Структура сложных сетей
Лекция № 4 курса
«Алгоритмы для Интернета»

Юрий Лифшиц*

19 октября 2006 г.

Содержание

1. Введение

1.1. Что такое сеть?

Сетью будем называть набор *вершин* и связей между ними — *ребер*. Для описания реальных сетей этих двух параметров может быть недостаточно. Подумаем, какие могут быть дополнительные характе-

*Законспектировал Иван Гунич.

ристики. Во-первых, ребра могут быть *ориентированными* или *неориентированными* (сети соответственно называются ориентированными или неориентированными). К примеру, сеть телефонных звонков или электронных сообщений должна иметь ориентированные ребра. *Двудольные сети* содержат вершины двух различных типов, а ребра соединяют только вершины различающихся типов. Такие сети также называют *сетями членства* (affiliation networks), в которых люди объединены в группы, и сеть состоит из вершин двух видов — людей и групп. Также ребра и вершины могут иметь различные числовые и качественные характеристики. К примеру, в сетях дружбы между людьми вершина может представлять либо мужчину, либо женщину. Или если в той же сети отношений мы хотим знать, насколько один человек хорошо знает другого. Ориентированные сети могут иметь циклы или быть *ациклическими*. Например, сеть цитирований в статьях ациклична, поскольку статья может ссылаться на другую статью лишь в том случае, когда она уже написана. Таким образом, все ребра в сети цитирований указывают на события, произошедшие раньше во времени, поэтому циклов быть не может (по крайней мере они появляются очень редко). Еще одной из основных характеристик является динамика сети, то есть процесс добавления и исчезновения новых вершин и ребер. Существуют и другие характеристики сетей.

1.2. Как изучать сложные сети?

Опишем существующие подходы к изучению сложных сетей в современной науке. Во-первых, необходимо собрать максимум статистической информации о свойствах интересующей нас сети (длины путей, степенные распределения), которые характеризуют структуру и поведение системы, и предложить подходящие способы для их измерения. Во-вторых, нужно разработать научный язык (терминологию) для описания свойств сетей и научиться говорить про свойства и параметры сетей из жизни научным языком.

Далее необходимо разработать математические модели, имеющие сходные с реальными сетями характеристики. Вы спросите, зачем это нужно? Дело в том, что очень дорого проводить эксперименты на практических сетях. Если же у нас есть упрощенная идеальная математическая модель, то над ней намного легче экспериментировать, и это не требует огромных финансовых затрат. Также такая модель дает возможность записывать некоторые свойства сетей и процессы, происходящие в сетях, в математических терминах. Основная цель изучения сетей — разработка алгоритмов для управления, оптимизации и предсказания процессов в сетях.

2. Сети вокруг нас

2.1. Социальные сети

Социальные сети — это сети социальных взаимоотношений между людьми. Вот несколько примеров социальных сетей.

1. Сеть дружбы (рис. 5).
2. Сеть соавторства ученых (рис. 7).

   Поль Эрдеш (Paul Erdös) — один из величайших математиков 20-го столетия — за всю свою жизнь был автором или соавтором 1486 научных работ. *Индекс Эрдеша* определяется следующим образом: Поль Эрдеш имеет индекс 0, его соавторы — индекс 1, а соавторы его соавторов — индекс 2 и т. д. Существует гипотеза, что почти у 98% математиков, писавших статьи в соавторстве, индекс Эрдеша будет не более 7. Например, у Юрия Лифшица индекс Эрдеша равен 3: Paul Erdös $\overset{1}{\to}$ Richard K. Guy $\overset{2}{\to}$ Юрий Матиясевич $\overset{3}{\to}$ Юрий Лифшиц.
3. Сеть сексуальных отношений (рис. 1, 6).
4. Браки между кланами.
5. Бизнес отношения.

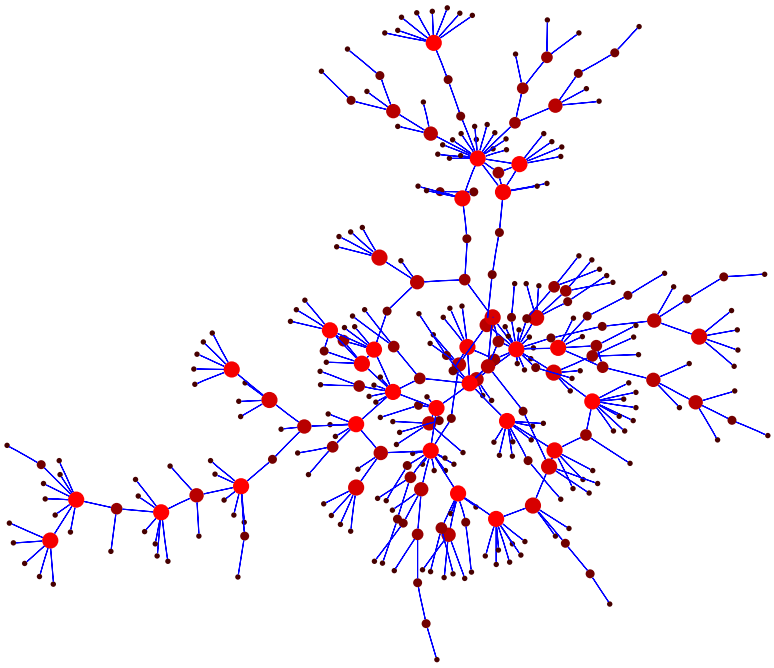

Рис. 1. Сеть сексуальных контактов [Potterat et al., 2002]. Наличие циклов из нечетного числа ребер говорит о существовании однополых связей.

