. И к настоящему времени придумано очень много разных моделей интернета, которые куда более адекватны реальности, чем модели Боллобаша–Риордана, Бакли–Остгуса или модели копирования. Это огромная увлекательная область теории *случайных графов*, которую еще предстоит развивать и систематизировать. Пафос рассмотренных нами примеров в том, что они как нельзя лучше демонстрируют, насколько простыми могут быть принципы, лежащие в основе весьма сложных явлений. А до полного решения проблемы далеко, и это только приятно: нынешнему читателю наверняка найдется, чем заняться, если он захочет исследовать веб-графы.

На этом можно было бы поставить точку, но мы еще скажем пару слов о приложениях моделей к практике поиска в интернете.

Об одном приложении

Давайте рассмотрим один из видов спамерской деятельности. А именно, поговорим о «линковых кольцах». Сразу заметим, что название сложилось исторически: когда-то спамеры, желая обмануть поисковую систему и повысить свои позиции в поисковой выдаче, цитировали друг друга по кругу, т.е. A_1 ставил ссылку на A_2, A_2 – на A_3,..., A_n – на A_1. Такую схему быстро научились изобличать, спамерам пришлось стать хитрее, но название «линковое кольцо» осталось.

Сейчас типичная конструкция – это своего рода двудольный граф (рис.2). Вершины из (условно) правой доли – это покупатели ссылок, те, кто таким образом надеется показать, что имеет высокий «индекс цитирования», и потому должен быть поднят на самый верх в поисковой выдаче. Вершины из левой доли – это продавцы ссылок. Последние – вовсе не обязательно какие-то «маргиналы». Напротив, часто ссылки покупают у вполне респектабельных владельцев, которым понадобились деньги на решение каких-либо проблем: с точки зрения поисковой системы, сайт, процитированный респектабельным сайтом, сам становится кандидатом в члены «клуба уважаемых граждан». Ребра в линковом кольце идут, в основном, слева направо; также много ребер может быть внутри левой доли (как просто за счет респектабельности, так и ради «накрутки»).

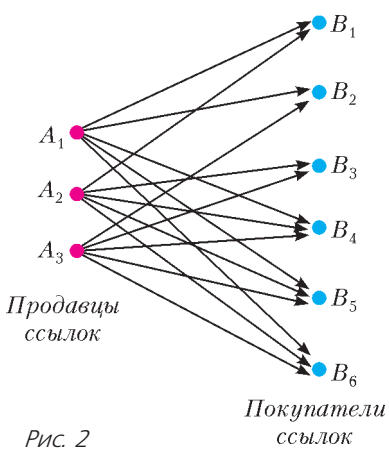

Рис. 2

Задача хорошей поисковой системы состоит в автоматическом выявлении недобросовестных владельцев. Почему автоматическом? А потому что тех же линковых колец *сотни тысяч* (!) и выловить их «руками» просто физически невозможно. Но как научить машину отличать кольцо от структуры, которая кольцом не является?

Представим себе, что у нас есть идеальная (или просто достаточно адекватная) модель интернета без спама. Подсчитаем в этой модели вероятность возникновения ребра между вершинами заданных входящих степеней. Иными словами, про вершины A и B известно, что $\mathrm{indeg}\ A = a$, $\mathrm{indeg}\ B = b$. И пусть, при условии этого знания, вероятность ребра из A в B равна $f(a,b)$ (соответственно, вероятность ребра из B в A равна $f(b,a)$).

Пусть теперь у нас есть кусок интернета, который, возможно, является кольцом. Обозначим A_1, ..., A_k – вершины левой доли, а B_1, ..., B_l – вершины правой доли. Величины входящих степеней обозначим, соответственно, a_1, ..., a_k, b_1, ..., b_l. Найдем $M = \sum_{i=1}^{k}\sum_{j=1}^{l} f(a_i,b_j)$. По сути, M – это ожидаемое число ребер, которые в «хорошем» интернете идут из левой доли нашей структуры в ее правую долю. Пусть, наконец, реальное количество ребер «слева направо» равно μ. И все понятно: остается лишь сравнить M и μ. Если ожидаемое число ребер меньше реального, то, наверное, структура аномальная.

Конечно, всегда есть вероятность ошибки. И цена вопроса может оказаться очень высокой. Поэтому обычно сайты в изловленных кольцах не напрямую понижают в выдаче; им лишь присваивают некоторую числовую характеристику, зависящую от M и μ, а машина аккуратно учитывает эту характеристику при ранжировании документов по запросу.

Список литературы

1. *А.М.Райгородский*. Модели случайных графов. – М: МЦНМО, 2011.

2. *B.Bollobás*. Random Graphs, Second Edition. – Cambridge Univ. Press, 2001.

3. *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 (2011), №2, p. 40–73.