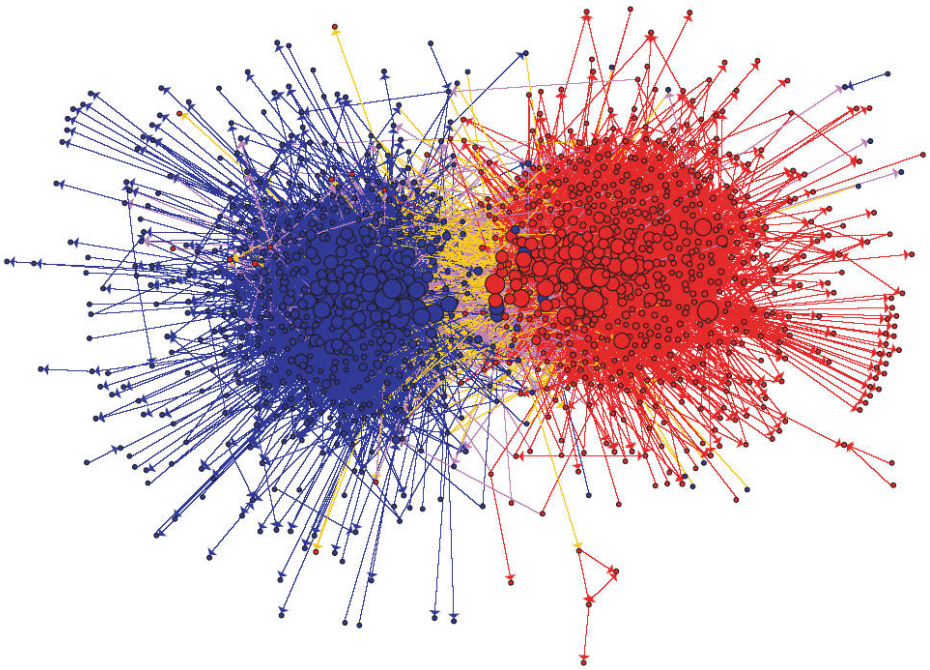
УНИВЕРСИТЕТ ИТМО

Н.Ф. Гусарова

АНАЛИЗ СОЦИАЛЬНЫХ СЕТЕЙ. ОСНОВНЫЕ ПОНЯТИЯ И МЕТРИКИ



Санкт-Петербург
2016

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

Н.Ф. Гусарова

АНАЛИЗ СОЦИАЛЬНЫХ СЕТЕЙ. ОСНОВНЫЕ ПОНЯТИЯ И МЕТРИКИ

Учебное пособие

Санкт-Петербург

2016

Гусарова Н.Ф. **Анализ социальных сетей. Основные понятия и метрики.** – СПб: Университет ИТМО, 2016. – 67 с.

Пособие охватывает теоретический материал, необходимый для усвоения программы курса «Анализ социальных сетей. Основные понятия и метрики». Рассматриваются вопросы, связанные с базовыми понятиями в анализе социальных сетей, такими как комплексные сети, степенные законы распределения, случайные графы, малый мир и т.д. Особое внимание уделяется метрикам, используемым при анализе социальных сетей, и их содержательной интерпретации.

Пособие адресовано студентам, обучающимся по направлениям подготовки 09.04.02 – Информационные системы управления и технологии, 45.04.04 – Интеллектуальные системы в гуманитарной среде.

Рекомендовано к печати Ученым советом факультета инфокоммуникационных технологий, протокол № 18/06 от 17.10. 2016.

УНИВЕРСИТЕТ ИТМО

Университет ИТМО – ведущий вуз России в области информационных и фотонных технологий, один из немногих российских вузов, получивших в 2009 году статус национального исследовательского университета. С 2013 года Университет ИТМО – участник программы повышения конкурентоспособности российских университетов среди ведущих мировых научно-образовательных центров, известной как проект «5 в 100». Цель Университета ИТМО – становление исследовательского университета мирового уровня, предпринимательского по типу, ориентированного на интернационализацию всех направлений деятельности.

Оглавление

Введение

Социальные сети являются примером сложных сетевых систем. При изучении социальных сетей рассматриваются три основных вопроса:

- анализ структуры сетей,
- формирование сети,
- процессы на сетях (например, распространение слухов, вирусов и пр.).

Изучение социальных сетей как научное направление возникло на стыке ряд научных дисциплин – социология, дискретная математика, Computer Science (алгоритмы на графах и сетях). В последнее время к этому списку добавились статистическая физика, где также были выявлены объекты типа социальных сетей, а также экономика, из которой пришли подходы независимых агентов, теории игр и др.

В силу неустоявшегося характера научного направления в ней присутствует сборная терминология:
network = graph = граф = сеть;
nodes = vertices, actors = узлы графа, акторы (социология, экономика);
links = edges, relations = связи в графе (relations – в экономике);
clusters = communities = сообщества = группы узлов, которые теснее связаны между собой, чем с остальным графом.

Рассмотрим пример социальной сети (рис. 1). На рисунке легко выделяются 4 сообщества.

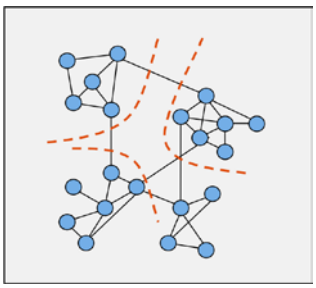


Рис. 1. Пример социальной сети

Что можно посчитать на этом графе, используя подход социальных сетей?

- количество узлов и ребер,
- средняя плотность,
- связность,
- наличие сообществ,

- принадлежность конкретного узла к конкретному сообществу,
- относительная роль каждого из узлов и/или ребер в графе.

В чем специфика социальных сетей как сложных сетевых образований (Complex networks)? Они, с одной стороны, не являются регулярными, но в то же время не могут рассматриваться как чисто случайные – в этом их сложность, что хорошо видно видно на рис. 2.

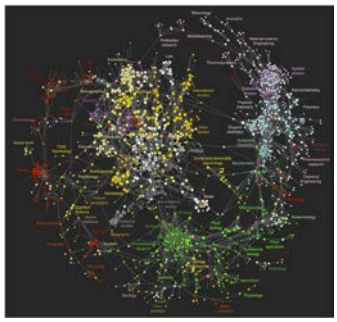


Рис. 2. Социальная сеть как сложное сетевое образование

Как известно, абсолютно случайные сети изучает статистическая физика, абсолютно регулярные (регулярные решетки) изучает математика. Социальные сети занимают промежуточное положение:

- у них присутствует некоторая топология, но она не тривиальна;
- они имеют до некоторой степени фрактальные свойства, т.е. проявляют самоподобие на разных масштабах (scale-free networks)
- они обладают некоторыми универсальными свойствами, независимо от типа – биологические сети, социальные сети, сети «покупатели–товары» как двудольный граф, рынок акций как «сетевой клубок», где узлами являются акции, а степень их скоррелированности отображается ребром с весом.

Нетривиальная задача – так построить изображение больших сетей, чтобы явно показать на нем необходимые характеристики сети. Удачные примеры изображений представлены на рис. 3–7.

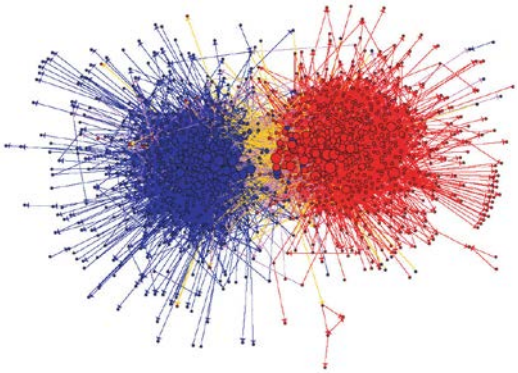


Рис. 3. Сеть связи политических партий: слева и справа – политические блоки, посередине (серым цветом) – журналисты, которые их связывают

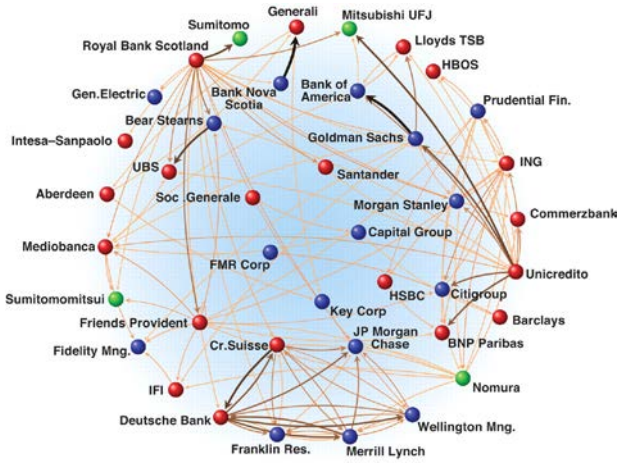


Рис. 4. Сеть заимствований банков друг у друга. Если один из банков «падает», то можно посчитать, как это отразится на других

Рис. 5. Транспортная карта Цюриха. Явно выделяются центральные узлы, которые выполняют функцию хабов в Интернете (связывают «всех со всеми»)

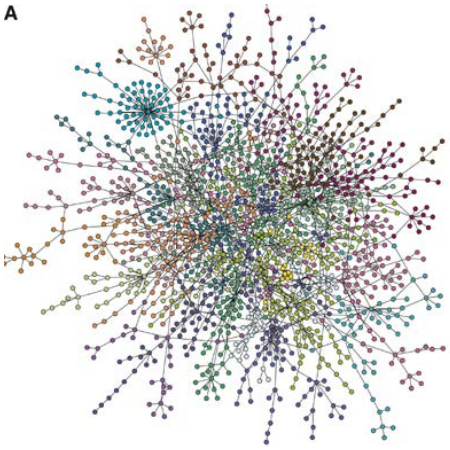


Рис. 6. Взаимодействие белковых молекул. Каждый узел – это отдельный белок. Видно, какие связи нужно порвать, чтобы целые фрагменты перестали взаимодействовать с остальными

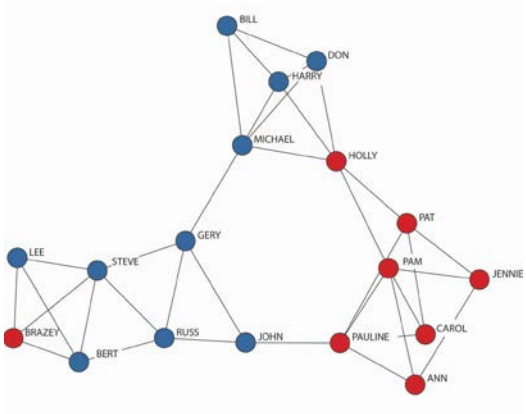


Рис. 7. Социальная сеть организации. Явно выделены сотрудники – связующие звенья, которые имеют уникальную позицию в организации

1. КОМПЛЕКСНЫЕ СЕТИ

1.1. ОСНОВНЫЕ ПОНЯТИЯ ИЗ ТЕОРИИ ГРАФОВ. МЕТРИЧЕСКИЕ СВОЙСТВА ГРАФА

Граф $G(V; E)$ – упорядоченное множество, состоящее из узлов V (vertices) и ребер E (edges). Число узлов и ребер обозначается соответственно как

$m = |V|$, $n = |E|$.

Путь – последовательность ребер, соединяющая последовательность узлов.

Выделяются следующие виды графов: направленные (ориентированные) и ненаправленные (неориентированные), простые (нет циклов и кратных ребер), взвешенные, двудольные, связные.

Маршрут в неориентированном графе – любая последовательность смежных ребер.

Цепь – маршрут, все ребра которого различны. В орграфе цепь называется *путем*.

Цикл – замкнутая цепь. В орграфе цикл называется *контуром*.

Граф называется *связным*, если любые две его вершины связаны маршрутом.

Компонента связности графа $G = (X, U)$ – его подграф $G' = (X', U')$, образованный на подмножестве всех вершин X', которые можно соединить произвольным маршрутом.

Связный граф состоит из единственной компоненты связности.

Бикомпонент (компонент сильной связности) – это максимальный по включению сильно связанный подграф графа.

Расстояние между вершинами графа – длина кратчайшей цепи, соединяющей эти вершины. Длина цепи – это количество входящих в нее ребер.

Если известны расстояния между каждой парой вершин графа $(x_i, x_j) \in X$, то можно определить его *диаметр* $d(G)$ как максимальное расстояние между двумя вершинами:

$d(G) = \max d(x_i, x_j)$.

Центр графа – это вершина x_0, для которой выполняется следующее условие:

$\forall x \subseteq X \, [\max d(x_i, x_j) \geq \max d(x_0, x_j)]$,

а величина $d(x_0, x_j)$ определяет радиус графа.

Эксцентриситет вершины графа − это максимальное расстояние от нее до других вершин (по количеству ребер или их весу). Тогда *диаметр графа* − это максимальный из эксцентриситетов вершин, а радиус − минимальный из эксцентриситетов. Центральные вершины графа − это такие,

эксцентриситет которых равен радиусу графа, а периферийные − такие, эксцентриситет которых равен диаметру графа.

Метрика диаметра графа чувствительна к наличию цепочек. Поэтому более показательной является метрика среднего кратчайшего пути.

Геодезическое расстояние – кратчайший путь между двумя узлами.

В графе типа «цепочка» среднее расстояние большое, а в графе типа «звезда» среднее расстояние всегда равно 2.

Степень узла k_i – число ближайших соседей.

Распределение степеней узла $P(k)$ – доля узлов со степенью k – описывает вероятность того, что у случайно выбранного узла ровно k соседей.

1.2. ФУНДАМЕНТАЛЬНЫЕ СВОЙСТВА КОМПЛЕКСНЫХ СЕТЕЙ

К фундаментальным относятся три свойства комплексных сетей.

1. Распределение степеней узлов, т.е. доля узлов, которые имеют данное количество соседей, является распределением с «длинным хвостом» и моделируется степенными распределениями.

Содержательно это означает, что в таких сетях много узлов, имеющих 1–3 соседа, но мало узлов, у которых тысячи соседей.

Пример – распределение сайтов в Интернете.

2. Сложные сети имеют очень небольшой диаметр (модель «маленького мира»).

Пример – известный феномен «6 рукопожатий» как среднее расстояние между незнакомыми людьми.

3. Локальная плотность (кластерный коффициент): если у вас есть в сети два друга, то велика вероятность, что они знают друг друга, т.е. в сети часты треугольники (плотные локальные структуры, dense local structure). Другими словами, в сложной сети велик коэффициент транзитивности (high transitivity/clustering coefficient)

Кроме того, комплексные сети имеют дополнительные свойства:

4. имеется гигантская связная компонента (giant connected component), т.е. больше 80% узлов связаны между собой;

5. имеются иерархии;

6. помимо треугольников, возникают более сложные кластерные образования (local community (cluster) structure).

Рассмотрим фундаментальные свойства более детально.

Первое свойство известно как распределение с длиным хвостом, Парето-принцип, принцип (80:20) и пр. Оно выражает степенной закон распределения (рис. 8).

Пример – книжный магазин. Есть бестселлеры, которые покупают все, но их мало; также есть длинный хвост, где мало покупателей, но много названий книг. Причем суммарная продажа всех популярных объектов – 20%, а длинного хвоста – 80%. Отсюда понятна популярность Интернет-

магазинов: они хранят все, т.е. могут делать деньги на «длинном хвосте», а обычные магазины делают свои продажи только на бестселлерах.

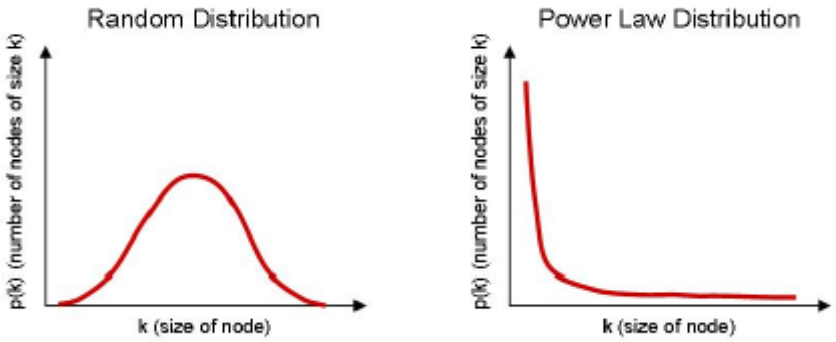


Рис. 8. Сравнение нормального и степенного распределения

Второе свойство – расстояние в 6 рукопожатий – иллюстрируется рис. 9. Как видно из графа, два любых актера связаны друг с другом не более чем через 6 промежуточных пар. Эта же закономерность была продемонстрирована на экспериментах «расстояние до Ландау» (через сколько соавторов данный автор связан с Ландау), а также на знаменитом эксперименте Милгрема (1967 год).

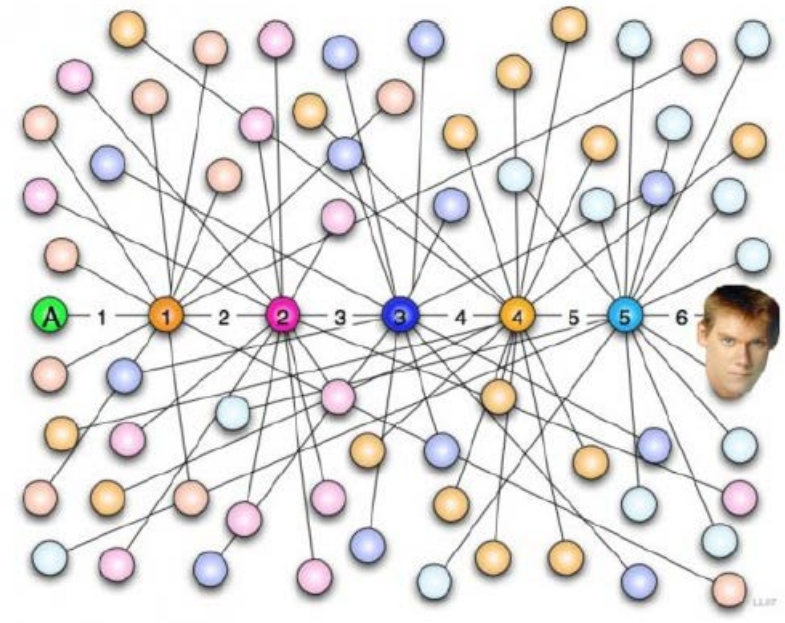


Рис. 9. Граф актеров. Связь означает, что два актера снимаются в одном фильме

Эксперимент Милгрема заключался в следующем.