6. Совместное появление киноактеров в фильмах.

   Есть известный американский сайт `www.imdb.com`, где можно найти абсолютно любой фильм, начиная от черно-белых триллеров Альфреда Хичкока и кончая третьесортными русскими боевиками. На базе этого сайта построена и тщательно изучена сеть совместного появления актеров в фильмах, то есть если один актер играл в одном фильме с другим актером, то между ними есть ребро в рассматриваемой сети.

7. Телефонные звонки, электронные письма, сеть icq-контактов.

2.2. Информационные сети

Информационные сети — это сети отношений между информационными объектами. Приведем примеры информационных сетей.

1. Цитирования в научных статьях.

   Это классический пример информационных сетей. Мы уже говорили о нем, когда вводили понятие ацикличности. Цитирования в научных статьях (citations) образуют граф, в котором вершины — это научные статьи, а ориентированные ребра означают, что одна статья ссылается на другую.

2. Граф ссылок WWW.

3. Цитирование в патентах.

4. Peer-to-peer сети.

   P2P сеть состоит из равноправных узлов, причем каждый из них взаимодействует лишь с некоторым подмножеством узлов сети, так как установление связи «каждый с каждым» невозможно из-за

ограниченности вычислительных ресурсов и пропускной способности. При этом передача информации между узлами, не связанными в данный момент непосредственно, может осуществляется как по своеобразной эстафете (от узла к узлу), так и путем установления временной прямой связи. Все вопросы маршрутизации сообщений, передаваемых по эстафете, лежат не на едином сервере, а на всех отдельных узлах. Такое определение также известно под названием Pure P2P.

5. Совместное употребление слов в текстах.

2.3. Технологические сети

Что же такое технологические сети? Можно сказать, что они показывают «физические» связи в нашем трехмерном мире. Вот несколько примеров.

1. Интернет как сеть компьютеров (рис. 2).

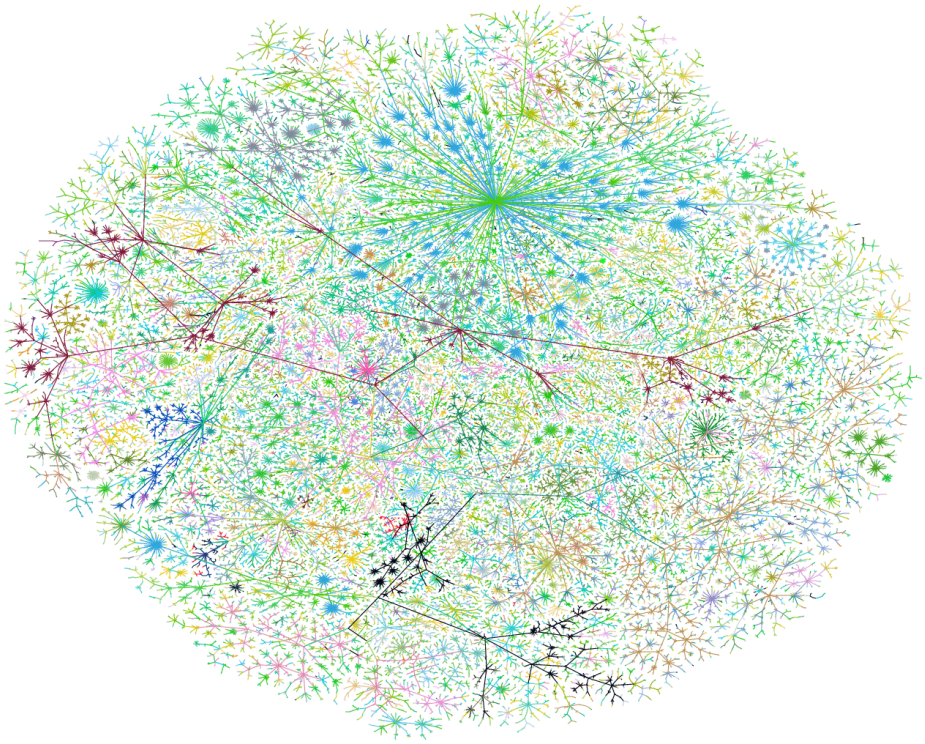

Рис. 2. Интернет как сеть компьютеров [Hal Burch, Bill Cheswick, Lumeta Corporation].

В данном случае мы имеем в виду кабели, соединяющие сервера, провайдеров и т. д.

2. Национальные электросети.
3. Телефонные линии, почтовые службы доставки.
4. Поезда, самолеты, автобусы.

2.4. Биологические сети

Под биологическими сетями будем понимать сети внутри и между животными, растениями, людьми.

1. Сети обменных процессов (metabolic networks).