Набрали волонтеров, им раздали письма, на письме было указано имя получателя, город, профессия, место рождения. Задача – постараться передать это письмо указанному человеку, но не напрямую, а через знако-

мых. Адресат был в Бостоне. Использованы два сценария: письмо было отдано волонтерам или непосредственно в Бостоне, или в Небраске, т.е. за 1200 миль от Бостона. Всего было задействовано 296 волонтеров – 196 в Небраске и 100 в Бостоне.

Эксперимент дал следующие результаты. Из отправленных писем достигли цели N = 64,29%, при этом средняя длина цепи составила $\langle L \rangle$ =5,2, причем из Бостона $\langle L \rangle$ = 4,4, а из Небраски $\langle L \rangle$=5,7. На рис. 10 видно, что распределение числа посредников при передаче имеет два пика – на 4 и 6 посредниках. Отметим недостаток эксперимента: в нем не учитывались письма, которые вообще не дошли (поэтому 6 – неточная цифра).

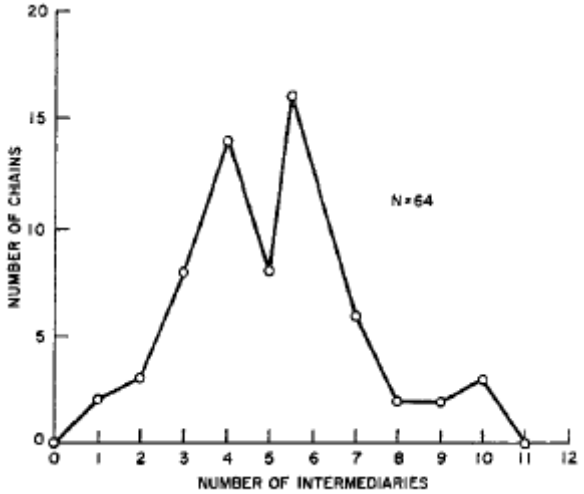


Рис. 10. Распределение числа посредников в эксперименте Милгрема

Неоднократно проводились повторы этого эксперимента, уже в Интернете, которые дали аналогичные результаты:

- Email graph: D. Watts (2001), 48,000 senders, $< L >= 6$
- MSN Messenger graph: J. Lescovec et al (2007), 240mln users, $< L >=$ 6,6
- Facebook graph: L. Backstrom et al (2012), 721 mln users, $< L >= 4{,}74$

Средняя величина – 5–6 шагов – указывает на то, что граф должен иметь некоторую структуру – не линейную и не решетчатую. На рис.11 представлен пример такой структуры – дерево Кейли или решетка Бете.

Решетка Бете – граф, топология которого полностью характеризуется координационным числом (т.е. числом ближайших соседей) Z. Дерево Кейли – бесконечный граф без циклов, где каждый узел связан с фиксированным количеством соседей, но при этом они не перекрываются. На рис. 11 $Z = 3$, т.е. у каждого узла есть 3 соседа, и возникают круги соседей.

Этот граф удобен как модельное представление тем, что для него легко получаются количественные оценки: количество узлов на каждом уровне $N_k = z(z-1)^{k-1}$, т.е. всего узлов в графе

$$N = 1 + \sum_{1}^{l} z(z-1)^{k-1}.$$

В этой сумме самое большое слагаемое $-z^l$, т.е. приближенно $z^l=N$. Тогда для $N \approx 6{,}7$ млн и z=50 (друзей) получаем $L \approx 5{,}8$, т.е. подтверждается идея о 6 рукопожатиях.

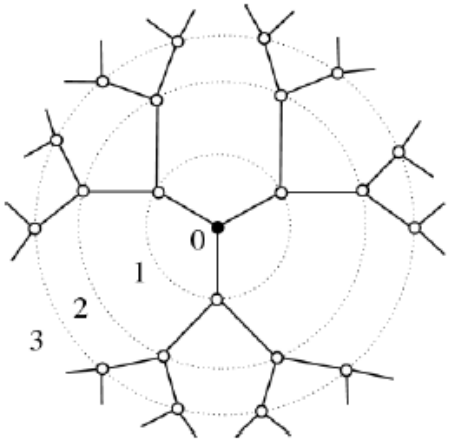


Рис. 11. Модельное представление для эксперимента Милгрема

2. СТЕПЕННЫЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

2.1. ОСНОВНЫЕ ХАРАКТЕРИСТИКИ СТЕПЕННЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

Одной из характеристик сложных или социальных сетей являются специфические законы распределения степеней узлов. В частности, в сетях большинство узлов имеют низкую степень, но в то же время встречаются узлы с высокой степенью, и распределение отличается от того, которое возникло бы в совершенно произвольных сетях. Наличие такого распределения является характеристическим свойством социальных сетей и одной из необходимых характеристик для классификации сети как сложной или социальной.

Оказывается, в природе встречается очень много случаев, когда законы распределения – степенные.

Вспомним основные сведения из теории вероятностей. Если у нас есть непрерывно распределенная случайная величина (т.е. величина может принимать любое значение), то для этой случайной величины можно ввести функцию плотности вероятности $f(x)$, которая определена следующим образом:

$$\Pr(a \le x \le b) = \int_a^b f(x)dx.$$

Интеграл от $f(x)$ в некотором интервале от a до b есть вероятность обнаружить случайную величину в этом интервале. Понятно, что эта функция должна быть неотрицательная и что интеграл должен быть нормирован на единицу.

Также будет полезным понятие функции распределения – вероятности того, что случайная величина примет значение меньше либо равное x:

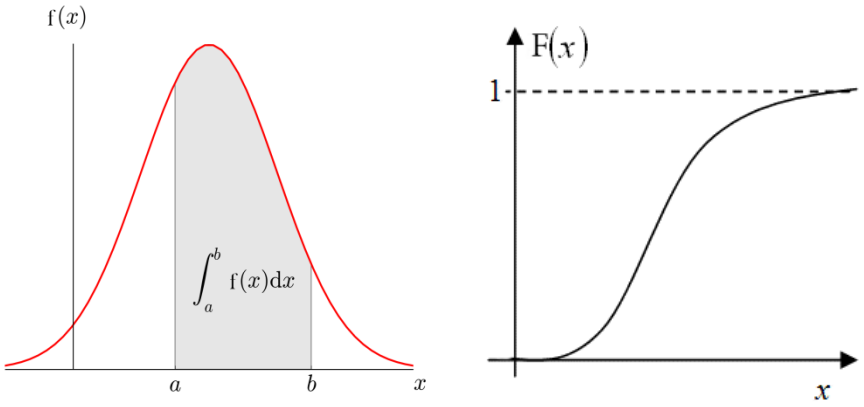
$$F(x) = \Pr(X \le x) = \int_{-\infty}^{\infty} f(x)dx$$

Мы помним, что если речь идет о непрерывной случайной величине, то вероятность выпадения конкретного значения равна нулю, т.е. мы всегда работаем с интервалами (рис. 12, а). Функция плотности вероятности непрерывна справа, предел функции, стремящийся к минус бесконечности, равен 0, а предел функции, стремящийся к бесконечности, равен 1:

$\lim_{\varepsilon \to 0-} F(x) = F(x);$

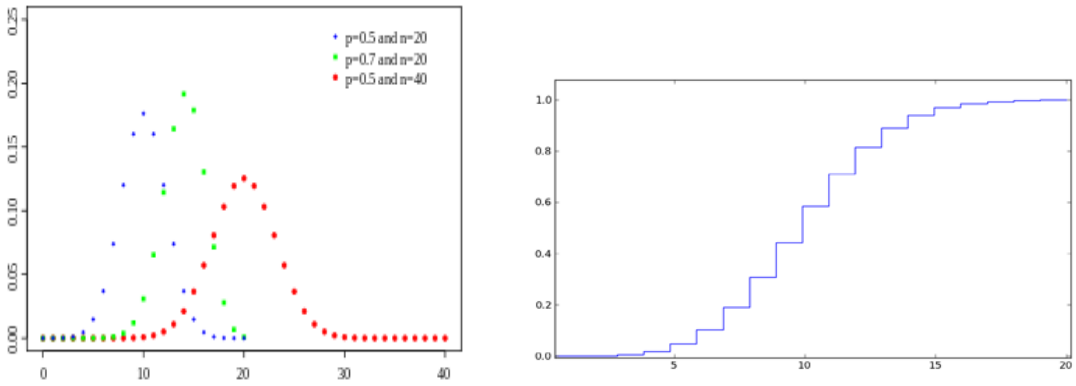
$\lim_{x \to -\infty} F(x) = 0;$

$\lim_{x \to \infty} F(x) = 0.$



а б

Рис. 12. Функции плотности вероятности (а) и распределения (б) для непрерывной случайной величины



а б

Рис. 13. Функции вероятности (а) и распределения (б) для дискретной случайной величины

Наиболее привычным является нормальное распределение, что является следствием центральной предельной теоремы: сумма большого количества случайных величин имеет распределение, близкое к нормальному. Пример – распределение роста людей на Земле.

Однако не все статистики оказываются такими. Например, если функция распределения населения городов не имеет максимального пика, а просто спадает и затухает (рис. 14). Это означает, что существует довольно много городов с маленьким населением и совсем мало городов с довольно большим населением. т.е. возникает «длинный хвост». Для того чтобы его лучше увидеть, его обычно рисуют в логарифмических шкалах (рис. 14, б). Если распределение действительно степенное, то в логарифмической шкале оно будет выглядеть как прямая.

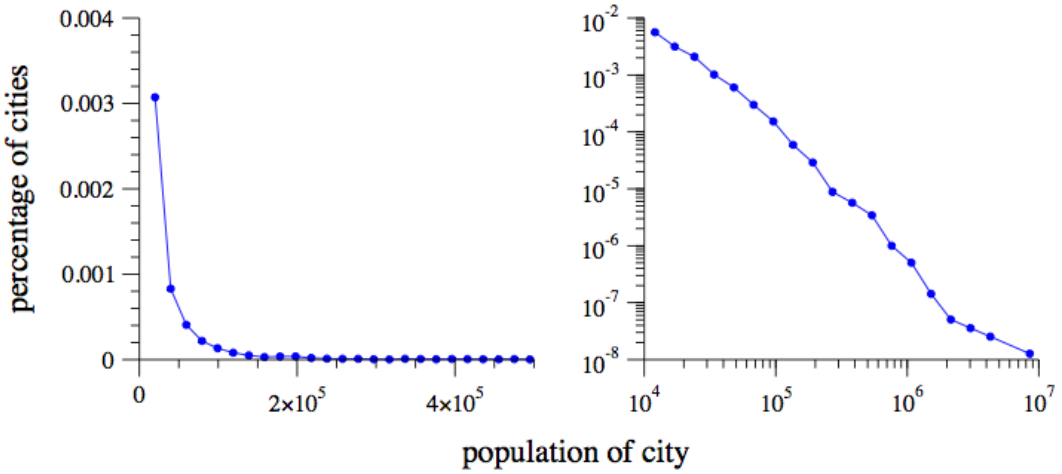


Рис. 14. Распределение населения городов: в линейном масштабе (а); в логарифмическом масштабе (б)

Оказывается, есть очень много явлений, которые обладают степенной функцией распределения – частота встречаемости слов в тексте, число цитирований, популярность ресурсов в Интернете, продажи книг, магнитуды землетрясений и пр. В логарифмическом масштабе они более-менее соответствуют прямой вида $1/x$, причем, возможно, не от нуля, а начиная с некоторого значения $1/x$. Это и есть степенныее распределения. Они и внешне, и по свойствам сильно отличаются от гауссова распределения.

Формально степенное распределение имеет вид

$$f(x) = Cx^{-\alpha} = \frac{C}{x^{\alpha}}, \quad x \geq x_{\min}$$

Здесь C является нормировочной константой. Также нужно ограничить x, так как в нуле график расходится. График обычно нормируется, чтобы площадь под всей кривой равнялась единице, что приводит к выражению

$$f(x) = Cx^{-\alpha} = \frac{\alpha - 1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha}.$$

Подчеркнем, что в этом законе 2 неизвестных параметра – α и $x_{\min}$.

Сравнение различных распределений представлено на рис. 15, а, в линейной шкале и на рис. 15, б – в логарифмической шкале.

Из рис. 15, а, видно, что экспоненциальный хвост затухает гораздо быстрее, чем степенной, т.е. экспоненциальное распределение стремится к нулю гораздо быстрее.

На рис. 15, б, представлены кривые с разными параметрами α, а на рис. 15, в, – отношение степенного закона к экспоненциальному закону.

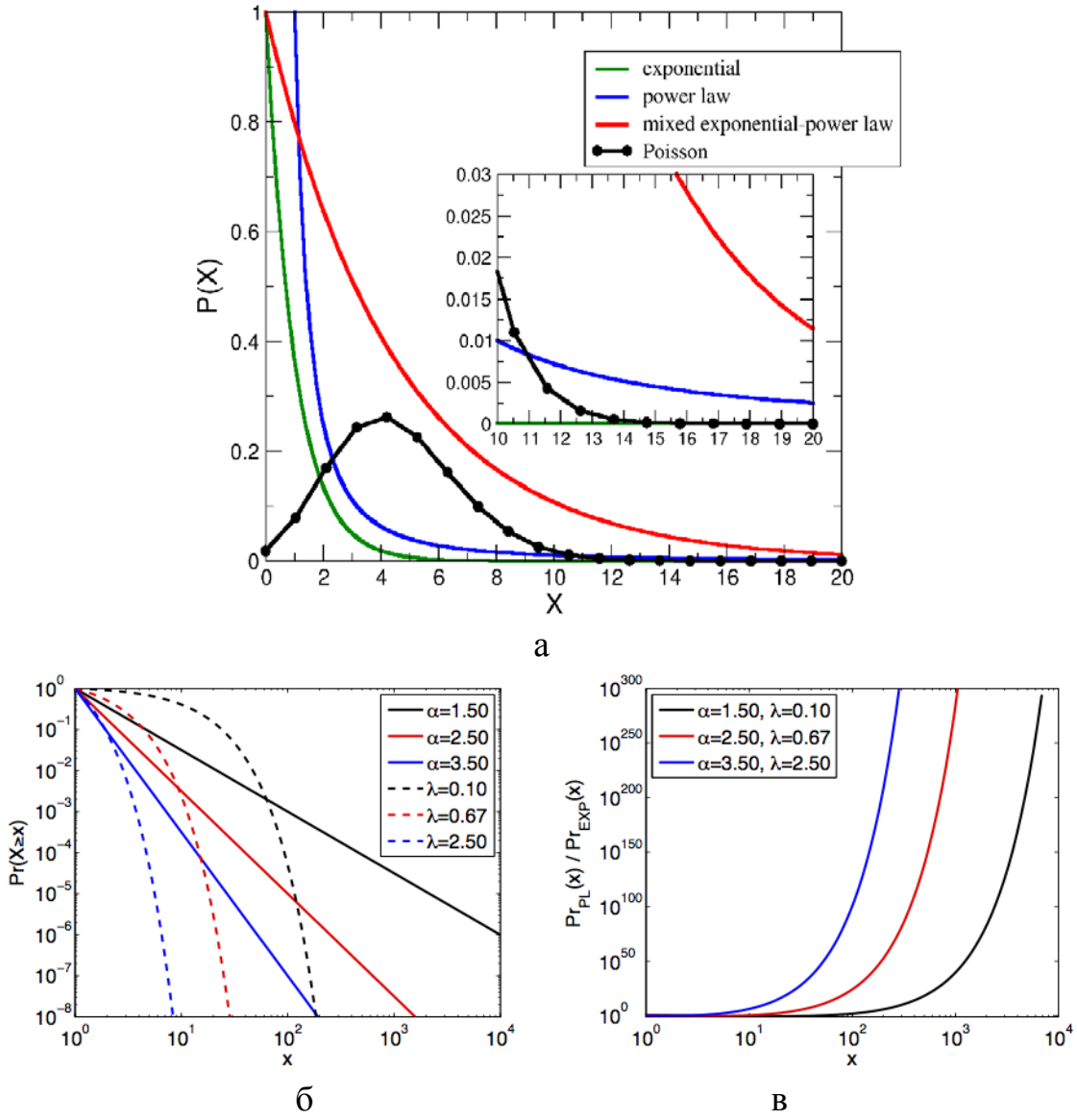


Рис. 15. Сравнение экспоненциального и степенного распределений

У степенного распределения есть несколько необычных свойств:

- среднее значение расходится, если само распределение имеет $\alpha<2$; при $\alpha>1$распределение существует, но среднее этого распределения неопределенно. На практике это означает, что среднее является очень нестабильной величиной, т.е. использовать его нельзя;
- дисперсия имеет смысл только при $\alpha>3$;
- в общем случае, чтобы n-ный момент распределения имел смысл, должно выполнятся условие $\alpha>(n+1)$;
- степенное распределение – единственное, которое выглядит одинаково независимо от масштаба. Пример: если файлы размером 2 кБ встречаются в 4 раза реже, чем файлы размером 1 кБ, то файлы размером 2 МБ встречаются в 4 раза реже файлов размером 1 МБ.

2.2. ПРИЛОЖЕНИЯ СТЕПЕННЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

Закон Ципфа (рис. 16) – эмпирическая закономерность распределения частоты слов естественного языка: если все слова языка (или просто достаточно длинного текста) упорядочить по убыванию частоты их использования, то частота *n*-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру *n* (так называемому рангу этого слова). Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и так далее.

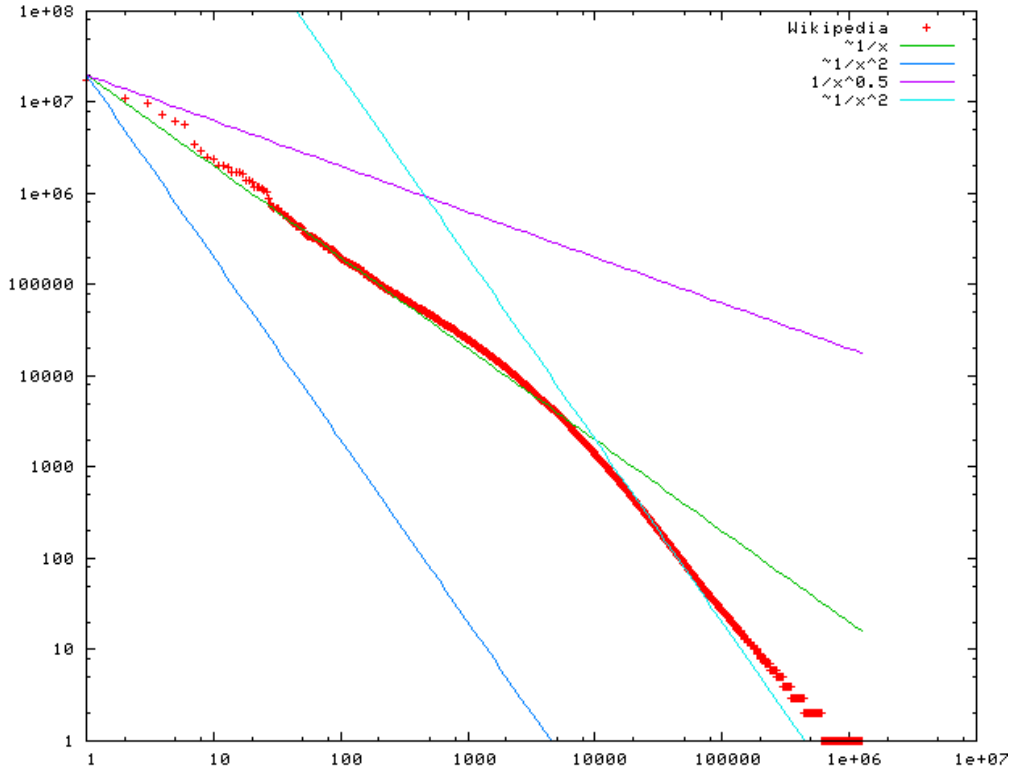


Рис. 16. Закон Ципфа

На рис. 16 представлен график «ранг – частота». Самое частое встречаемое слово в тексте будет иметь ранг 1, менее встречаемое – ранг 2 и т.д. Измерив угол наклона графиков и скорректировов его на единицу, мы фактически получаем показатель степенного закона.

Степенные законы в социальных сетях. В любой социальной сети есть узлы, и у этих узлов есть соседи, таким образом, у узлов есть степени k=1, 2, … n, в данном случае это дискретные значения. Показано, что для многих экспериментальных и сложных сетей зависимость $P(k)$ является степенной:

$$P(k) = \frac{n_k}{n} = \frac{C}{k^{\alpha}}.$$

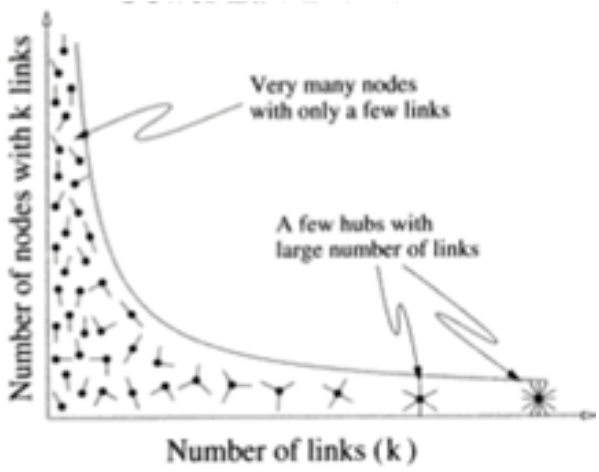


Рис. 17. Степенной закон для социальных сетей

Когда мы работаем с реальными данными, у нас есть список узлов и их степеней. Однако, построив график зависимости степеней от их количества, мы не всегда сможем точно определить его параметры α и $x_{\min}$. Связано это с тем, что на графике всегда присутствует шум. В этом случае нужно использовать специальные подходы, описание которых можно найти в литературе.

3. МОДЕЛИ ФОРМИРОВАНИЯ И РОСТА СЕТЕЙ

Напомним характерные эмпирические факты для реальных сетей:

- степенной закон распределения степеней узлов, то есть у сети очень много узлов с малым количеством соседей и в то же время существуют узлы с очень большим количеством соседей;
- маленький диаметр сетей (маленькое среднее расстояние);
- высокий кластерный коэффициент;
- наличие гигантской связной компоненты;
- иерархическая структура.

Другими словами, у реальных сетей есть много характерных признаков. И, естественно, стали возникать модели, которые пытаются объяснить появление сетей и, соответственно, позволяют сгенерировать эти сети.

Основные модели, объясняющие появление сетей:

- модель случайных графов. Одна из первых сетевых моделей, которая была создана в 1959 году Erdos и Renyi. Многие сегодняшние модели получили свое начало в модели Erdos и Renyi;
- модель малого мира (Watts, Strogatz, 1998);
- модель предпочтительных присоединений (Barbasi, Albert, 1999).

3.1. МОДЕЛЬ СЛУЧАЙНЫХ ГРАФОВ (МОДЕЛЬ ЭРДОША–РЕНЬИ)

Рассматривается граф $G\{E, V\}$, где E – ребра, V – множества узлов.

Существует два типа модели Эрдоша–Реньи – $G_{n,m}$ модель и $G_{n,p}$ модель.

$G_{n,m}$ модель. У данной модели есть 2 параметра – n (число узлов) и m (число ребер). Тогда:

$n(n–1)/2$ – количество пар узлов, которые могут быть соединены ребром (количество паросочетаний);

$C^m_{n(n-1)/2}$ – количество способов выбора m ребер из $n(n–1)/2$ числа ребер.

О случайном графе можно думать как об одном графе, выбранном случайным образом из этого множества $C^m_{n(n-1)/2}$.

$G_{n,p}$ модель. У данной модели есть 2 параметра – n (число узлов) и p – вероятность того, что любая случайно выбранная пара узлов соединена ребром. Если $p = 0$, то нет ребер вообще в графе, если $p = 1$, то получаем полный граф, все узлы соединены ребрами. Тогда число ребер – это случайное число, которое имеет математическое ожидание $\langle m \rangle$

$$\langle m \rangle = p \cdot \frac{n(n-1)}{2}$$

Средняя степень узла – количество его соседей:

$$\langle k \rangle = \frac{1}{n}\sum_i k_i = \frac{2\langle m \rangle}{n} = p\ (n-1) \approx pn$$

Для случайных графов легко посчитать функцию распределения степени узлов: она описывается формулой Бернулли

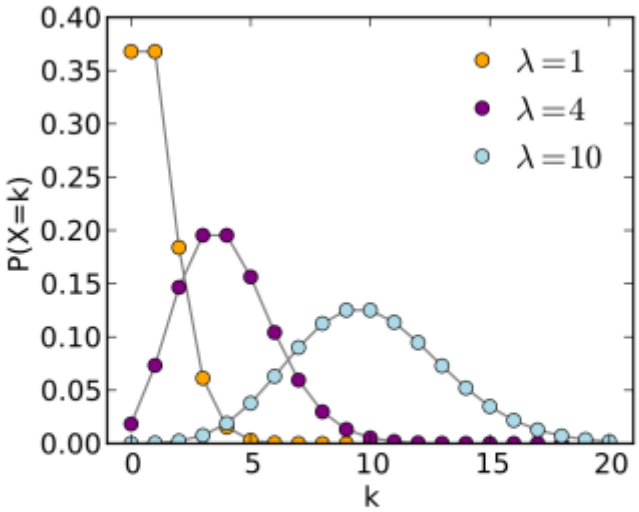
$$P(k_i = k) = P(k) = C_{n-1}^{k} p^k (1-p)^{n-1-k}$$

Известно, что у распределения Бернулли есть предельный случай: если $n \to \infty$ но при этом зафиксировать среднее начение $\langle k \rangle = pn = \lambda$), тогда это распределение превращается в распределение Пуассона (рис. 18).

$$P(k) = \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!} = \frac{\lambda^k e^{-\lambda}}{k!}$$

Это распределение дискретно, т.е. k имеет целочисленное значение.

Видно, что распределение симметрично относительно своего пика. В данной модели все узлы графа будут иметь степень, близкую к средней степени, и очень мало сильно отличающихся узлов. В таком графе найти узлы с маленькой степенью почти невозможно. И эта модель расходится с экспериментами.



$$P(k_i = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda = pn$$

Рис. 18. Распределения Пуассона и Бернулли

Интересное свойство модели – наличие фазового перехода, при котором плавное изменение параметра приводит к скачкообразным изменениям свойств системы. Это свойство послужило к созданию новой науки – Complex systems.

Рассмотрим граф $G_{n,p}$ как функцию от p. При p=0 имеем пустой граф (нет связей), при p =1 – полный граф. Начнем понемногу увеличивать p от значения p=0. В промежутке от 0 до 1 происходит структурное изменение – появляется возможность попасть из одного узла в любой другой за какое-то количество шагов. Другими словами, происходит фазовый переход, который заключается в том, что появляется связность. В какой-то момент времени возникает гигантская связная компонента (большая часть графа, которая связана между собой).

Момент скачка из несвязанного состояния к появлению гигантской связной компоненты и есть фазовый переход. Т.е. плавное увеличение p приводит к тому, что неожиданно возникает связь между всеми ребрами (рис. 19).

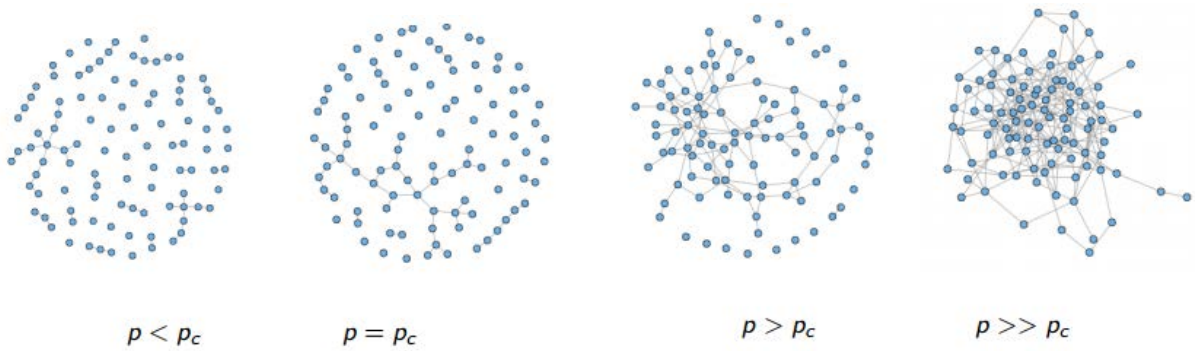


Рис. 19. Возникновение гигантской связной компоненты

Размер гигантской связной компоненты можно определить, решая трансцедентное уравнение

$1 - s = e^{-\lambda s}$.

Здесь s – это доля узлов, которые принадлежат гигантской связной компоненте, и $\lambda = pn$. Можно показать, что

при $\lambda \to \infty$ $s \to 1$;

при $\lambda \to 0$ $s \to 0$;

другими словами, если средняя степень узла очень велика, то получается плотный граф, и все узлы принадлежат гигантской связной компоненте.

Критическую точку – точку перехода системы из одного состояния в другое – можно определить из графического решения уравнения (рис. 20). Слева показана просто прямая x=s, справа – насыщающаяся экспонента. В зависимости от λ могут возникать различные кривые: если λ маленькая, то кривая плоская, если λ возрастает, кривая становится более выпуклой.

Пересечение возможно в двух точках, что дает нулевое решение (при s=0) и ненулевое решение. Критической точкой для возникновения ненулевого решения является значение средней степени узла λ:

$\lambda_c = p_c n = 1$.

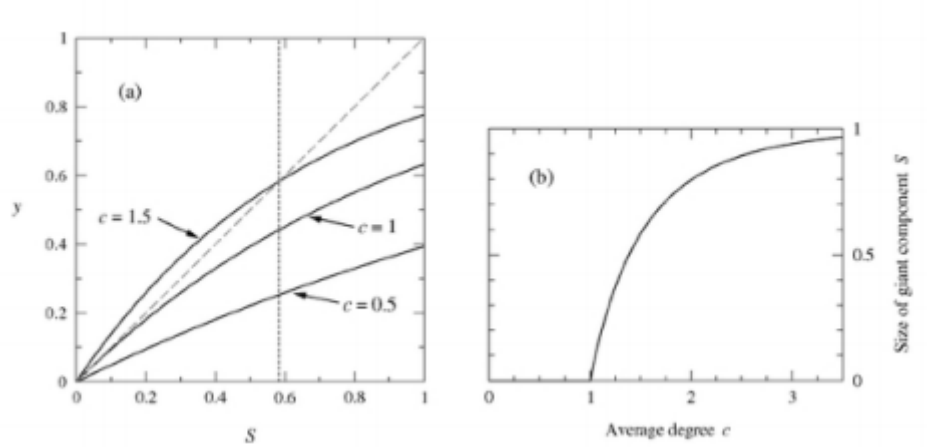


Рис. 20. Графическое решение уравнения

Если $0 < \lambda < 1$, то гигантская связная компонента отсутствует, если $\lambda > 1$, то гигантская связная компонента резко возрастает. Выбирая критическое значение p, мы можем это контролировать.

У графа $G_{n,p}$ есть другие свойства. Если p ведет себя как степенная функция от n в степени, то это отражается на структуре графа (рис. 21).

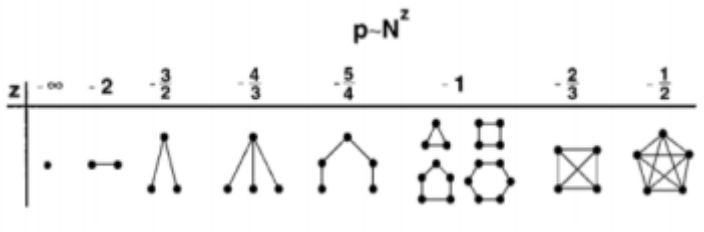


Рис. 21. Изменение структуры связей в зависимости от p

Меняя p и n , можно гарантировать не только то, что возникает гигантская связная компонента, но и то, что возникают определенные структуры (циклы, полные субграфы порядка 4, 5 и т.д.). Другими словами, можно определить те моменты, когда в графе появляются определенного типа поведение. Преимущество модели заключается в том, что эти моменты математически хорошо просчитывается.

Мы можем определить диаметр такого графа. Возьмем два узла с их окружениями (рис. 22). Обозначим d_{ij} – расстояние между узлами i и j и l – длину пути между узлами,

$$l = s + t + 1$$

(s – одно окружение, t – второе окружение и 1 для того, чтобы их соединить). Тогда выражение

$$P(d_{ij} > s + t + 1)$$

означает вероятность того, что между окрестностями (окружениями) этих графов не существует связей, т.е. нельзя от i перейти в j.

Диаметр графа – это то наименьшее значение l , когда $P(d_{ij} > l) = 0$, т.е. для любой пары узлов существует нулевая вероятность того, что они раз-

делены большим расстоянием, чем l. Учитывая, что в момент появления гигантской связной компоненты граф становится деревом, можно показать, что у случайного графа диаметр невелик и равен

$$d = \frac{\ln n}{\ln \langle k \rangle}.$$

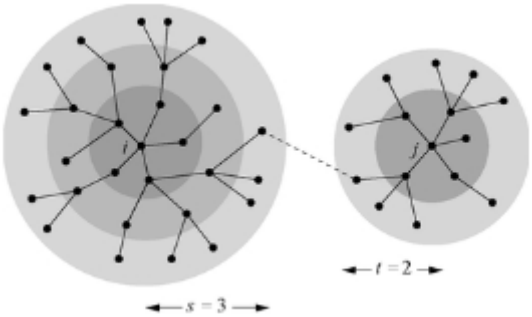


Рис. 22. К расчету диаметра графа

Кластерный коэффициент – это вероятность того что два соседа в графе связаны между собой, т.е. это количество узлов между ближайшими соседями, деленное на максимальное количество узлов между ближайшими узлами:

$$C(k) = \frac{\#\text{of links between NN}}{\#\text{max number of links NN}} = \frac{pk(k-1)/2}{k(k-1)/2} = p$$

или

$$C = p = \frac{\langle k \rangle}{n}$$

Когда n велико, кластерный коэффициент стремится к нулю, в больших графах кластерный коэффициент очень маленький.

Модель случайного графа легко видоизменить и подстроить под нее функцию распределения, тем самым подгоняя свойства модели под имеющиеся экспериментальные данные.

Модель Эрдоша–Реньи – первая модель, предложенная для анализа сетей. Эта модель очень разумна. Она предусматривает всего два параметра – заданное (зафиксированное изначально) количество узлов и вероятность наличия связи между парами этих узлов. У этой модели есть интересные свойства, такие как, например, наличие гигантской связанной компоненты и эффект малого мира/малого диаметра. Ее недостатками являются нестепенное распределение степеней узлов и маленький кластерный коэффициент.

Другими словами, эта модель хороша тем, что она полностью просчитывается аналитически, т.е. легко ее настраивать и менять ее свойства. Но, с другой стороны , она слабо соответствует результатам эксперимента.

3.2. МОДЕЛЬ ПРЕДПОЧТИТЕЛЬНОГО ПРИСОЕДИНЕНИЯ (МОДЕЛЬ БАРАБАШИ–АЛЬБЕРТА)

Мотивацией этой модели послужили растущие сети, которые развиваются со временем и которые нельзя описать моделью Эрдоша–Реньи. Примеры таких сетей приводятся ниже.

Сети цитирования формируются следующим образом. Имеется статья некоторого автора, потом другой автор пишет статью, цитируя эту статью (в конце каждой статьи есть целый список авторов, потому как автор цитирует много статей), а потом проходит время, и уже его начинают цитировать, и т.д. Таким образом, цитируемость работы зависит от времени: чем раньше работа была написана, тем больше у нее шансов быть цитированной. В то же время сами узлы не будут заданы изначально, они появляются с течением времени. Имеет место характерная динамика: появляется новый узел и появляются связи, при этом сначала появляются связи, входящие в этот узел, затем появляются связи, от этого узла идущие дальше.

Сети взаимодействия (коллаборации). Похожая картина имеет место для сетей взаимодействия. Например, чем дольше вы работаете в какой-то области, тем больше у вас связей, тем с большим количеством людей вы сотрудничаете.

Web-граф. В самом начале его развития было мало страниц и было мало связей. С течением времени появлялись новые страницы, на них ставились линки, указывающие на эти страницы. Понятно, что в большинстве своем степень более старых страниц в Интернете росла, если они были популярны.

Социальные сети. Есть люди, у которых мало друзей, а есть те, у кого очень много друзей. Если у вас много друзей, то вам проще приобрести новых друзей, и, соответственно, никто не хочет дружить с лузерами/одиночками. Таким образом, опять видна явная динамика: социальная сеть начинается с малого количества узлов и небольшого числа связей между ними, а с течением времени новые люди присоединяются к социальной сети, с ними устанавливаются связи, но у тех, кто давно в социальной сети, гораздо больше друзей, как правило. Очень важно, что время в этой системе отсчитывается от момента присоединения узла к системе.

Ясно, что в настоящих сетях происходят и другие процессы:

- в коллаборационных сетях с течением времени люди могут переставать взаимодействовать с теми, с кем взаимодействовали раньше;
- некоторые веб-страницы перестают поддерживаться;
- умирая, люди уходят из социальных сетей.

Однако сети цитирования – это красивый и чистый случай, где узлы и связи не уходят, поэтому первоначально модель Барабаши–Альберта создавалась именно для этого случая, а потом ее расширили и стали успешно использовать для web-графа как одну из первых моделей.