   Данные сети состоят из веществ, вступающих в реакцию, и продуктов реакции. Два вещества соединяются ребром, если существует реакция между ними, в ходе которой образуется уже известное вещество.

2. Реакции между протеинами.

   Между двумя протеинами будет ребро в сети, если между ними возможна реакция.

3. Сеть нейронов в мозге.

4. Кровеносные сосуды.

5. Граф «хищник-жертва» (рис. 3)

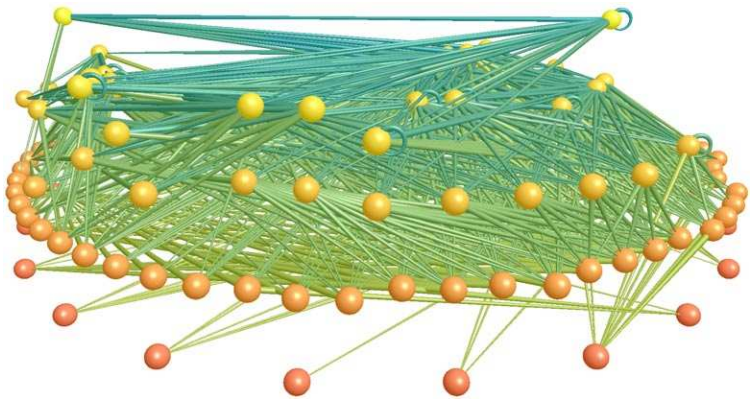

Рис. 3. Пищевая сеть «хищник-жертва» в пресной воде.

   Это очень интересная сеть, в которой вершина представляет собой какой-то конкретный вид животного. Между вершинами существует ребро, если один вид питается другим.

6. Реки, озера, океаны.

3. Вспоминаем теорию графов

3.1. Эффект «тесного мира»

Вы, наверное, слышали об известных экспериментах, проводимых в 60-х годах Стэнли Милгрэмом[1]. В первом эксперименте участвовало около 150 человек. После обработки полученных данных оказалось, что от первого отправителя биржевого брокера отделяла цепочка длины от двух до десяти звеньев. В среднем получилось 5 звеньев. «*Эффект тесного мира*» («*Six degrees of separation*») можно сформулировать так: «Двух случайно взятых людей отделяет друг от друга цепочка из не более чем 6 звеньев (или рукопожатий)».

[1] Добровольные участники эксперимента получали следующие инструкции:

- отправьте письмо заранее известному биржевому брокеру в Бостон;
- если вы вдруг с ним не знакомы, то отправьте письмо своему приятелю, который, как вам кажется, с ним знаком;
- в письме вы должны написать свое имя и адрес (это нужно, чтобы отследить всю цепочку и чтобы не образовались циклы);
- на отдельном листе напишите свои данные и данные тех, кто был в цепочке до вас (это нужно для того, чтобы отследить путь писем, которые по какой-либо причине не дойдут до адресата).

Вернемся опять к сайту `www.imdb.com`! Анализ его базы данных показал, что этот эффект применим и к голливудскому миру: как ни странно, цепочка от одного актера до другого состоит в среднем менее чем из 3-х звеньев. В свое время в Америке была даже такая игра «Kevin Bacon's game»[2]: каждый американец подсчитывал, сколько рукопожатий отделяют его от Кевина Бэкона. Например, Кевина Бэкона отделяют от Чарли Чаплина всего три звена: сам Бэкон играл с Лоуренсом Фишберном, который в свою очередь играл с Марлоном Брандо, который снимался с Чарли Чаплином.

Теперь подумаем, как выразить эффект тесного мира в числах. Введем такое понятие, как *среднее кратчайшее расстояние* и обозначим его l:

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geqslant j} d_{ij},$$

где d_{ij} — кратчайшее расстояние от вершины i до вершины j, измеренное в числе ребер. Мы включили в формулу расстояние от вершины до нее самой, поэтому мы делим на $\frac{1}{2}n(n+1)$, а не на $\frac{1}{2}n(n-1)$, как было бы в формуле для среднего арифметического по всем расстояниям.

Теперь представим ситуацию, когда в нашем графе больше одной компоненты связности. Тогда расстояние между некоторыми вершинами будет равно ∞. Что же делать в таких случаях? Конечно, можно взять среднее по всем связанным парам. Но есть альтернативный вариант — посчитать *среднее гармоническое кратчайшее расстояние*:

$$l^{-1} = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i > j} d_{ij}^{-1}.$$

Тогда слагаемые с бесконечным расстоянием d_{ij} вносят нулевой вклад в сумму.

3.2. Транзитивность

У этого понятия также есть другое название — *кластеризация*. Было замечено, что многие сети обладают следующим свойством: если вершина A соединена ребром с вершиной B, а вершина B соединена ребром с вершиной C, то высока вероятность того, что вершина A тоже соединена с вершиной C. Если говорить языком социальных сетей, то это свойство звучит так: «Друг моего друга — мой друг!» Другими словами, людей можно разделить на группы так, что будет много ребер внутри группы и мало ребер вне группы.

Транзитивность означает, что в сети присутствует значительное количество связанных троек (три вершины, связанные друг с друг с другом). Поэтому введем такое понятие, как *коэффициент кластеризации*, и обозначим его через C:

$$C = \frac{3 \times \text{число треугольников в сети}}{\text{число «вилок»}},$$

где под «вилкой» (triple) мы будем понимать вершину с двумя ребрами, идущими от этой вершины к двум другим. Также коэффициент кластеризации можно записать в виде:

$$C = \frac{6 \times \text{число треугольников в сети}}{\text{число путей длины 2}}.$$

Существует другой подход к вычислению коэффициента C. Посчитаем количество треугольников, в которые входит вершина, и количество «вилок» для этой вершины, а затем поделим одно на другое:

$$C_i = \frac{\text{число треугольников с вершиной } i}{\text{число «вилок», центром которых является вершина } i}.$$

Затем берем среднее арифметическое всех кластеризаций:

$$C' = \frac{1}{n} \sum_i C_i.$$

[2]Кевин Бэкон — известный голливудский актер.

Проиллюстрируем эти формулы на небольшом примере (рис. 4).

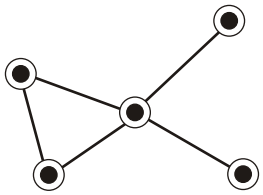

Рис. 4. Пример вычисления коэффициента кластеризации. Данная сеть состоит из одного треугольника и восьми «вилок», поэтому $C = 3 \times \frac{1}{8} = \frac{3}{8}$. Индивидуальные вершины имеют свой собственный коэффициент кластеризации C_i: 1, 1, $\frac{1}{6}$, 0 и 0. Отсюда получаем, что $C' = \frac{13}{30}$.

3.3. Распределение степеней

Это одна из важнейших характеристик графа. В последние годы ей уделяют очень много внимания. Степенное распределение резко отделяет практические сети, о которых мы говорили ранее, от случайных графов. Иначе говоря, если взять случайный граф, то в нем распределение степеней будет сильно отличаться от распределения степеней реальных графов. Определим *степенное распределение* p_k следующим образом: p_k — это доля вершин в сети, имеющих степень k. Также p_k — это вероятность того, что случайно выбранная вершина имеет степень k.

Если представить распределение степеней p_k как функцию от величины k, то главный недостаток будет заключаться в том, что эта функция негладкая, особенно для больших чисел. Рассмотрим сеть ссылок в Интернете. Возможна ситуация, когда не существует сайтов с 1000 ссылок, нет сайтов с 1001 и 1002 ссылками, но существует, к примеру, 5 сайтов с 1003 ссылками. Этот пример показывает, что наша функция «скачет». Как же добиться гладкости функции распределения степеней?

Мы можем усреднять p_k по возрастающим интервалам. Например, для каждого $2^t \leqslant k < 2^{t+1}$ определим:

$$p'_k = \frac{1}{2^t} \sum_{j=2^t}^{2^{t+1}-1} p_j.$$

Если мы хотим проверить предположение, то есть мы теоретически на модели вывели какую-либо функцию и хотим проверить ее на практике, то лучше использовать формулу для p'_k.

При рассмотрении моделей сетей выделяют два степенных распределения. Первое из них — это *пуассоновское распределение*:

$$p_k = \frac{z^k \, e^{-z}}{k!},$$

где z — это некоторая константа, а k — степень. Напомним, что для любого распределения выполнено свойство $\sum_k p_k = 1$. Фактически, данное распределение имеет вид:

$$p_k = \frac{c_1^k \, c_2}{k!},$$

где c_1 и c_2 — некоторые константы. Эта функция убывает экспоненциально быстро. При $k = 1000$ p_k будет очень маленькой величиной (с огромной скоростью убывает вероятность иметь 1000 и более ссылок), что не соответствует действительности. Например, в сети электронных сообщений мы сможем найти спамеров, у которых степень будет гораздо больше, чем у обычных вершин этой сети.

Второй важный закон — это распределение по *степенному закону* (*power-law*):

$$p_k = \frac{k^{-\alpha}}{\zeta(\alpha)},$$

где α — константа, $\zeta(\alpha)$ — дзета-функция Римана, которая служит для того, чтобы выполнялось равенство $\sum_k p_k = 1$. Степенному закону подчиняются эмпирические распределения степеней во многих реальных сетях. В частности, для Интернета $\alpha = 2.5$, а для сети голливудских актеров $\alpha = 2.3$. В таблице 1 приведены коэффициенты α для некоторых сетей.