Идея модели – модернизировать модель случайных сетей в направлении описания динамики развития:

- в начальный момент времени (t=0) есть m несвязанных узлов;
- на каждом шаге (t=1, 2, 3…) будем добавлять новый узел с m ребрами;
- хотя количество связей, с которым приходит новый узел в граф, фиксировано (m), к какому именно узлу он присоединяется – выбирается случайным образом, т.е. рост идет случайно.

Можно показать, что ожидаемая степень i-го узла равна

$$k_i(t) = m\left(1 + \log\frac{t}{i}\right),$$

где i – индекс (номер) узла, соответствующий тому времени, когда узел присоединился. Рис. 23 показывает поведение этой модели как функцию времени для m=20, i=10, 20, 40.

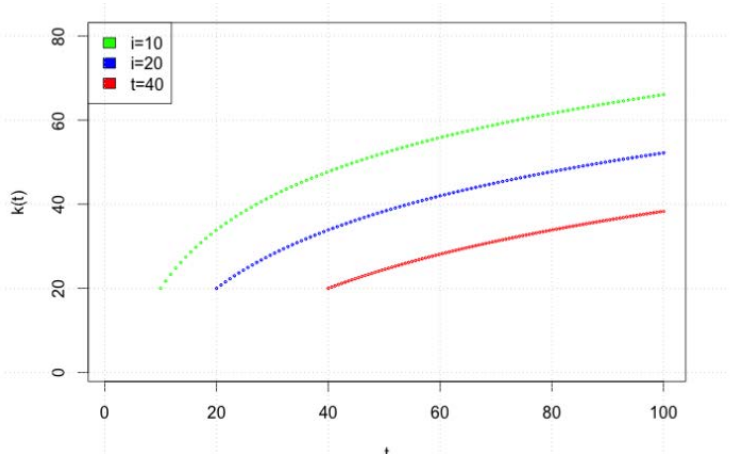


Рис. 23. Поведение модели Барабаши–Альберта как функция времени

Обратим внимание, что те узлы, которые присоединились раньше, имеют большую степень с течением времени, чем позже присоединившиеся узлы. Действительно, все эти кривые не пересекаются, что и означает преимущество присоединившегося ранее. Это преимущество в таких моделях победить нельзя: когда узел только приходит в систему, у него конкуренция за новые связи очень небольшая, т.е. любой новый приходящий узел в любом случае с ним свяжется; когда же система (граф) становится большой, то при появлении нового узла он может установить с другими узлами только ограниченное количество связей. Другими словами, шанс случайного узла получить связь становится все меньше и меньше.

Рис. 24 показывает поведение этой же модели как функцию от i для m=20, t=50, 100, 200.

Теперь посмотрим на граф по той же самой формуле, только как функцию от номера узла i.

Этот граф будет означать, что мы проследим за несколькими узлами и посмотрим на эти узлы (на поведение/степень этих узлов) как функцию от i (от номера узла). Например, мы хотим найти все те узлы, которые в

момент времени t имеют степень меньше какой-то данной степени, т.е. в конечном счете посчитать для такого графа функцию распределения степеней узлов. Можно показать, что функция распределения степеней узлов в этом случае равна

$$P(k) = \frac{1}{m} e^{-\frac{m-k}{m}}.$$

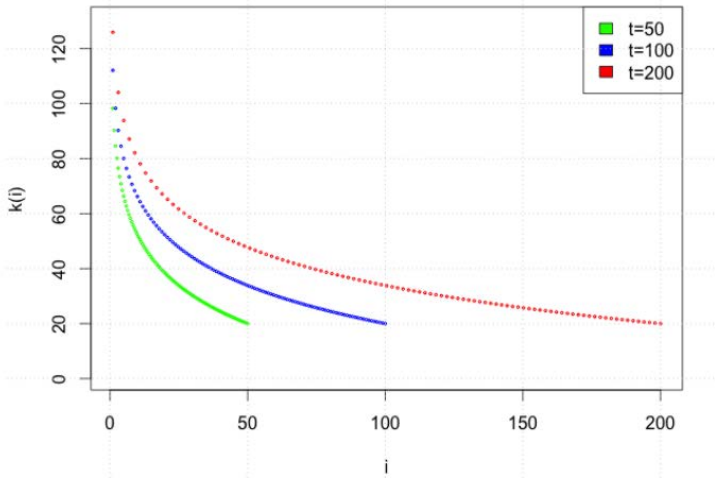


Рис. 24. Поведение модели Барабаши–Альберта как функция номера узла

Следующий шаг был сделан Барабаши и Альбертом в 1999 году при переходе к описанию web-графов. Они пытались придумать модель, которая описывает распределение степени узлов. Они увидели, что степени узлов распределены не по Бернулли или Гауссу, не экспоненциально, а как-то иначе. Поэтому они взяли модель динамически растущего случайного графа и предложили сделать небольшое изменение к этой модели, которое, тем не менее, очень сильно изменило свойства.

Сохраним ту же самую ситуацию, когда вначале есть некоторое количество узлов и дальше с течением времени присоединяются новые узлы, но изменим одно условие. Вместо того, чтобы присоединение к новым узлам происходило просто случайным образом, они ввели метод предпочтительного присоединения. Таким образом, вероятность у нового узла присоединиться к существующему узлу пропорциональна степени этого существующего узла.

У старых узлов и так было преимущество перед новыми узлами – они с течением времени имели степень большую, чем новые узлы. Добавление предпочтительного присоединения только усилило этот эффект. Наглядно представить эту систему можно с помощью рис. 25.

Можно показать, что с введением предпочтительного присоединения функция распределения степеней узлов становится равной

$$P(k) = \frac{2m^2}{k^3}$$

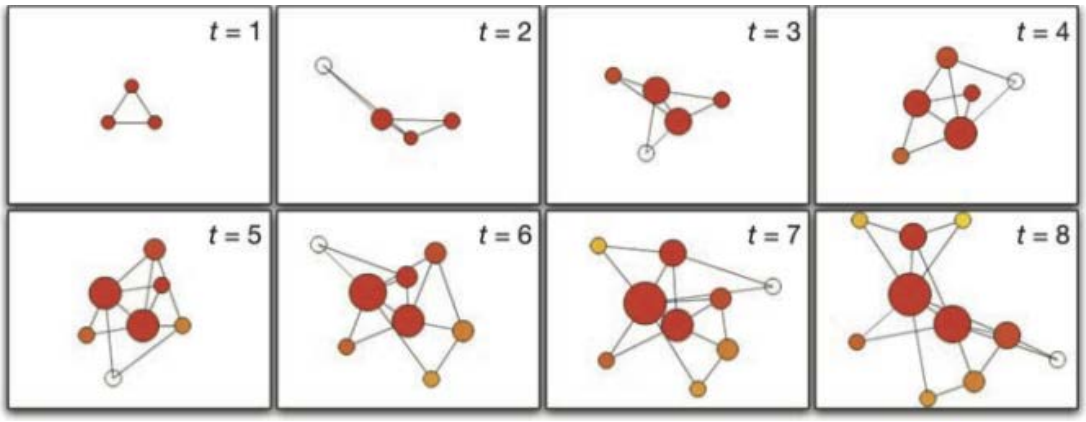
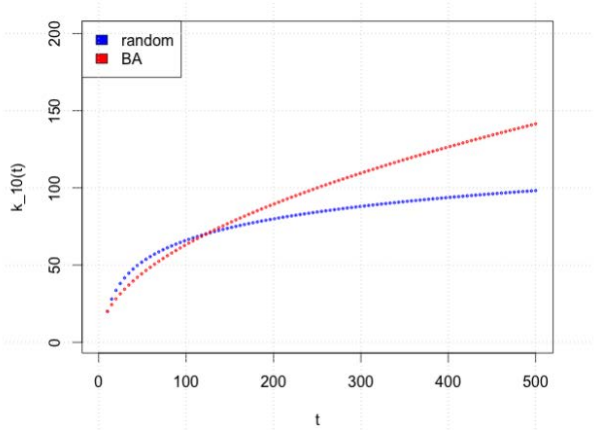


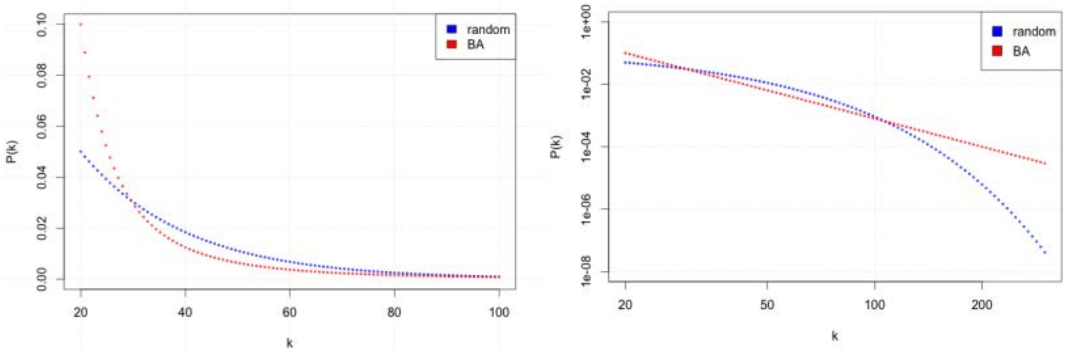
Рис. 25. Добавление предпочтительного присоединения

Сравнение моделей Барабаши–Альберта и случайно растущего графа представлено на рис. 26, а, б. По осям показаны степень узла и функция распределения $P(k)$.



$$BA: k_i(t) = m\left(\frac{t}{i}\right)^{1/2}, \quad RG: k_i(t) = m\left(1 + \log\left(\frac{t}{i}\right)\right)$$

а



$$BA: P(k) = \frac{2m^2}{k^3}, \qquad RG: P(k) = \frac{1}{m}e^{-\frac{k-m}{m}}$$

б

Рис. 26. Сравнение моделей Барабаши–Альберта и случайного графа: по степени узлов (а), по функции распределения (б)

Видно, что в модели Барабаша–Альберта присутствует больше узлов с малой степенью, чем в модели случайного графа.

Если вспомнить, что степенной закон выглядит как прямая при логарифмировании, и прологарифмировать оба выражения, мы получим прямую в одном случае и кривую в другом. Стоит отметить, что эта кривая будет прижиматься сильнее к осям, чем прямая, что говорит о вероятности наличия узлов с очень высокой степенью. Другими словами, модель Барабаши–Альберта демонстрирует гораздо большую вероятность иметь узлы с высокой степенью, чем модели просто случайно растущего графа, т.е. является более предпочтительной.

Для этой модели можно посчитать среднюю длину пути:

$\langle L \rangle \sim \log(N) / \log(\log(N))$

Она примерно ведет себя как логарифм N – немного послабее, но, все равно, присутствует логарифмическая связь. Другими словами, эта модель удовлетворяет концепции «малого мира», и это ее сильное преимущество.
В то же время кластерный коэффициент просчитать по формулам нельзя, это только результат численных экспериментов:

$C \sim N^{-0.75}$.

Как видно, при больших N он уходит к нулю. Фактически, когда N становится очень большим, кластерный коэффициент становится очень маленьким. Кластерный коэффициент пропорционален числу треугольников, которые возникают в системе. Следовательно, согласно модели Барабаши–Альберта, если система становится очень большой, то в ней остается все меньше и меньше треугольников – локальных сильных связей. В этом она противоречит тому, что наблюдается.

Реальное поведение кластерного коэффициента – единственное свойство, которое модели не удается воспроизвести. Это обстоятельство иллюстрирует рис. 27, где присутствует мало треугольников.

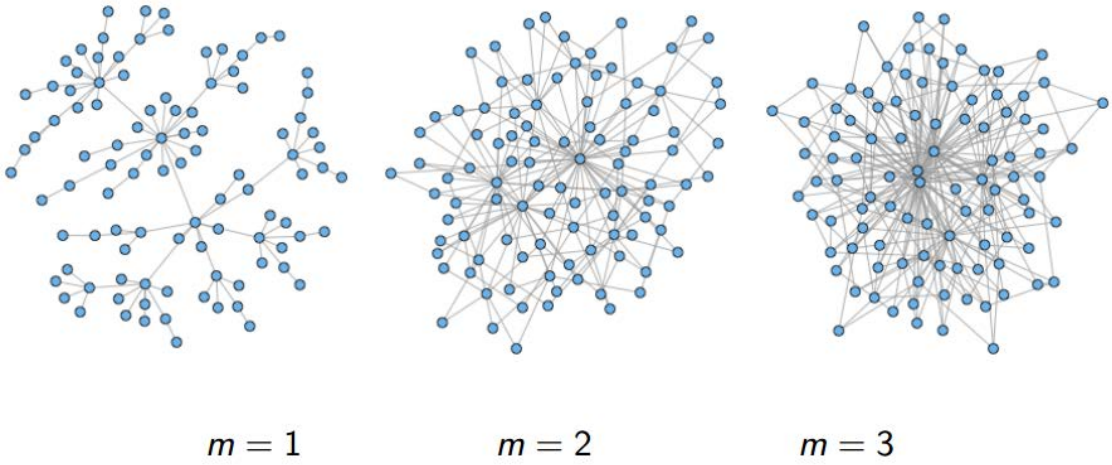


Рис. 27. Поведение модели Барабаши–Альберта при разных степенях присоединяемых узлов m

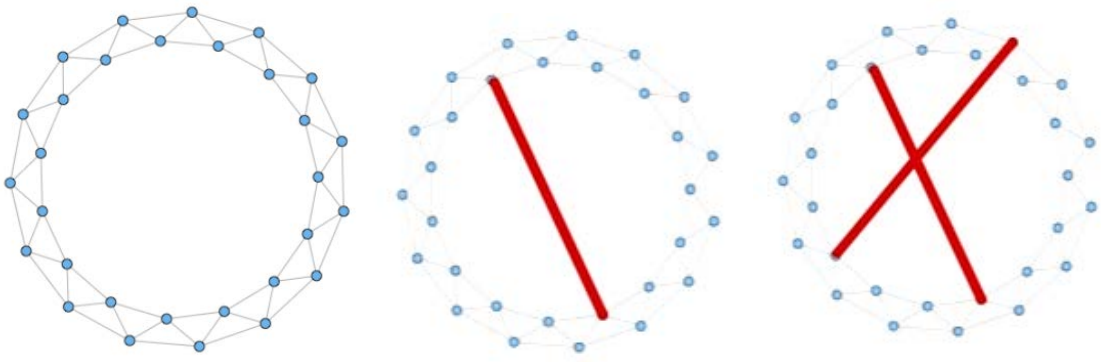
3.3. МОДЕЛЬ МАЛОГО МИРА (МОДЕЛЬ УОТТСА И СТРОГАТЦА)

Модель малого мира была разработана на год раньше модели Барабаши–Альберта. Модель достаточно простая, но аналитически просчитывается только в очень простом случае.

Мотивация модели – требуется воспроизвести реальную сеть, причем на первое место поставим маленький диаметр и высокий кластерный коэффициент, а не функцию распределения. Поэтому целесообразно строить модель с ситуации, для которой характерен высокий кластерный коэффициент. В примере (рис. 28, а) взят граф в форме круга (авторы использовали графы на круге, чтобы избежать ограничительных условий). Связи в форме треугольников создают высокий кластерный коэффициент.

Кластерный коэффициент – это фактически отношение числа связей, которые имеют ближайшие соседи узла, к максимально возможному числу связей между ними. У произвольного узла на графе рис. 28 – 4 соседа. Максимальное количество связей между ними равно 4×3/2=6, а реальное число связей между соседями – 3, т.е. кластерный коэффициент для произвольного узла равен 3/6=1/2. Так как граф абсолютно симметричен, то полный кластерный коэффициент для него также равен 1/2.

Однако диаметр этого графа высокий. Спрашивается, можно ли сохранить высокий кластерный коэффициент и одновременно уменьшить диаметр? Да, можно. Для этого достаточно в этот граф добавить некоторые длинные связи – например, как показано на рис. 28, б, в. Введя несколько таких связей, мы резко уменьшим диаметр графа.



а б в

Рис. 28. Модель малого мира: а – исходный граф;
б, в – добавление длинных связей

Однако, вводя такие длинные связи, мы изменим среднюю степень узла в графе (оставляем количество узлов, но добавляем новые связи). Можно попробовать сохранить эту среднюю степень узла фиксированной. Для этого, вместо того, чтобы просто добавлять какие-то связи, можно попробовать взять одну связь, локально оторвать ее и перенаправить в другое

место. Тогда полное количество связей не меняется, поэтому средняя степень узла сохраняется, но мы получим связи на большом расстоянии.

Перечислим еще раз этапы построения модели:

– Начинаем с регулярной решетки с некоторым количеством ближайших соседей.
– Вводим некоторый параметр *p*, модель становится однопараметрической.
– Когда *p*=0, имеем регулярную решетку (идеальное колечко). А когда *p*=1, то получается случайный граф.

Параметр *p* отображает долю пересоединенных ребер. Наглядный пример приводится на рис. 29.

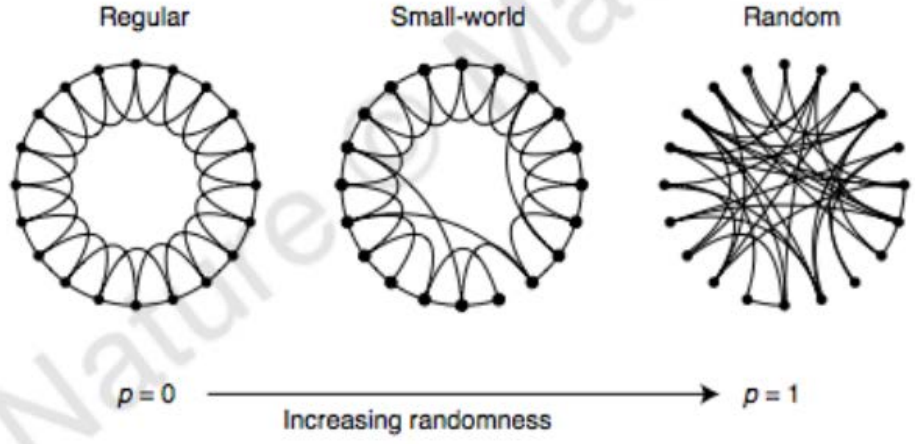


Рис. 29. Эволюция модели с пересоединением ребер

Легко видеть, что, если пересоединить все ребра, то мы получим случайный граф. Когда все связи регулярны, кластерный коэффициент высокий, но диаметр графа также высокий, а в случайном графе он близок к нулю. Поэтому следует ожидать, что существует золотая середина, где, с одной стороны, сохраняется высокий кластерный коэффициент, а с другой – появляется момент возникновения малого мира.