	Сеть	Тип	n	m	l	α	C	C'
Социальные	Голливудские актеры	Неориент.	449 913	25 516 482	3.48	2.3	0.20	0.78
	Директора компаний	Неориент.	7 673	55 392	4.60	–	0.59	0.88
	Соавторство в математике	Неориент.	253 339	496 489	7.57	–	0.15	0.34
	Соавторство в физике	Неориент.	52 909	245 300	6.19	–	0.45	0.56
	Соавторство в биологии	Неориент.	1 520 251	11 803 054	4.92	–	0.088	0.60
	Сеть телефонных звонков	Неориент.	47 000 000	80 000 000		2.1		
	Сеть электронных сообщений	Ориент.	59 912	86 300	4.95	1.5/2.0		0.16
	Отношения между студентами	Неориент.	573	477	16.01	–	0.005	0.001
	Сексуальные контакты	Неориент.	2 810				3.2	
Информационные	WWW nd.edu	Ориент.	269 504	1 497 135	11.27	2.1/2.4	0.11	0.29
	WWW Altavista	Ориент.	203 549 046	2 130 000 000	16.18	2.1/2.7		
	Цитирование в статьях	Ориент.	783 339	6 716 198		3.0/–		
	Словари Роджета	Ориент.	1 022	5 103	4.87	–	0.13	0.15
	Совместное употребление слов	Неориент.	460 902	17 000 000		2.7		0.44
Технологические	Интернет (физический)	Неориент.	10 697	31 992	3.31	2.5	0.035	0.39
	Электрическая сеть	Неориент.	4 941	6 594	18.99	–	0.10	0.080
	Маршруты поездов	Неориент.	587	19 603	2.16	–		0.69
	Комплекты ПО	Ориент.	1 439	1 723	2.42	1.6/1.4	0.070	0.082
	Электронные цепи	Неориент.	24 097	53 248	11.05	3.0	0.010	0.030
	P2P сети	Неориент.	880	1 296	4.28	2.1	0.012	0.011
Биологические	Обменные сети	Неориент.	765	3 686	2.56	2.2	0.090	0.67
	Реакции между протеинами	Неориент.	2 115	2 240	6.80	2.4	0.072	0.071
	Пищевая сеть (морские виды)	Ориент.	135	598	2.05	–	0.16	0.23
	Пищевая сеть (пресноводные)	Ориент.	92	997	1.90	–	0.20	0.087
	Нейронные сети	Ориент.	307	2 359	3.97	–	0.18	0.28

Таблица 1. Характеристики некоторых сетей: тип сети (ориентированная или неориентированная сеть); суммарное число вершин n; суммарное число ребер m; среднее кратчайшее расстояние l; коэффициент α из формулы для степенного закона («−» означает, что в данной сети этот закон не действует; для ориентированных сетей приводятся распределения входящих и исходящих степеней); коэффициенты кластеризации C и C' [1].

Для других типов графов степенное распределение может иметь более сложный вид. К примеру, для двудольных графов можно определить два степенных распределения, по одному для отдельного типа вершин. В ориентированных графах у каждой вершины есть своя входящая и исходящая степень, поэтому вводится функция распределения p_{jk}.

3.4. Корреляции

Пусть в сети есть вершины разных типов. В сети наблюдается эффект *assortative mixing*[3], если ребра «чаще» соединяют вершины внутри одного типа, чем между разными типами.

[3]Можно перевести как *эффект предпочтительного связывания*.

Пусть e_{ij} — доля ребер между типами i и j в множестве всех ребер сети ($i = \overline{1,N}$, $j = \overline{1,N}$, N — число вершин в сети). Обозначим через $P(j|i)$ условную вероятность того, что мой друг представляет класс j при условии, что я сам представляю класс i:

$$P(j|i) = \frac{e_{ij}}{\sum_k e_{ik}}.$$

Тогда assortative mixing можно измерить коэффициентом Q:

$$Q = \frac{[\sum_i P(i|i)] - 1}{N - 1}.$$

Вы спросите, зачем мы делим на $N - 1$? Это необходимо для нормировки коэффициента: $0 \leqslant Q \leqslant 1$. Значения $Q = 0$ и $Q = 1$ в данном случае будут иметь как раз правильный смысл. Давайте в этом убедимся. Какие бывают два крайних случая?

Первый случай: все дружат только с представителями своего класса. Тогда каждая вероятность $P(i|i)$ будет равна 1. Следовательно:

$$Q = \frac{N - 1}{N - 1} = 1.$$

Второй случай: «дружба» совершенно не зависит в теоретико-вероятностном смысле от того, какого класса мой «приятель». Это значит, что в нашей сети доли исходящих ребер, ведущих в свой класс и во все остальные, будут одинаковы. Получаем:

$$Q = \frac{N\frac{1}{N} - 1}{N - 1} = \frac{1 - 1}{N - 1} = 0.$$

В таблице 2 приведен классический пример assortative mixing для социальной сети расового смешивания.

		Женщины			
		Чернокожие	Испанцы	Белые	Другие
Мужчины	Чернокожие	506	32	69	26
	Испанцы	23	308	114	38
	Белые	26	46	599	68
	Другие	10	14	47	32

Таблица 2. Расовая принадлежность супругов, изученная на примере пар из Сан-Франциско, штат Калифорния [Catania et al., 1992]. Один из недостатков формулы для вычисления Q состоит в том, что можно вычислить два различных значения коэффициента Q в зависимости от того, мужчинам или женщинам соответствуют индексы i и j в формуле для e_{ij}. Величина одного из значений коэффициента assortative mixing: $Q = \left(\left[\frac{506}{565} + \frac{308}{400} + \frac{599}{829} + \frac{32}{164}\right] - 1\right) / (4 - 1) = 0.53$. Неясно, какое из двух значений является «корректным» для рассматриваемой сети. Поэтому была введена другая формула для коэффициента assortative mixing, которая не обладает данным недостатком [1].