В модели Воттса– Строгатца отчетливо виден переход от регулярного мира к случайному состоянию через состояние малого мира. Иначе говоря, одним параметром (параметром добавления случайных связей) можно регулярную решетку превратить в малый мир.

Авторы модели провели численные расчеты характеристик:

– распределение степеней узлов – пуассоновское
– средняя длина пути $\langle L(p)\rangle$:

  $p \to 0 \quad \langle L(0)\rangle = 2n/k$ *круговая решетка*

  $p \to 1 \quad \langle L(1)\rangle = \log(n)/\log(k)$ *случайный граф*
– кластерный коэффициент $C(p)$:

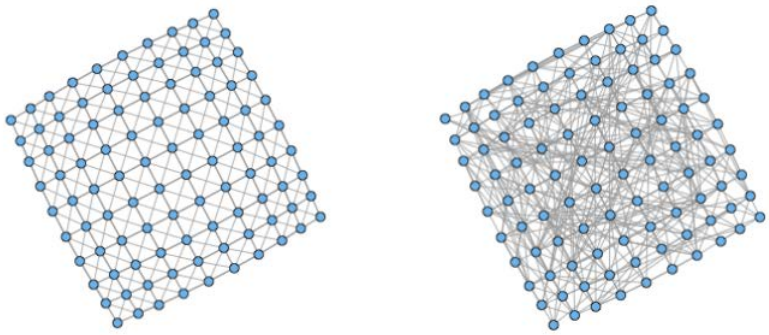
$p \to 0 \quad \langle C(0) \rangle = 3/4 = \text{const} \quad$ *круговая решетка*

$p \to 1 \quad \langle C(1) \rangle = k/n \quad$ *случайный граф*

В модели Уоттса– Строгатца не заложен никакой контроль над функцией распределения степеней узлов. Эта функция изменяется от пуассоновской для случайного графа к отдельной точке (значению степени узла) для регулярной решетки. Ни тот, ни другой крайний случай не соответствует эмпирическим распределениям степенного закона, наблюдаемым на практике, поэтому не стоит ожидать, что, смешав их, можно получить степенной закон. Поэтому функцию распределения по этой модели не вычисляют, а основное внимание уделяют вычислению средней длины пути и кластерного коэффициента.

Интересно, что существует некоторый интервал значений *p*, в котором уже кластерный коэффициент не нулевой (недостаточно упал), а в то же время средняя длина пути уже становится достаточно маленькой. То есть можно с помощью одного параметра перейти из полного порядка к случайности и пройти через некоторую зону, где появляются эффекты малого мира.

Еще один пример приведен на рис. 30. Имеется идеальная решетка с высоким кластерным коэффициентом 0,49 (много треугольников), а средняя длина пути 3,58. Перебрасывая 20% связей, мы уменьшили кластерный коэффициент до 0,19 и, в то же время, среднюю длину пути уменьшили до 2,32 и тем самым перешли от регулярной решетки (а) до состояния (б). Эффект не настолько яркий, потому что *n* невелико, но он заметен.



а б

Рис. 30. Эффект появления малого мира (б) из регулярной решетки (а)

Приведем сводку рассмотренных моделей.

В таблице представлены три основных модели (случайного графа , Барабаши–Альберта, Уоттса–Строгатца) и одна вспомогательная (эмпирические сети).

	Random	BA model	WS model	Empirical networks
P(k)	$\frac{\lambda^k e^{-\lambda}}{k!}$	k^{-3}	poisson like	power law
C	$\langle k \rangle / N$	$N^{-0.75}$	const	large
$\langle L \rangle$	$\frac{\log(N)}{\log(\langle k \rangle)}$	$\frac{\log(N)}{\log\log(N)}$	$\log(N)$	small

Таблица 1. Сравнение основных моделей

– Модель случайного графа. Функция распределения оказалась функцией распределения Бернулли. Она не соответствует эмпирическим законам. Кластерный коэффициент очень низок, когда N велико, он стремится к нулю.
– Модель Барабаши–Альберта – единственная, которая дает хорошую картину для функции распределения, а именно степенной закон, который наблюдается в реальной жизни. Однако кластерный коэффициент стремится к нулю при большом N, что плохо.
– Модель малого мира. Функция распределения близка к типу Пуассона. Кластерный коэффициент может быть достаточно большим.

Как оказалось, смоделировать короткий диаметр и малый мир не так сложно. Все перечисленные модели это делают. Однако все они, в той или иной мере, имеют проблемы с кластерным коэффициентом.

Рассмотренные модели являются базовыми, они отличаются очевидностью и простотой. К настоящему времени создан целый ряд других моделей, но они получались более сложными и не на столько очевидными, как эти три основные.

4. СТРУКТУРНАЯ ЭКВИВАЛЕНТНОСТЬ

4.1. ТИПЫ ЭКВИВАЛЕНТНОСТИ

Существует три вида эквивалентности:

- структурная;
- автоморфная;
- регулярная.

Два актора являются ***структурно эквивалентными***, если они имеют одинаковые отношения со всеми другими элементами сети.

Два актора являются ***автоморфно эквивалентными***, если существуют автоморфное отображение этих узлов, т.е. акторы остаются на том же расстоянии от всех других участников, если они поменялись местами, и акторы других классов также были заменены. Идея состоит в том, что наборы акторов могут быть эквивалентными, если они имеют одни и те же закономерности связей – «параллельные» структуры.

Два актора ***регулярно эквиваленты***, если они имеют одинаковые связи с другими регулярно эквивалентными акторами.

На рис. 31 показаны все три вида эквивалентности.

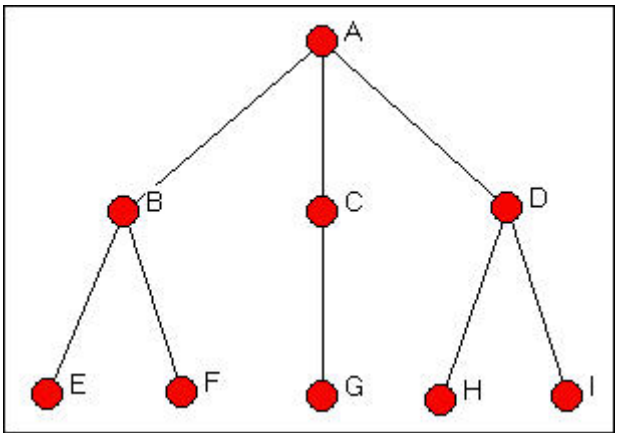


Рис. 31. Различные типы эквивалентности

Структурная эквивалентность (одни и те же отношения к другим узлам): {A}, {B}, {C}, {D}, {E, F}, {G}, {H, I}. Поясним этот выбор.

A, B, C, D – каждый из этих субъектов имеет уникальный набор связей с другими, так что они образуют три класса, каждый из которых с одним членом. E и F – каждый из них имеет одну связь (с актором *B*), имеют точно такую же картину связей со всеми другими участниками, они структурно эквивалентны. Актор G – класс сам по себе, его профиль связей с другими узлами в диаграмме является уникальным. H и I подобны E и F.

Автоморфная эквивалентность (параллельные структуры): {A}, {B, D}, {C}, {G}, {E, F, H, I}.

Для пояснения предположим, что на рисунке описана франшиза сети ресторанов гамбургеров. Актор A – центральная штаб–квартира, акторы B, C, D – менеджеры трех различных магазинов. Акторы E и F являются рабочими в одном магазине; G – рабочим во втором магазине; H и I – рабочими в третьем магазине. Акторы B и D не являются структурно эквивалентными (они имеют одного и того же хозяина, но не одних и тех же рабочих), они «эквивалентны» в ином смысле. Оба менеджера B и D отчитываются перед боссом (A), и каждый из них имеет ровно двух рабочих. Если их поменять местами, а также сменить четырех рабочих, то все расстояния между всеми участниками в графе будут абсолютно идентичны. Поэтому акторы B и D образуют автоморфный класс эквивалентности.

Регулярная эквивалентность (идентичные модели связей с другими классами): {A}, {E, F, G, H, I},{B, C, D}.

Самый простой класс – пять акторов по всей нижней части диаграммы (E, F, G, H, I). Эти акторы регулярно эквивалентны друг другу, так как:
1. они не имеют связь с любым актором в первом классе (с актором A);
2. каждый из них имеет связь с актором во втором классе (B, C или D).

Акторы B, C, и D образуют класс, потому что:
1. каждый из них имеет связь с членом первого класса (с актором A);
2. каждый из них имеет связь с членом третьего класса. B и D на самом деле имеют связи с двумя членами третьего класса, в то время как актор C имеет связь с только одним членом третьего класса; это не имеет значения, поскольку есть связь с каким–то членом третьего класса.

Актор A – класс сам по себе, так как он:
1. имеет связь с членом второго класса;
2. не имеет связь с членом третьего класса.

Различия между различными типами эквивалентности можно увидеть на рис. 32.

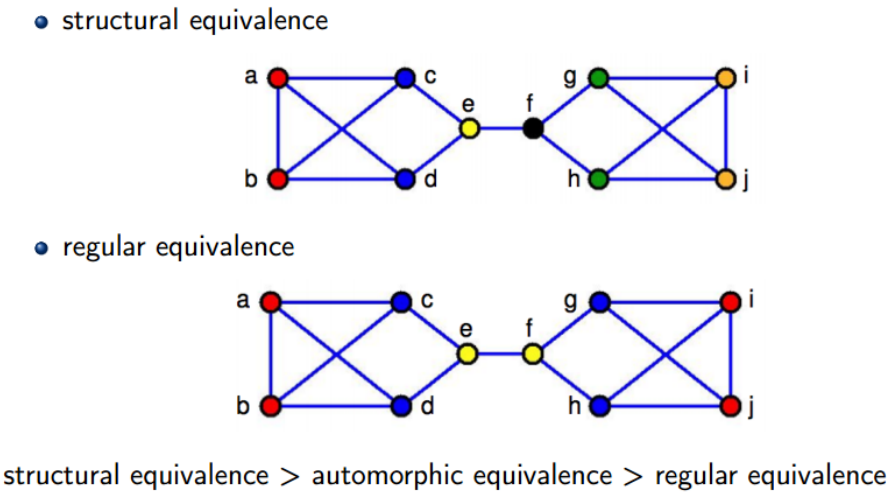
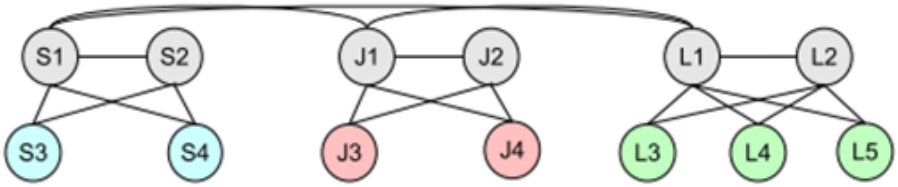


Рис. 32. Графическое представление структурной и регулярной эквивалентностей

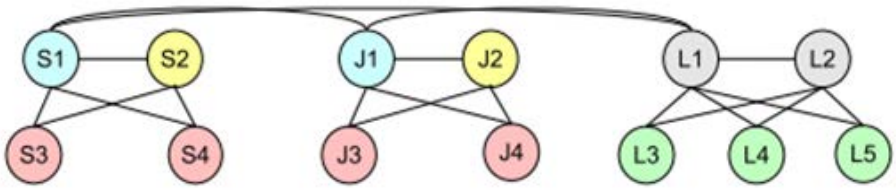
Еще один пример представлен на рис. 33. Рассмотрим расширенную семью. S1, J1 и L1 являются братьями и сестрами, S2, J2 и L2 являются супругами, а остальные узлы – их дети. Дети каждой семьи – {S3, S4}, {J3, J4} и {L3, L4, L5} – образуют нетривиальный класс эквивалентности. Узлы с тем же цветом находятся в том же классе, за исключением того, что серые узлы представляют одноэлементные классы. Каждый серый узел является собственным классом. Эта модель является слишком строгой.

Рассмотрим три семьи –Смит {S1, S2, S3, S4}, Джонс {J1, J2, J3, J4}, Ли {L1, L2, L3, L4, L5}. Различить семьи Смит и Джонс невозможно, а семейство Ли отличается, потому что в нем трое детей вместо двух. Следовательно, классы автоморфной эквивалентности {S1, J1}, {S2, J2}, {S3, S4, J3, J4}, {L1}, {L2} и {L3, L4, L5}.

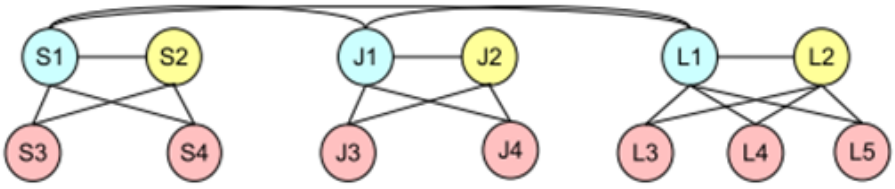
Однако при использовании регулярной эквивалентности все три семьи могут быть эквивалентными, при этом возникают три класса эквивалентности: родной брат – родитель {S1, J1, L1}, супруг – родитель {S2, J2, L2} и дети.



а



б

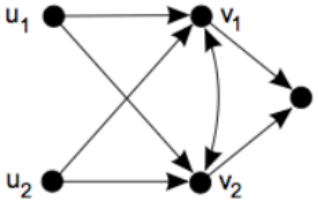


в

Рис. 33. Сравнение эквивалентностей:
а – структурная, б – автоморфная, в – регулярная

Две вершины структурно эквивалентны, если они имеют одинаковые отношения со всеми другими элементами сети. Строки и столбцы матрицы смежности структурно эквивалентных узлов идентичны, «соединяются с

теми же соседями». На рис. 34 видно, что вершины U_1 и U_2 эквивалентны, вершины V_1 и V_2 тоже эквиваленты.



	u1	u2	v1	v2	w
u1	0	0	1	1	0
u2	0	0	1	1	0
v1	0	0	0	1	1
v2	0	0	1	0	1
w	0	0	0	0	0

Рис. 34. Граф и матрица смежности

4.2. ХАРАКТЕРИСТИКИ ЭКВИВАЛЕНТНОСТИ

Евклидово расстояние. Пусть A_{ik} – количество связей между акторами n_i и n_k. Определим дистанцию структурной эквивалентности для акторов n_i и n_j как евклидово расстояние связей между этими акторами. Для акторов n_i и n_j это будет суммирование различий между i-и и j-й строками и i-м и j-м столбцами матрицы смежности графа:

$$d(v_i, v_j) = \sqrt{\sum_k ((A_{ik} - A_{jk})^2 + (A_{ki} - A_{kj})^2)}$$

Если акторы n_i и n_j структурно эквивалентны, то соответствующие строки и столбцы матрицы смежности будут равны между собой, а евклидово расстояние будет равно 0. Если же они абсолютно неэквивалентны, то и величина евклидова расстояния станет большой.

Расстояние Хэмминга – число позиций, где векторы различны:

$$h(v_i, v_j) = \sum_k |A_{ik} - A_{jk}|$$

Максимально возможное расстояние Хэмминга:

$$max(d_{ij}^2) = k_i + k_j$$

Нормализованное расстояние Хэмминга:

$$d_{ij\,N}^2 = \frac{d_{ij}^2}{k_i + k_j} = \frac{\sum_k (A_{ik}^2 - 2A_{ik}A_{jk} + A_{jk}^2)}{k_i + k_j} = 1 - \frac{2n_{ij}}{k_i + k_j}$$

Меры подобия (similarity measures)

– Косинусная мера измеряется как расстояние между векторами в n-мерном пространстве:

$$\sigma(v_i, v_j) = cos(\theta_{ij}) = \frac{v_i v_j}{||v_i||||v_j||} = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{\sum A_{ik}A_{ki}}\sqrt{\sum A_{jk}A_{kj}}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

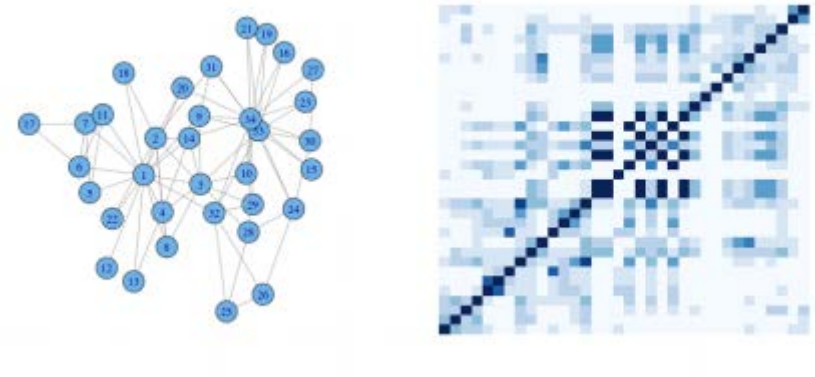
– Мера Жаккара

$$J(v_i, v_j) = \frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|}{|\mathcal{N}(v_i) \cup \mathcal{N}(v_j)|}$$

– Коэффициент корреляции Пирсона

$$r_{ij} = \frac{\sum_k (A_{ik} - \langle A_i \rangle)(A_{jk} - \langle A_j \rangle)}{\sqrt{\sum_k (A_{ik} - \langle A_i \rangle)^2} \sqrt{\sum_k (A_{jk} - \langle A_j \rangle)^2}} = \frac{n_{ij} - \frac{k_i k_j}{n}}{\sqrt{k_i - \frac{k_i^2}{n}} \sqrt{k_j - \frac{k_j^2}{n}}}$$

На рис. 35 представлены граф и его матрица подобия.



а б

Рис. 35. Граф (а) и его матрица подобия (б)

Изоморфность. Два графа G и H называются изоморфными (рис. 36), если существует биекция (взаимно однозначное отображение одного множество в другое) такая, что

ψ: V(G) → V(H), ∀u, v ∈ V(G) (u, v) ∈ E(G) ⇔ (ϕ(u), ϕ(v)) ∈ E(H).

Другими словами, два графа изоморфны, если существует взаимно однозначное отображение вершин и для каждого ребра в одном графе есть единственное ребро в другом графе между соответствующими вершинами (оба графа имеют «одинаковую структуру»).

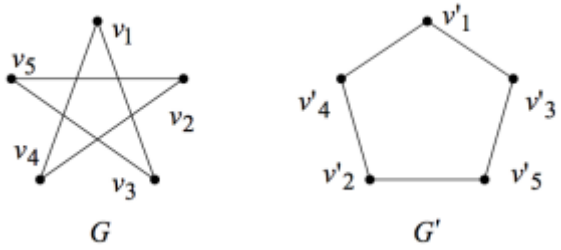


Рис. 36. Изоморфные графы

Если изоморфизм ψ отображает граф G сам на себя, то такое отображение называется автоморфизмом (рис. 37). Автоморфизм – это такое взаимно однозначное отображение вершин, при котором для каждого ребра на графе имеется уникальное ребро между соответствующими отображенными вершинами.