Перейдем к вопросу о том, как определять наличие assortativity для степеней. Во-первых, можно построить график «средняя степень друзей у вершин степени k», а затем посмотреть на убывание/возрастание (кривая возрастает с увеличением k в случае assortativity). Во-вторых, можно сосчитать коэффициент Пирсона корреляции степеней на концах всех ребер.

3.5. Другие свойства

Вершина в графе может обладать свойством (*betweenness*). Оно показывает, на сколько часто данная вершина лежит на кратчайших путях между другими вершинами. Например, в графе перелетов большим значением betweenness будут обладать крупные международные аэропорты.

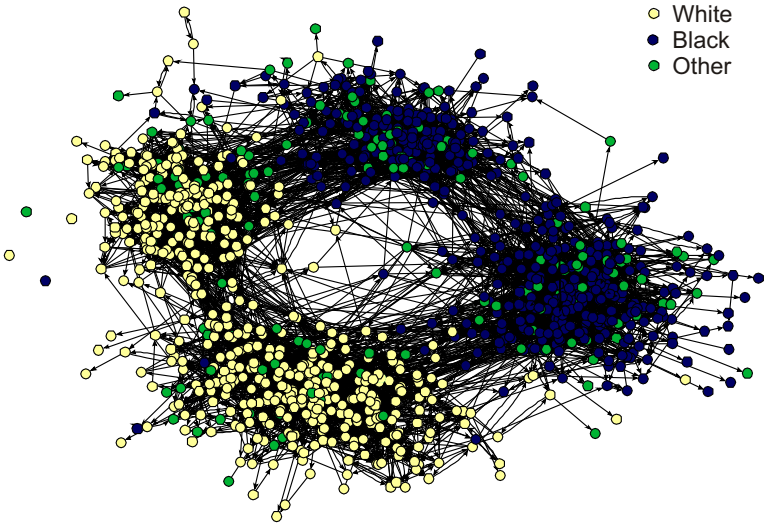

Рис. 5. Сеть дружбы в отдельно взятой американской школе как пример «структуры сообщества» социальной сети [James Moody]. При построении сети добавлялось ориентированное ребро из A в B, если школьник A говорил, что B — его друг (но не наоборот). Вершины в графе раскрашены в соответствии с расовой принадлежностью школьника. Явное разделение между верхней и нижней частью сети связано с тем, что в школе есть ученики младших и старших классов, которые мало общаются между собой.

Существует несколько вариантов определения свойства betweenness. Свойство, которое мы определили, обычно называют *промежуточностью по кратчайшим путям* (*shortest-path betweenness*).

Для случайной вершины i *промежуточность случайного блуждания* (*random-walk betweenness*) (рис. 6) равна усредненному по всем вершинам s и t количеству случайных путей из вершины s в t, проходящих через вершину i [2].

В некоторых графах интересно изучать уязвимость связности сети при выкидывании нескольких вершин (рис. 7). К примеру, для «физического» Интернета важно, насколько быстро теряется связность в случае выхода из строя некоторых вершин.

Также очень интересная характеристика — размер наибольшей компоненты, поскольку есть много сетей, в которых эта большая компонента должна занимать почти все пространство. К примеру, в графах цитирования и ссылок Интернета есть огромная связанная компонента и маленькие группы изолированных вершин.

4. Математические модели сетей

4.1. Случайные пуассоновские графы

Solomonoff и Rapoport, а также независимо от них Erdös и Rènyi, предложили следующую простую модель:

– фиксируем параметр p;

– фиксируем число n вершин графа;

– независимо для каждой пары вершин с вероятностью p проводим ребро между ними.

Эту модель Эрдеш и Рении назвали $G_{n,p}$.

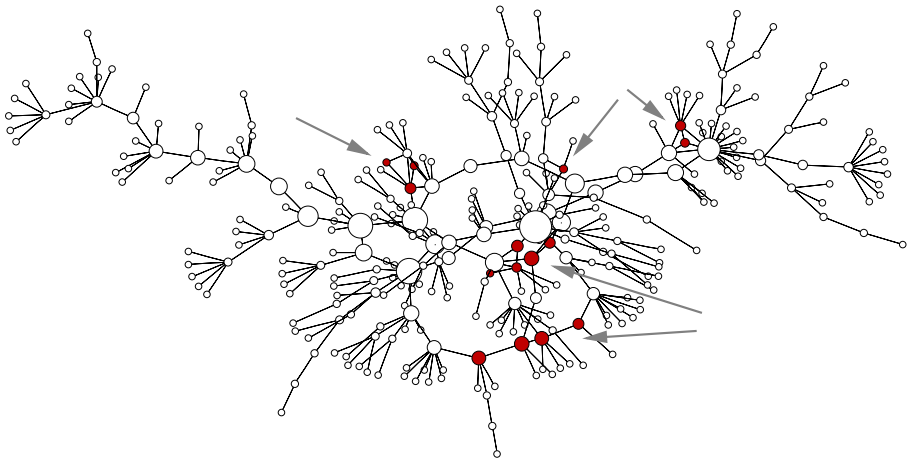

Рис. 6. Самая большая компонента сети сексуальных контактов между актерами в Колорадо Спрингз. Размеры вершин пропорциональны значению random-walk betweenness соответствующего актера. Красные вершины, на которые указывают стрелки, показывают, что для этого актера значение random-walk betweenness значительно больше, чем его shortest-path betweenness.

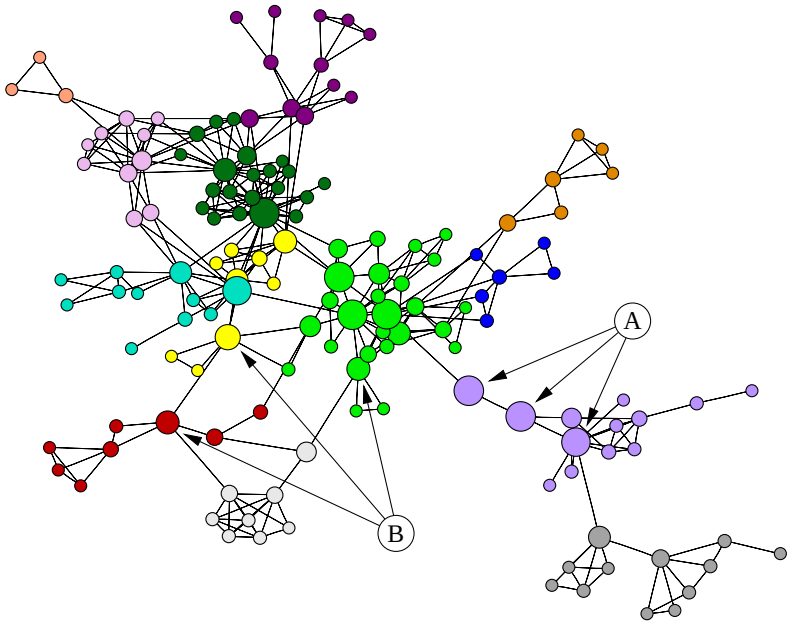

Рис. 7. Самая большая компонента сети соавторства ученых, изучающих сложные сети. Размеры вершин пропорциональны значению random-walk betweenness. Символы «A» и «B» указывают на вершины, при удалении которых сеть теряет связность.

Ими также была предложена другая похожая модель $G_{n,m}$:

– фиксируем общее число ребер m;

– берем случайный граф с n вершинами и m ребрами (каждый возможный граф может появиться с одинаковой вероятностью).

4.2. Конфигурационная модель

Как уже отмечалось, случайные графы не обладают многими свойствами реальных сетей. *Конфигурационная модель* определяется следующим образом:

– фиксируем распределение степеней p_k (напомним, что p_k — это доля вершин, имеющих степень k);

– выбираем n чисел k_i согласно распределению ($i = \overline{1,n}$, n — число вершин в будущей сети);

– у каждой вершины i в нашем графе нарисуем k_i «хвостов» (заготовок для будущих ребер);

– случайно выбираем пары «хвостов» и соединяем их ребром.

Таким образом, этот процесс с равной вероятностью генерирует любую возможную конфигурацию сети с заданным распределением степеней вершин.

4.3. Улучшения конфигурационной модели

1. Можно построить ориентированный граф с заданным распределением входящих и исходящих ребер.

   Каждая вершина в таком графе имеет как входящую (j), так и исходящую (k) степень, и распределение становится «двойным» (p_{jk}). К примеру, для Интернета эти два распределения будут различаться: на каждом сайте ограниченное число исходящих ссылок (страница имеет какой-то разумный размер), а входящих ссылок может быть огромное количество. Также можно привести пример с цитированиями в статьях: Марк Ньюмэн процитировал 429 статей [1], но почти никакая статья не будет цитировать более 1000 статей. При этом существует много научных работ, которые были процитированы более чем 1000 авторами.

   При построении такого ориентированного графа необходимо проверять, что математические ожидания входящих заготовок равняются математическим ожиданиям исходящих заготовок. Это нужно для того, чтобы в результате не получился граф, у которого входящих «хвостов» будет намного больше, чем исходящих.

2. Можно порождать конфигурационные модели двудольных графов.

   Для двудольного графа также задаются два степенных распределения. В данном случае «хвосты» одной группы вершин соединяются с «хвостами» другой группы вершин.

3. Марковские случайные графы. Мы начинаем с какого-нибудь случайного графа. Потом мы начинаем блуждать по «графу графов» с вероятностями, пропорциональными тем свойствам, которые мы хотим получить. К примеру, пусть у нас есть измеренные характеристики реальной сети, которые должны присутствовать в нашей модели, такие, как число ребер, степенной закон, кластеризация или assortative mixing. Тогда чем сильнее соседний граф будет соответствовать заданным свойствам, тем больше будет вероятность, что мы к нему перейдем.

4.4. Модель «тесного мира»

Уотс и Строгатс (Watts and Strogatz) предложили следующую модель.

1. Возьмем n вершин и расположим их по кругу.

2. Затем соединяем каждую вершину со всеми, находящимися на расстоянии не более чем k. Получается сеть из n вершин и nk ребер (рис. 8(a)).