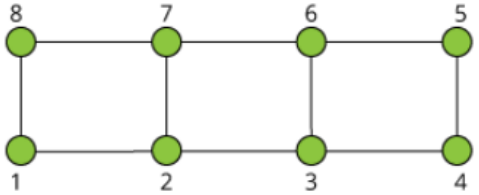


Рис. 37. Автоморфизм графа

Ассортативное смешивание (assortative mixing), или гомофилия – тенденция узлов сети формировать связи с аналогичными узлами.

Термин возник в социологии, в частности, при изучении закономерностей формирования супружеских пар. Социологические исследования показали, что сети друзей также формируются на основе общего языка, расы, возраста, уровня образования и доходов и т.п. В таких случаях говорят, что соответствующие сети обладают свойствами ассортативного смешивания.

Ассортативное смешивание по степени – это принцип, согласно которому вершины высокой степени соединены с аналогичными вершинами высокой связанности, а вершины низкой степени соединены с другими менее связанными вершинами.

В сети возможна ситуация, когда узлы, имеющие большую степень («звезды»), преимущественно связаны с узлами, имеющими большую степень. Иными словами, «звезды» предпочитают быть связанными со «звездами». Такие сети называют ассортативными. Возможна также обратная ситуация: «звезды» связаны с другими «звездами» через цепочки узлов, имеющих малое число соседей. Такие сети называют дисассортативными.

Коэффициент ассортативности r – коэффициент корреляции Пирсона между степенью соседних узлов. По определению, $-1 \le r \le 1$. Для ассортативных сетей $r > 0$, для дисассортативных сетей $r < 0$. Ассортативное смешивание по степени узла находится как

$$AM(G) = \frac{M \sum_{e \in E} k_1^e k_2^e - [\sum_{e \in E} k_1^e]^2}{M \sum_{e \in E} (k_1^e)^2 - [\sum_{e \in E} k_1^e]^2},$$

где G – исследуемый граф, E – множество всех ребер графа, k_2^e – количество ребер графа, e – ребро графа, k_1^e – количество связей первой вершины из двух концов ребра e, k_2^e - количество связей второй вершины из двух концов ребра e.

Например, супруги в браке обычно имеют приблизительно один возраст, уровень образования, этнос и т.д. Ассортативность проявляется также в том, что формируются сообщества (кластеры) людей, связанных по увлечениям и интересам (коллекционеры, книголюбы) и профессиям. Знаменитые люди обычно знакомы друг с другом. Для богатых эта закономерность известна как «явление элитарного клуба».

В таблице 2 приведены значения коэффициента Пирсона (коэффициента ассортативности) для реальных сетей различной природы.

Тип сети	Сеть	Размер сети	Ассортативность r
Социальные	соавторов по физике	52 909	0,363
	соавторов по биологии	1 520 251	0,127
	соавторов по математике	253 339	0,120
	сотрудничества актеров кино	449 913	0,208
	директоров компаний	7 673	0,276
	адресов электронной почты	16 881	0,092
Технологические	сеть электростанций	4 941	–0,003
	Интернет	10 697	–0,189
	«всемирная паутина» (WWW)	269 504	–0,067
Биологические	взаимодействий белков	2 115	–0,156
	метаболическая сеть	765	–0,240
	нейронная сеть	307	–0,226
	морская пищевая сеть	134	–0,263
	пресноводная пищевая сеть	92	–0,326
	сеть сообщества дельфинов	62	–0,044

Таблица 2. Значения коэффициента ассортативности для реальных сетей

5. СЕТЕВЫЕ СООБЩЕСТВА

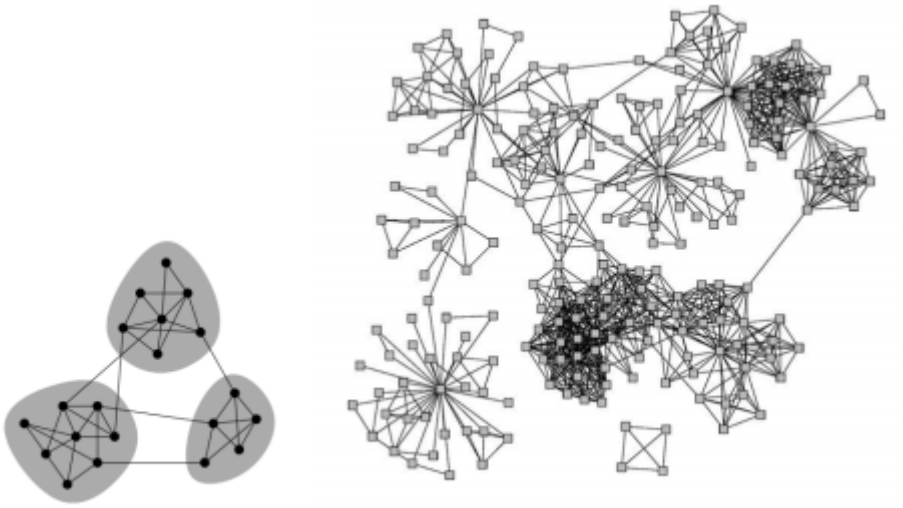
5.1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Сетевые сообщества – это довольно большая, важная и активно разрабатываемая тема.

Четкого математического определения сетевого сообщества не существует. Обычно про сообщество говорят как про некоторую группу узлов или некоторую группу людей, которые связаны внутри группы сильнее между собой, чем со всем остальным миром. Однако четкий математический порог здесь отсутствует.

Часто также используется термин «кластер» – кластер узлов (пользователей) в графе.

Несмотря на отсутствие четкой формулировки того, что такое сообщества, во многих работах предлагаются различные модифицированные определения, которые позволяют, в свою очередь, уточнить основные определения и таким образом ввести параметр, который позволяет определить, что является сообществом, а что нет.



а б

Рис. 38. Разрезы графа: идеальный случай (а), реальный случай (б)

Можно говорить о нескольких путях нахождения сообществ. В первую очередь посмотрим на рис. 38, а, на котором присутствуют 3 сообщества. Мы можем их отделить друг от друга. Это называется разрезом графа. Достаточно очевидно, что тут мы разделяем граф на 3 части.

В реальных графах (рис. 38, б) это не совсем так. Тут даже не совсем понятно, что является сообществом, а что – нет. Например, отстоящий кусок – это сообщество, но остальной граф – связный, и непонятно, как в нем

выделить сообщества. То есть в реальных графах может возникать много проблем с содержательным определением сообществ.

Возникают и дополнительные проблемы. Во-первых, можно рассматривать разные задачи – разбиение графа на части (сообщества) или выделение одного конкретного сообщества из графа. Во-вторых, можно рассматривать задачу в сильной постановке, когда сообщества считаются неперекрывающимися (узел принадлежит только одному сообществу) или ослабить эту формулировку. В последнем случае задача усложнится.

5.2. РАЗДЕЛЕНИЕ ГРАФА НА ЧАСТИ

Интуитивно понятно, что при выделении сообщества мы сравниваем среднюю плотность связей внутри кластера и между кластерами, т.е. сообщество – это та часть графа, в которой средняя плотность между узлами превышает плотность связей между сообществами. Другими словами, внутри сообщества связей должно быть больше, чем между сообществами. Запишем это в математической форме, через плотность связей в графе.

Как известно, плотность связей в графе задается формулой

$$\rho = \frac{m}{\left| n(n-1)/2 \right|},$$

где m – количество ребер в графе, n – количество узлов в графе, $n\,(n-1)$ – максимальное количество ребер в графе. Тогда для какого-то сообщества плотность связей равна количеству ребер в этом сообществе m_c, деленному на максимально возможное количество ребер в сообществе:

$$\delta_{\text{int}}(C) = \frac{m_c}{\left| n_c(n_c-1)/2 \right|},$$

и внешняя плотность вычисляется аналогично:

$$\delta_{\text{ext}}(C) = \frac{m_{ext}}{\left| n_c(n_c-1)/2 \right|},$$

где m_{ext} – полное число внешних связей.

Для выделения сообщества плотность внутренних связей должна быть больше плотности внешних связей, при этом плотность внутренних связей должна быть больше, чем средняя плотность графа, а плотность внешних связей должна быть меньше, чем средняя плотность графа. Таким образом, получаем следующие условия:

$$\delta_{\text{int}} > \rho, \quad \delta_{\text{ext}} < \rho.$$

Будем находить кластеры путем анализирования максимума разницы внутренних и внешних плотностей, т.е. введем математический параметр

$$\max(\delta_{\text{int}} - \delta_{\text{ext}}).$$

Однако возникает проблема: для максимизации такой формулы нужно вначале выделить необходимое сообщество, и при «лобовом» решении

получится факториально большое количество таких делений. На практике такая максимизация невозможна, и требуется подобрать подходящую эвристику.

Для этого, во-первых, введем ограничение по типам графов: мы смотрим только на разреженные графы (плотность очень маленькая), для которых

$m<<n^2$.

Это реальное ограничение: большинство сетей так и устроены. Во-вторых, будем считать, что все сообщества – связные, так как задача нахождения несвязных сообществ тривиальна. Таким образом, задача свелась к нахождению компонентов связности в разреженном графе. Для этого нужно выделить часть узлов, которая удовлетворяет условиям, наложенным нами на сообщества, т.е. ввести критерий и выбрать для него метод оптимизации.

Решение этой задачи – NP-сложное. Действительно, если граф разбивается на 2 части, то каждый узел может принадлежать либо одной, либо другой части. Количество возможных разбиений очень велико даже в этом случае. Если делить его на большее количество частей, то задача становится еще более сложной. Поэтому, как любой NP-сложный алгоритм, такая задача решается приближенно:

- используются эвристики (нет гарантии схождения, но на практике это работает);
- используется жадный алгоритм (нет гарантии нахождения глобального минимума);
- приближенно ищется глобальный минимум (это задача комбинаторной оптимизации);
- используется рекурсивное разбиение графа на 2 части.
- выполняется сбалансированное разбиение (аналогично задаче загрузки процессоров при высоконагруженых вычислениях).

Критерий позволяет выбрать конкретную постановку задачи.

5.3. РАЗРЕЗЫ В ГРАФЕ

Один из оптимизационных критериев основан на разрезах графа на части. Мы можем сделать разрез графа и поделить его на 2 части. Каждую из частей мы также можем поделить пополам и получить 4 сообщества, и т.д. При этом мы используем минимальный разрез.

Разрез графа на части – минимальное количество ребер, которое мы можем удалить, чтобы разделить граф на части. Формулировка разреза графа на части проста, однако нужно понимать, что мы не хотим удалять части, маленькие сами по себе. Для этого мы используем балансировку.

Какие варианты разреза можно использовать?

- минимальный разрез

$$Q = cut(V_1, V_2) = \sum_{i \in V_1,\, j \in V_2} e_{ij}\,.$$

В этом случае можно отрезать очень мало узлов, так как нет условия балансировки.

– квотированный разрез (quotent cut)

$$Q = \frac{cut(V_1, V_2)}{\|V_1\|} + \frac{cut(V_1, V_2)}{\|V_2\|}.$$

Идея в том, что мы берем разрез и нормируем его на размеры половинок графа (число узлов в них).

– нормализованный разрез

$$Q = \frac{cut(V_1, V_2)}{\sum_{i \in V_1,\, j \in V} e_{ij}} + \frac{cut(V_1, V_2)}{\sum_{i \in V_2,\, j \in V} e_{ij}}$$

– более сильный вариант. Мы берем тот же самый разрез и нормируем его на количество ребер, выходящих из каждой из сторон графа.

5.4. РАЗДЕЛЕНИЕ ГРАФА НА ОСНОВЕ МОДУЛЯРНОСТИ

Другой набор метрик, который стал популярен в последнее время – это метрики, которые связаны с соотношением плотности внутри сообщества с внешней плотностью. Они опираются на понятие модулярности.

Идея заключается в следующем. Пусть граф можно разбить на классы (подграфы), и у каждого класса (и узла, принадлежащего классу) есть своя метка (например, разные цвета). Тогда, чем больше отличается подграф, соответствующий сообществу, от случайного подграфа, тем лучше разбиение.

Вводится понятие модулярности. Модулярность – это скалярная величина из отрезка [-1, 1], которая показывает то, насколько при заданном разбиении графа на сообщества плотность связей внутри сообществ больше плотности связей между сообществами, т.е. определяет качество разбиения сети на сообщества. Для вычисления модулярности мы вычисляем связи между узлами, принадлежащими разным кластерам, и узлами одного кластера:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta\left(c_i, c_j\right).$$

Здесь A – матрица смежности графа (число ребер из вершины i в вершину j; A_{ij} – элемент матрицы; d_i – степень i-й вершины графа (количество связей у i-й вершины); C_i – метка вершины (номер сообщества, к которому относится вершина); m – общее количество ребер в графе;

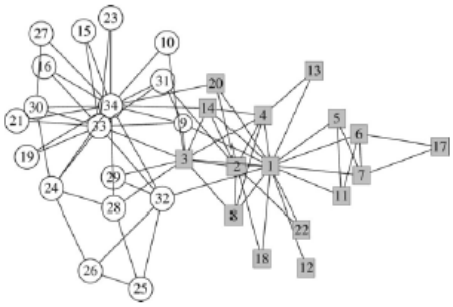
$\delta\left(C_i, C_j\right)$ – дельта-функция (равна единице при $C_i = C_j$ и нулю в других случаях), т.е. равна 1, если узлы принадлежат одному кластеру, и 0 в противном случае. В этой формуле просчитывается количество связей, находящихся внутри одного кластера, и потом они складываются.

Обратим внимание, что здесь также присутствует член $\frac{k_i k_j}{2m}$. Зачем нужен этот фактор? Идея в том, что мы хотим сравнить данное разбиение на сообщества со случайным графом: если у нас есть случайный граф, то у нас нет хороших кластеров. Таким образом, плотность в таком графе можно использовать как относительную метрику: у сообщества плотность должна быть строго больше, чем в случайном графе.

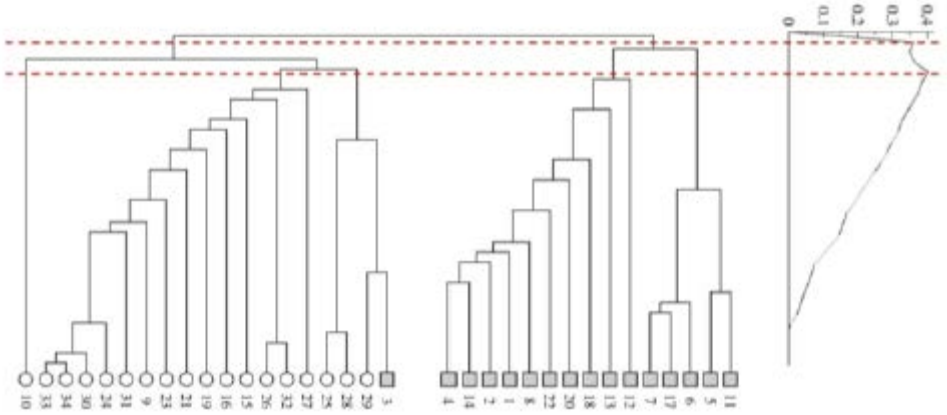
Если имеется узел со степенью $k(i)$, то к нему присоединено $k(i)$ ребер. Вероятность того, что произвольное ребро присоединено к узлу, равна $k(j) / 2m$, а вероятность связи между ребрами равна $\frac{k_i k_j}{2m}$. Иначе говоря, мы вычисляем разницу между реальным количеством ребер и ожидаемым. При этом сумма не равна нулю, если мы работаем в пределах одного класса.

Таким образом, при хорошем разбиении на классы модулярность высокая, если класс один, то модулярность равна 0.

Рассмотрим пример на рис. 39.



а



б

Рис. 39. Граф (а) и его дендрограмма (б)

Для графа (рис. 39, а) производится разбиение на сообщества рекурсивным образом. На рис. 39, б, представлена дендрограмма. Как мы видим, в какой-то момент произошло разбиение даже на 1 узел, что уже явно излишне. Для ответа на вопрос, когда остановить рекурсию, ищем максимум модулярности (рис. б справа). Как известно, вопрос о количестве кластеров в графе всегда является неформализуемым и обычно решается некоторой метрикой, в качестве которой в нашем случае выбрана модулярность.

Отметим, что модулярность можно не только использовать как критерий, но и просто отслеживать как показатель развития сообщества.

Приведем сводку наиболее распространенных подходов к разбиению графов:

- Алгоритмы разреза графов (Graph cut algorithms)
  - Kernighan–Lin
  - Spectral normalized cuts
  - s-t flow
  - Multilevel graph partitioning
- Алгоритмы, основанные на модулярности
  - Жадные (Greedy)
  - Спектральная максимизация модулярности (Spectral modularity maximization)
- Эвристические алгоритмы
  - Edge betweenness
  - Случайные блуждания (Random walks)

Широкое распространение получил спектральный алгоритм оптимизации модулярности, который основан на вычислении максимального собственного значения матрицы модулярности:

Algorithm: Spectral moduldarity maximization: two-way parition

Input: adjacency matrix $\mathbf{A}$

Output: class indicator vector $\mathbf{s}$

compute $\mathbf{k} = deg(\mathbf{A})$;

compute $\mathbf{B} = \mathbf{A} - \frac{1}{2m}\mathbf{k}\mathbf{k}^T$;

solve for maximal eigenvector $\mathbf{B}\mathbf{x} = \lambda\mathbf{x}$;

set $\mathbf{s} = sign(\mathbf{x}_1)$

Recursive bisection

На рис. 40 показаны результаты работы алгоритма.

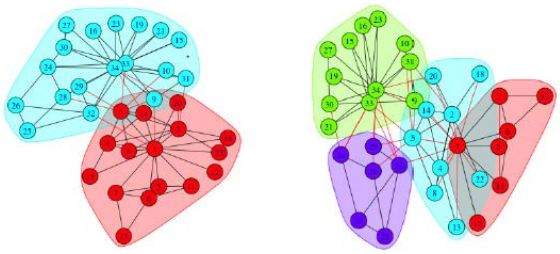


Рис. 40. Разделение графа с помощью спектрального алгоритма оптимизации модулярности

Еще один алгоритм выделения сообществ относится к разряду эвристических алгоритмов и отличается идеологической простотой. Он основывается на понятии edge betweenness – количество кратчайших путей, проходящих через ребро:

$$C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}.$$

Идея алгоритма заключается в том, что ребра, через которые проходит наибольшее количество кратчайших путей, являются «мостами» между кластерами в графе. Поэтому для этих ребер описанная метрика будет высокой. Мы можем найти эти ребра, удалить их, и тогда граф распадется на сообщества.

В ходе выполнения алгоритма вычисляем edge betweenness, удаляем ребро с максимальным значением этой метрики. Каждый раз метрика должна быть пересчитана, так как при удалении ребер она может меняться. Делаем это до тех пор, пока в графе остаются ребра. Если хотим поделить граф на 2 части, то останавливаем процесс в тот момент, когда у нас есть 2 компоненты. Несмотря на то, что алгоритм жадный, он очень интуитивен и дает неплохие результаты.

Пример представлен на рис. 41. При первом разбиении отделятся красные и оранжевые, потом отделятся все зеленые и т.д.

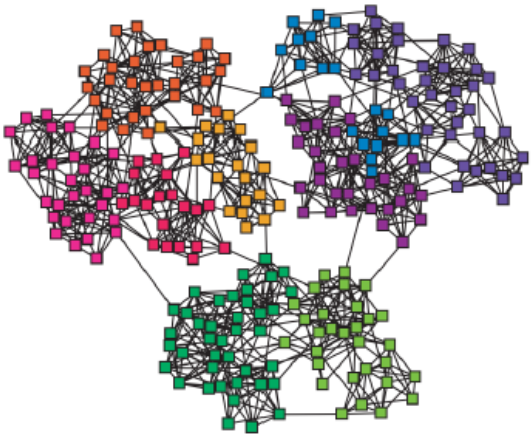


Рис. 41. Разделение графа с помощью алгоритма edge betweenness

Алгоритм имеет и недостатки: для не очень больших графов он может работать не очень хорошо, а также отличается сложностью вычислений.

6. АЛГОРИТМЫ РАЗБИЕНИЯ ГРАФА

Для рассмотрения алгоритмов разбиения графа напомним основные понятия.

Разрез графа – разделение узлов графа на два непересекающихся подмножества.

Размер разреза – количество ребер, которые необходимо убрать для разделения графа.

Минимальный разрез – наименьший возможный разрез графа.

Если в графе нет длинных одиночных цепочек и нет одиночных узлов, то как правило, минимальный разрез дает неплохое представление о том, как граф может быть разделен на части (рис. 42).

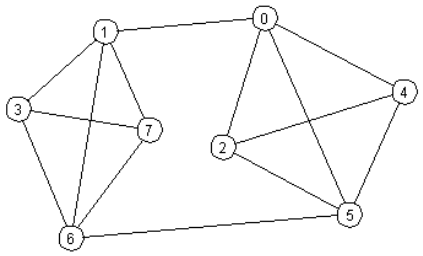


Рис. 42. Минимальный разрез графа

6.1. АЛГОРИТМ СЛУЧАЙНОГО МИНИМАЛЬНОГО РАЗРЕЗА

Случайность алгоритма заключается в том, что при новых исполнениях алгоритма результат может быть разным. Соответственно, мы можем получить как правильный ответ, так и нет, но мы получаем вероятностные границы того, что полученный ответ верный. Многократными итерациями выполнения алгоритма вероятность ошибки уменьшается.

Алгоритм использует единственную операцию соединения (схлопывания) двух узлов (ребра), при которой пара вершин заменяется на одну. (рис. 43). Такая процедура, если довести ее до логического завершения, приведет к тому, что останется два узла и некоторое количество ребер между ними. И это число ребер будет приближением к минимальному разрезу.

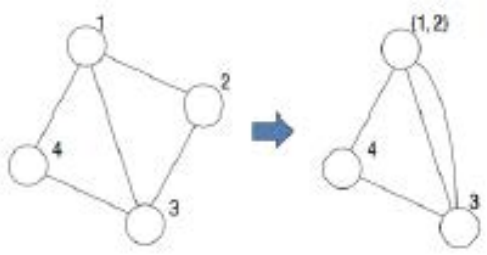


Рис. 43. Схлопывание ребер

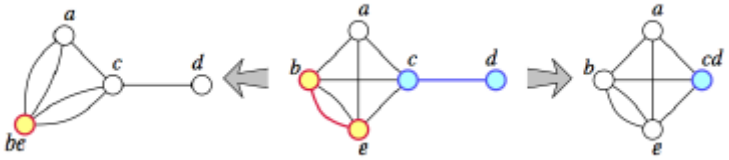


Рис. 44. Развитие минимального разреза

Рассмотрим граф посередине (рис. 44). Мы видим, что минимальный разрез приходится между узлами *c* и *d*. Пусть случайным образом выбирается ребро *be*, и в результате этого два узла *b* и *e* схлопываются. Остается один узел *be* и получается левый граф, при этом величина минимального разреза в графе не изменилась.

Пусть теперь случайным образом выбраны узлы *c* и *d*, по той же схеме получаем правый граф. Как мы видим, в исходном графе минимальный разрез шел по ребру *cd*, и его величина была равна 1. В новом графе этот разрез исчез, и минимальный разрез уже будет равен трем. В данном случае минимальный разрез после схлопывания увеличился, в первом же варианте минимальный разрез не изменился.

Можно показать, что схлопывание ребер не может уменьшить величину минимального разреза, т. е. минимальный разрез модифицированного графа будет больше или равен минимальному разреза исходного графа.

Рассмотрим еще один пример (рис. 45). Минимальный разрез исходного графа равен 2. На первом шаге выберем ребро bf, на втором шаге gh, затем dgh, ae, abef и cdgh. Получился граф из двух наборов узлов, и между ними происходит разрез. В данном случае минимальный разрез совпал с минимальным размером исходного графа.

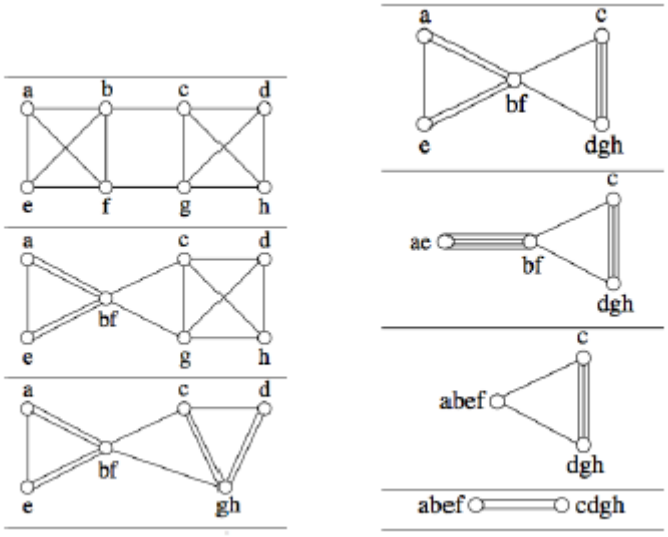


Рис. 45. Развитие минимального разреза

Оценим вероятность того, что полученный разрез будет минимальным. Вероятность успешного решения, т.е. того, что на всех шагах выбирались ребра не из минимального разреза, равна

$$P(success) \geq \frac{2}{n(n-1)}$$

Неуспех алгоритма означает, что в ходе поиска минимального разреза схлопнули ребро, которое к этому разрезу относится. Вероятность неуспеха высчитывается как

$$P(error) \leq \left(1 - \frac{2}{n(n-1)}\right)^{n^2/2} \sim \frac{1}{e} = 0.37$$

если мы проведем большое количество испытаний, то максимально возможная вероятность ошибки составляет 37%. Если использовать еще большее количество итераций, то вероятность ошибки можно еще больше уменьшить:

$$P(error) \leq \frac{1}{n^c}$$

.

Пример работы алгоритма показан на рис. 46.

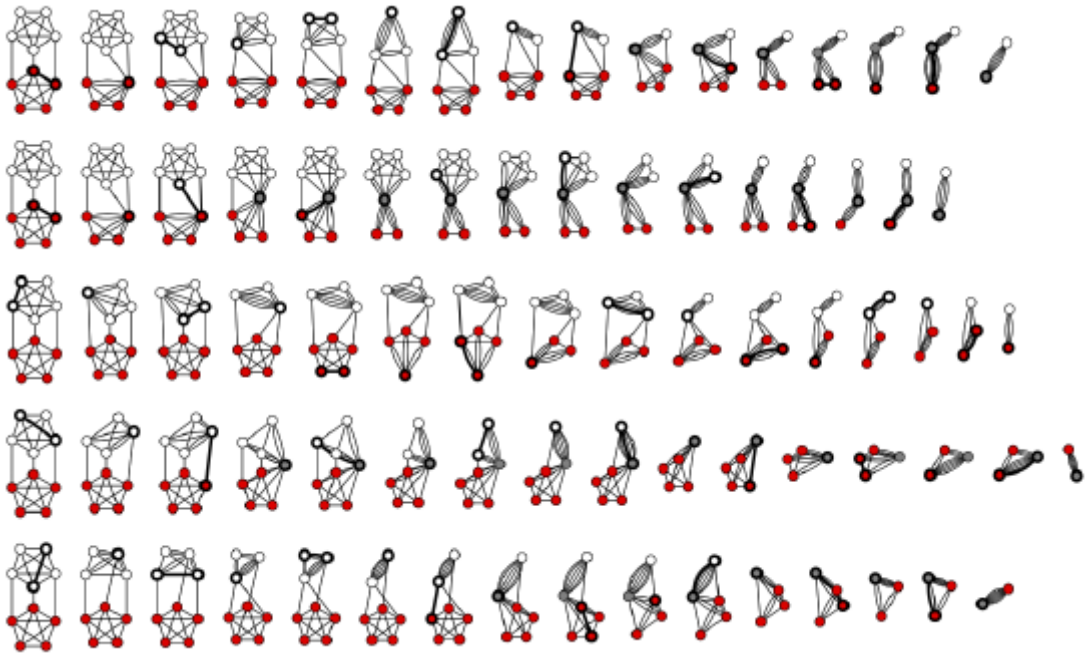


Рис. 46. Последовательность работы алгоритма случайного минимального разреза

6.2. АЛГОРИТМ МНОГОУРОВНЕВОГО РАЗБИЕНИЯ ГРАФА

Алгоритм многоуровневого разбиения графа предложен в 1998 г. Дж. Каруписом и его командой. Алгоритм является чисто эвристическим и не имеет каких-либо оценок. Идея алгоритма иллюстрируется рис. 47.

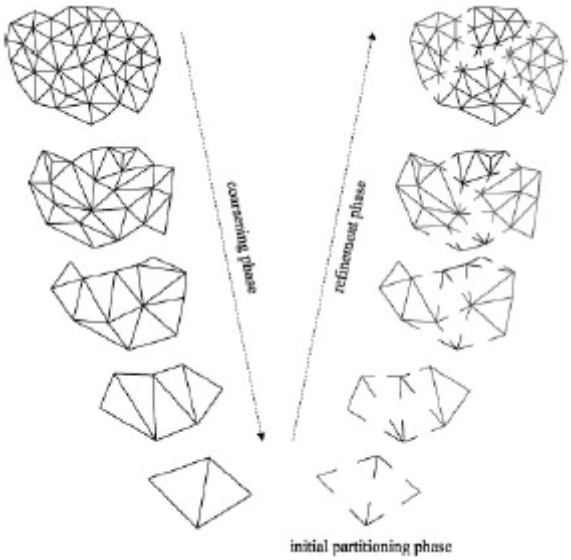


Рис. 47. Алгоритм многоуровневого разбиения графа

Идея схожа с алгоритмом случайного разбиения и состоит из двух фаз: опуститься вниз и подняться наверх. «Опуститься вниз» означает, что мы берем граф и производим процедуру огрубления графа. Процедура аналогична схлопыванию ребер, и в конечном итоге получается некоторый граф. Однако мы не доходим до состояния, когда остается только два узла, а производим огрубление, т.е. останавливаемся на некотором небольшом размере, на котором уже можем найти разрез без каких-либо сложностей.

После нахождения разреза мы его распространяем обратно, восстанавливая граф, и на каждом шаге восстановления пытаемся провести локальную оптимизационную процедуру, уточняя расположение этого разреза – сдвигая узлы, которые лежат вблизи разреза, на одну или другую сторону, можно попробовать улучшить качество разреза.

Огрубление (coarsening) производится с помощью процедуры matching. Matching – это набор независимых ребер (не имеют общих узлов). На рис. 48 они выделены красным цветом. Максимальный matching – состояние, когда невозможно добавить еще одно ребро, но не подразумевается, что выбрано наиболее оптимальное расположение ребер.

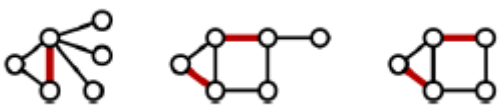


Рис. 48. Matching

Для matching’а можно брать узлы, узлы и их окружение или маленькие кластеры. Если есть веса на ребрах, можно учитывать и их.

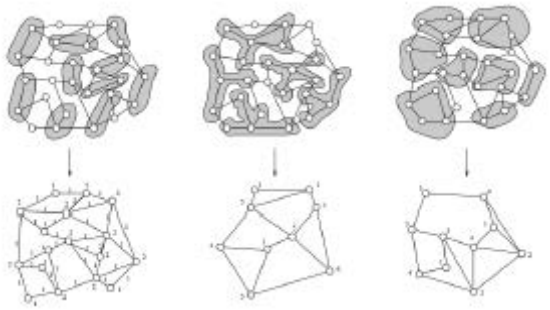


Рис. 12. Варианты matching’а

6.3. АЛГОРИТМ НАХОЖДЕНИЯ ЛОКАЛЬНЫХ КЛАСТЕРОВ

Алгоритм нахождения локальных кластеров (рис. 49) используется в том случае, если мы не хотим разделять весь граф на части, а, например, нужно выделить только один кластер.

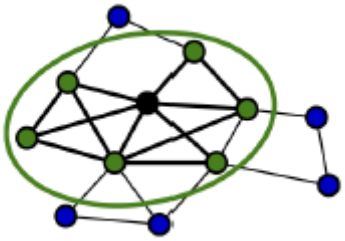


Рис. 49. Выделение локальных кластеров

В этом алгоритме используется метрика проводимости. Проводимость (conductance) – это вероятность того, что мы выбираем ребро из множества всех ребер, принадлежащих данному набору узлов *S*, и это ребро пересекает разрез:

$$\phi(S) = \frac{cut(S, V \backslash S)}{\min(vol(S), vol(S \backslash V))}$$

Также это показатель того, что случайное блуждание, начатое в графе, за один шаг уйдет из кластера. Чем меньше проводимость, тем меньше вероятность, что при случайном блуждании мы выйдем из кластера.

Для работы с алгоритмом нужен начальный граф, начальный узел и значение проводимости.

Выполняется фиксированное количество итераций. На каждой итерации начальный вектор умножается на матрицу переходов, которая представляет собой несколько измененную матрицу смежности. Изменение заключается в нормировании вероятности выхождения за узел; кроме того, добавлена вероятность того, что можно либо пойти вперед, либо остаться в этом же узле (нижняя формула).

На каждом шаге алгоритма вычисляется вероятность распределения по новым узлам, затем сравнивается со степенью узла, и если эта вероятность больше какого-то критического значения, то она сохраняется, если меньше, то обращается в ноль.

После вычисления вероятностей распределения производится сортировка узлов по вероятностям. Затем, начиная сверху, производится вычисление проводимости между i-ым количеством узлов и всем остальным графом. Если в какой-то момент находится разрез, который удовлетворяет всем критериям, то получаем искомое сообщество.

Algorithm: Nibble

Input: Graph $G, q_0(v_0), \phi_0$

Output: Graph partition S

for $t = 1 : t_{last}$ **do**

$q_t = Mr_{t-1}$;

$r_t(i) = q_t(i)$ *if* $q_t(i)/d(i) > \epsilon$, *else* 0;

sort i descending based on $r_{t_{last}}(i)/d(i)$;

sweep from top $\phi(S\{i = 1..j\}) < \phi_0$ or $\phi(S\{i = j + 1..n\}) < \phi_0$;

- Random walk:

$$M = (AD^{-1} + I)/2, \quad D = diag(d(i))$$

Сложность алгоритма пропорциональна размеру разреза, который мы находим. Пример работы алгоритма изображен на рис. 50.

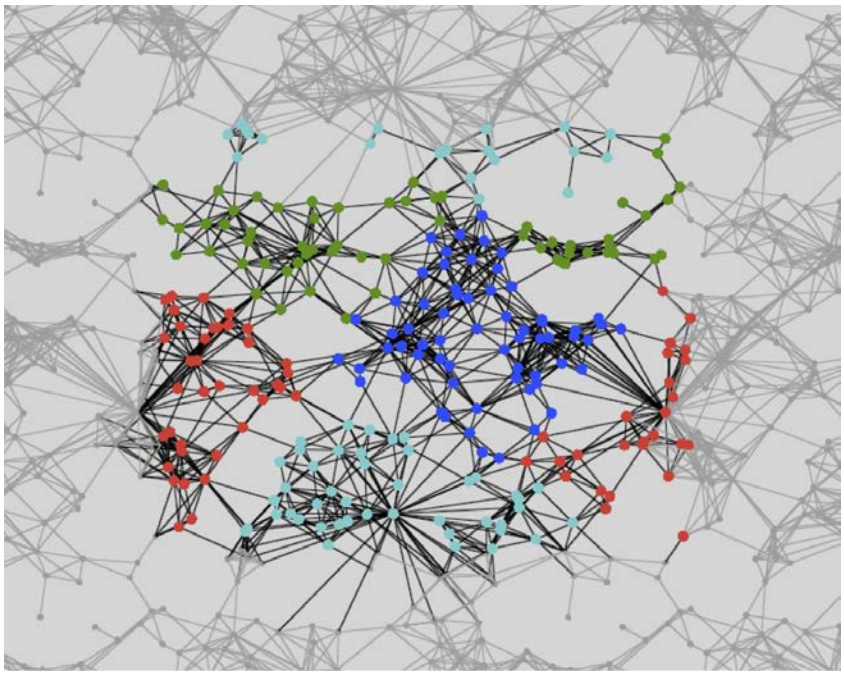


Рис. 50. Пример работы алгоритма нахождения локальных кластеров со случайными блужданиями

7. СЕТЕВЫЕ СТРУКТУРЫ

7.1. НАХОЖДЕНИЕ СООБЩЕСТВ КАК ХАРАКТЕРИСТИК СТРУКТУРЫ СЕТИ

Ранее были рассмотрены алгоритмы для нахождения сообществ. Были выделены несколько крупных классов алгоритмов для нахождения сообществ.