3. Каждое ребро мы с некоторой вероятностью p переадресовываем: один из его концов заменяем на случайную вершину (рис. 8(b)). При этом должно соблюдаться ограничение, согласно которому двойные ребра и петли не должны появиться.

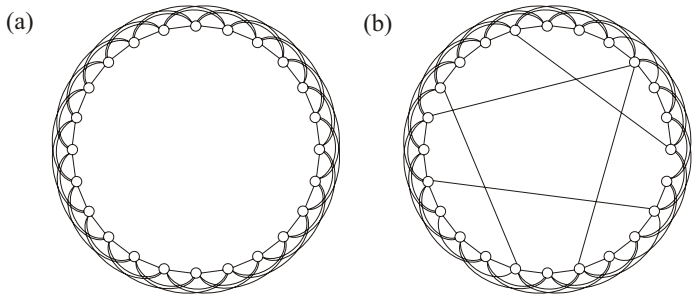

Рис. 8. (a) Соединяем все вершины, находящиеся на расстоянии $k = 3$. (b) С вероятностью p один из концов каждого ребра заменяем на случайную вершину.

Также существует другая модель, предложенная Монассоном (Monasson), Ньюмэном и Уотсом. В этой модели вместо переадресации мы просто добавляем еще одно ребро. Данная модель обладает тем свойством, что ни одна вершина никогда не будет отделена от остальной части сети, следовательно, расстояние между любыми вершинами всегда будет конечной величиной.

4.5. Модель Прайса

Эта модель считается наиболее известной. Она основана на разумном принципе: «Богатый становится богаче», — введенном в 50-х годах Гербертом Саймоном. Сейчас этот эффект называется *предпочтительностью присоединения* (preferential attachment)[4]. Основным вкладом Прайса являлось то, что он взял идеи Саймона и перенес на сети, в частности, на сети цитирования в научных статьях.

Рассмотрим модель, которую предложил Прайс. Пусть есть ориентированный граф с n вершинами (к примеру, сеть цитирования), p_k — доля вершин с входящей степенью k ($\sum_k p_k = 1$). Мы создаем новую вершину (к примеру, первая статья написана). Затем закрепляем за ней какое-то конкретное число исходящих ребер — количество статей, на которые она ссылается. Количество исходящих ребер может различаться для каждой вершины, поэтому введем параметр m, который задает среднее число исходящих ребер добавленных ранее вершин. Также это число задает среднее число входящих ребер: $\sum_k kp_k = m$, поскольку каждое ребро имеет два конца. Величина m может меняться от вершины к вершине, следовательно, она может принимать дробные значения, не превосходящие 1.

Вероятность присоединения одного из наших новых ребер к старой вершине (вероятность того, что созданная статья ссылается на старую статью) пропорциональна входящей степени k старых вершин. Сразу же появляется проблема: если вершина рождается с нулевой входящей степенью, то она всегда будет иметь равную нулю вероятность присоединения новых вершин. Для решения этой проблемы Прайс предложил сделать так, чтобы вероятность присоединения вершины была пропорциональна $k + 1$ (то есть при «рождении» вершины одно цитирование ей всегда гарантируется). Вероятность стать адресатом

[4]В социологии этот эффект называется *эффектом Матвея* в силу библейского изречения: «Ибо всякому имеющему дастся и приумножится...». Новый завет от Матвея, 25:29.

нового ребра для вершины степени k равна:

$$\frac{(k+1)p_k}{\sum_i (i+1)p_i} = \frac{(k+1)p_k}{m+1}.$$

Эта формула равносильна:

$$\frac{k+1}{\sum_i (i+1)p_i} = \frac{k+1}{m+1}.$$

Вероятность процитировать статью, которая уже имеет k цитирований пропорциональна $(k+1)/(m+1)$.

4.6. Расширение базовых моделей роста

В модель Прайса можно вносить следующие изменения.

1. Строить неориентированный граф. Вероятность стать адресатом ребра будет пропорциональна полной степени вершины.

2. Изменить добавочный коэффициент: сделать вероятности присоединения к вершинам степени k пропорциональными $k + k_0$, где k_0 — некоторая константа.

3. Вы наверное заметили, что в реальном мире, когда мы хотим сослаться на какую-либо статью, то мы руководствуемся не только количеством уже сославшихся на нее авторов, но и содержанием данной статьи. Поэтому было предложено при «рождении» вершины ввести случайный коэффициент «привлекательности» η_i этой вершины. Тогда вероятности проведения ребра будут пропорциональны $\eta_i k_i$ или $\eta_i + k_i$.

4. Вероятность присоединения может быть не линейной относительно k, а пропорциональной k^γ, где γ — константа.

Источники

[1] M. E. J. Newman. The structure and function of complex networks
http://www.santafe.edu/files/gems/paleofoodwebs/Newman2003SIAM.pdf

[2] M. E. J. Newman. A measure of betweenness centrality based on random walks
http://aps.arxiv.org/pdf/cond-mat/0309045.pdf

[3] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure of the Web
http://www.people.cornell.edu/pages/dc288/Paper1.pdf

[4] Страница курса
http://logic.pdmi.ras.ru/~yura/internet.html