Один из таких классов – методы глобальной оптимизации или алгоритмы разделения. Идея заключается в том, что есть большой граф, и мы его пытаемся разделить на части с помощью различных метрик – минимального разреза, проводимости и др. – которые нужно оптимизировать.

Другой класс основывается на идее модулярности – функции, сравнивающей плотность связей внутри графа с плотностью связей в случайном графе. Кластером в этом случае является область, где средняя плотность выше, чем в случайном графе.

Рассматривались также эвристические методы, в том числе edge betweenness, при котором убираются отдельные ребра графа с наибольшим edge betweenness, предполагая, что такие ребра являются мостами.

Глобальные методы можно использовать для получения плоской кластеризации или иерархической кластеризации.

На рис. 51 показан пример иерархической кластеризации. Мы видим, что на разных этапах мы получаем сообщества разного размера. На основании этого строится дендрограмма – схема, показывающая разбиение на сообщества. На различных этапах кластеризации можно использовать различные алгоритмы.

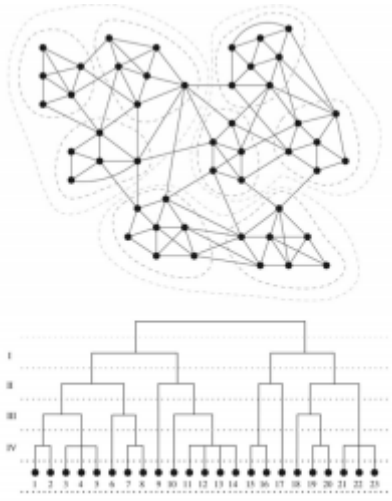


Рис. 51. Иерархическая кластеризация

Еще один класс алгоритмов кластеризации основан на идее, что сообщество – это группа узлов, которые похожи между собой. В качестве

признака «похожести» можно использовать структурную эквивалентность: узлы структурно эквивалентны, если они связаны с одними и теми же узлами. Такие алгоритмы называются blockmodeling.

Но полностью структурно эквивалентных узлов бывает очень мало. Поэтому рассчитывают примерную эквивалентность, для чего используются типовые метрики схожести (similarity):

– косинусная метрика – рассматриваем каждый узел как строку в матрице и находим косинус угла между этими векторами.

$$\sigma(v_i, v_j) = cos(v_i, v_j) = \frac{v_i v_j}{||v_i||||v_j||} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum A_{ik} A_{ki}} \sqrt{\sum A_{jk} A_{kj}}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

– метрика Жаккара – находим пересечение всех общих узлов и делим на несовпадающих соседей. В результате мы находим метрики похожести, и задача становится задачей кластеризации.

$$J(v_i, v_j) = \frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|}{|\mathcal{N}(v_i) \cup \mathcal{N}(v_j)|}$$

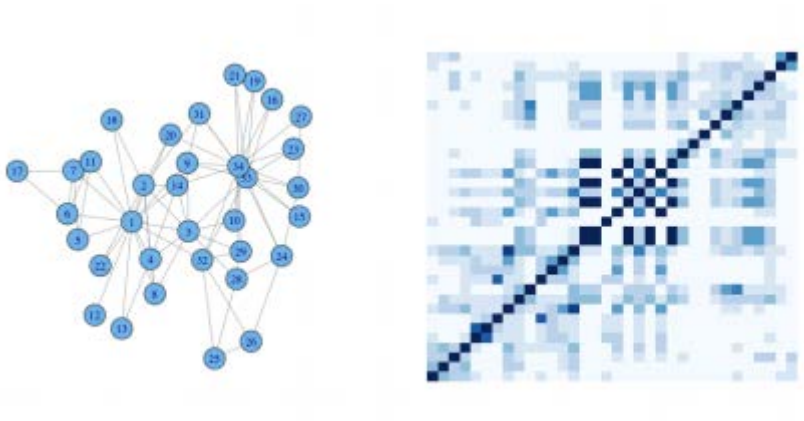


Рис. 52. Граф (а) и его матрица схожести (б)

На рис. 52 представлен граф (а) и матрица схожести для него (б). Чем темнее элемент, тем более эквивалентны узлы. Каждый квадратик посчитан через косинус угла. Далее можно использовать любой алгоритм кластеризации, в том числе агломеративный, k-means, спектральный и др. В результате кластеризации мы получим сообщества, в котором члены сообщества примерно одинаково связаны с остальным графом.

При выделении сообществ часто возникает проблема отрезания очень маленьких кусочков. Для противодействия ей предложена следующая идея: провести предобработку данных и вначале найти ядро графа. Достигается это путем итеративного отрезания узлов со степенью, равной единице (но это может привести к полному распаду графа). Процедура называется нахождением ядра графа.

k-core – ядро степени k – означает, что степень всех входящих в него узлов не меньше k. (k+1)-core всегда является подграфом k-core (рис. 53).

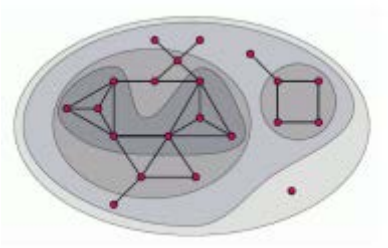


Рис. 53. Ядра графа разных степеней

Пример выделения ядер различных степеней для реальных сообществ представлен на рис. 54.

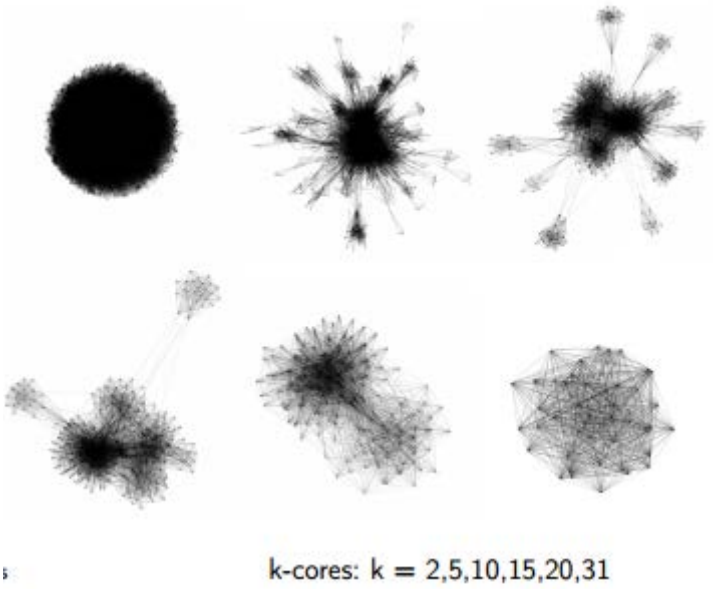


Рис. 54. Последовательное выделение ядер графа
с повышением степени

С практической точки зрения интересно, что, если вычесть ядро, граф может сам распасться на сообщества. Этот механизм вполне применим на реальных больших данных (рис. 55).

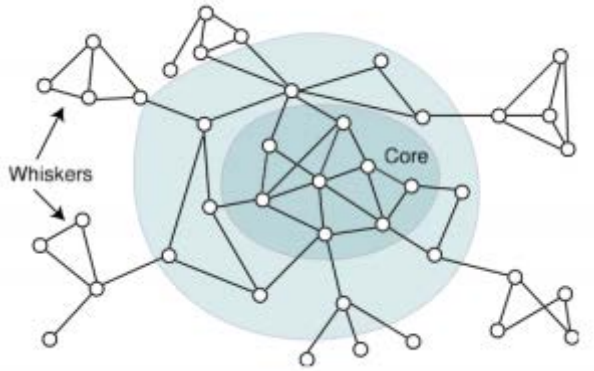


Рис. 55. Выделение сообществ путем удаления ядра

7.2. ДРУГИЕ СПОСОБЫ ОПРЕДЕЛЕНИЯ СТРУКТУРЫ СЕТЕВОГО ГРАФА

Кроме поиска сообществ, можно выделить другие задачи определения структуры графа. Классическая задача – это нахождение клик. Клика – полный подграф, в котором все ребра связаны между собой (рис. 56). В реальном мире очень мало полных клик, однако если допустить, что клика – это не полностью полный подграф, то задача сводится к поиску сообществ. Более интересная задача – найти клику максимального размера.

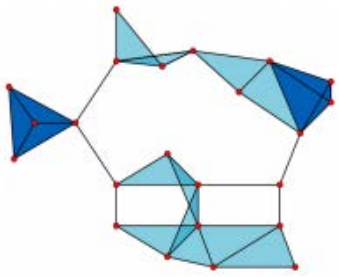
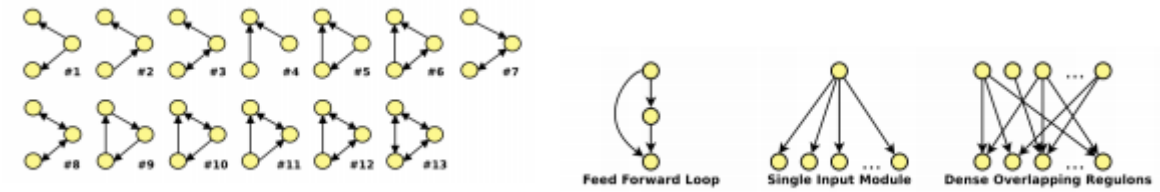


Рис. 56. Клики в графе

Еще одна распространенная задача – поиск часто встречаемых подграфов. Используется термин «сетевой мотив» – подграф, который встречается чаще, чем ожидалось. Примеры сетевых мотивов представлены на рис. 57.



а б

Рис. 57. Сетевые мотивы: а – триады; б – более сложные мотивы

Выбирается структура (мотив), которая часто встречается в графе, и оценивается ее статистическая важность. Оценивается, сколько раз она встречается в данном графе, генерируется набор случайных графов и вычисляется средняя частота подграфа в случайном графе. Если частота в нашем графе отличается существенно, то структура считается статистически значимой.

8. СПЕЦИАЛЬНЫЕ ТИПЫ ГРАФОВ

8.1. ДВУДОЛЬНЫЕ ГРАФЫ

Двудольный граф (биграф) – это такой граф, ребра которого могут быть разделены на два непересекающихся подмножества *U* и *V* таким образом, что ребро соединяет вершину из подмножества *U* с вершиной из подмножества *V*.

Свойства двудольного графа:

– У двудольного графа не бывает циклов нечетной длины. Если граф содержит цикл нечетной длины, то он не может быть двудольным.
– Двудольный граф может соединять вершины двух разных цветов. Например, треугольник не может быть биграфом, т.к. он не может быть раскрашен двумя цветами.

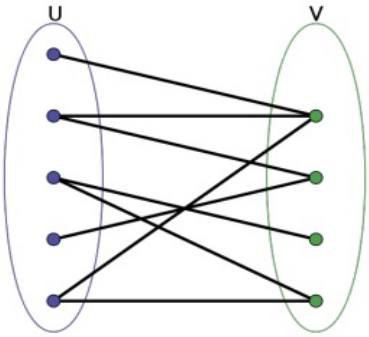


Рис. 58. Пример двудольного графа

Следовательно, чтобы проверить, является ли граф двудольным, нужно попробовать его раскрасить или проверить все циклы в нем.

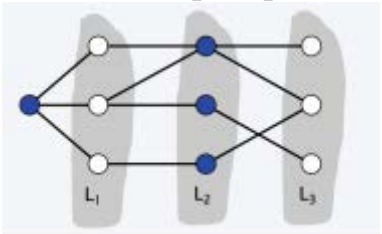


Рис. 59. Алгоритм поиска в ширину

Для «расслоения» графа используем алгоритм поиска в ширину (рис. 58). Каждый слой закрашиваем своим цветом. Проверяем, есть ли ребра с одинаковой окраской на концах и есть ли связи внутри слоев. Если в обоих случаях ответ «нет», то граф – двудольный.

Двудольные графы описывают те реальные ситуации, для которых существуют два типа узлов и связи между узлами разных типов. Примеры таких комбинаций:

– люди и группы,

– кандидаты и вакансии,
– авторы и статьи,
– актеры и фильмы,
– директоры – советы директоров (рис. 60).

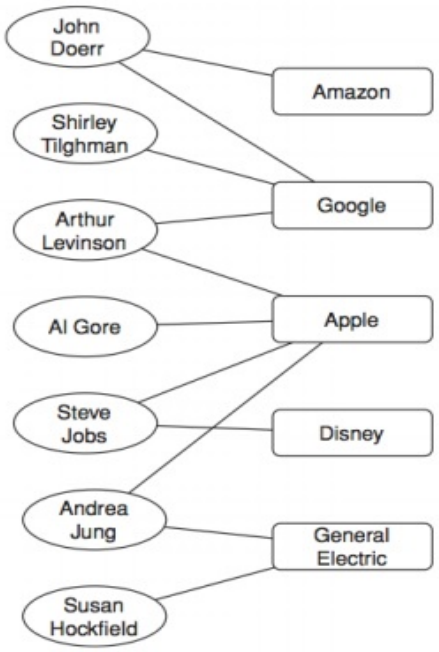


Рис. 60. Пример двудольного графа. Советы директоров

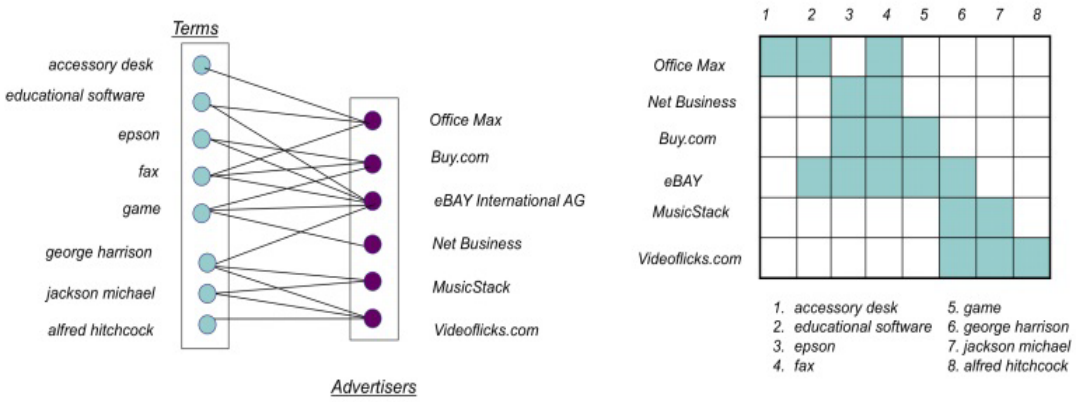


Рис. 61. Пример двудольного графа. Рекламодатели и ключевые слова

Граф на рис. 61 показывает связь между рекламодателями и ключевыми словами, которые они выбирают для показа их рекламы. Слева на примере – термины, справа – рекламодатели, которые за эти слова соревнуются.

Такой двудольный граф очень удобно изобразить в виде матрицы (incidence matrix). Эта матрица не есть матрица смежности. По столбцам расположены узлы одного типа, а по рядам – узлы другого типа (на примере по рядам – рекламодатели, по столбцам – слова). В ячейках на пересечении

можно обозначить либо наличие ребра (0 и 1), либо его вес (например, цену, которую рекламодатель готов заплатить за показ рекламы на данное ключевое слово).

Обратим внимание на то, что матрица не симметрична, и число ее ненулевых значений равно количеству ребер двудольного графа.

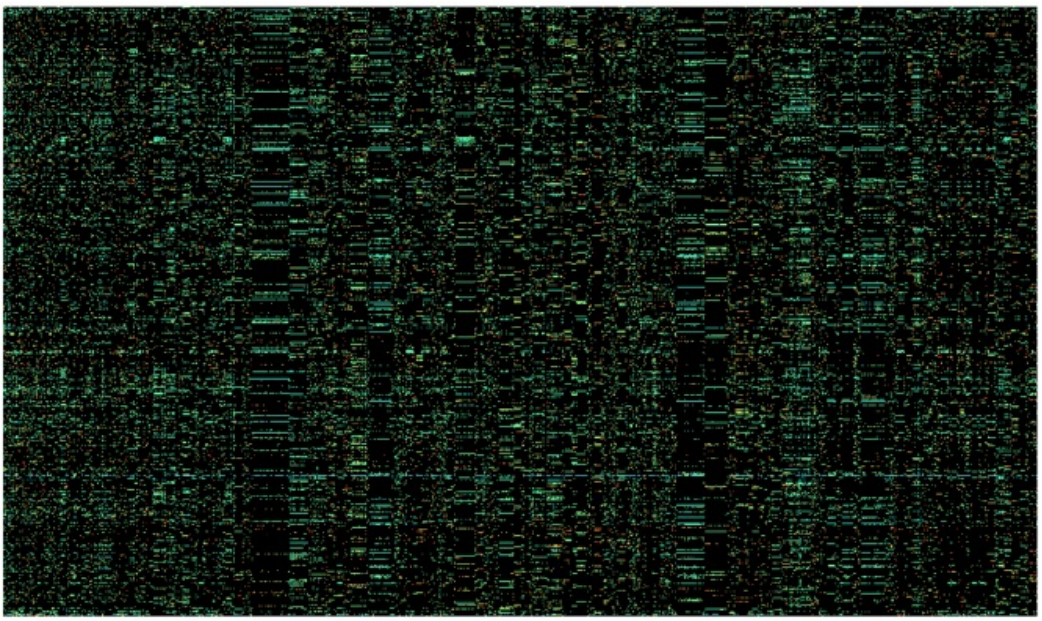


Рис. 62. Пример реальной матрицы для рекламного аукциона

На рис. 62 изображен фрагмент матрицы инциденций для рекламного аукциона. Масштабы – порядка 2000 акционеров и порядка 3000 различных поисковых терминов, на которые они ставят наивысшую цену спроса. Зеленая точка показывает, что существует ребро между акционером и словом. Здесь хорошо видно, что эта матрица – достаточно разреженная и совсем не симметричная.

Более формально матрица инциденций определяется следующим образом. Имеется m пользователей и n групп; формируем матрицу B размером m на n, в которой элемент B_{ij} равен 1, если пользователь i принадлежит группе j, и 0 в противном случае.

$$B_{ij} = \begin{cases} 1: & \text{if user } i \text{ belongs to group } j \\ 0: & \text{otherwise} \end{cases}$$

Можно также сгруппировать узлы обоих типов – сначала узлы, отвечающие пользователям, затем узлы, отвечающие группам. В результате этого получается один длинный пронумерованный вектор узлов, который можно использовать для того, чтобы построить матрицу смежности (adjacent matrix) всего графа:

$$A = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}.$$

В этой матрице $m+n$ колонок и $m+n$ строк, причем первые m и строк, и столбцов – это пользователи, а дальше – ключевые слова. Нули показывают, что сами пользователи с самими пользователями не связаны. Пользователи с группами связаны с помощью матрицы инциденций.

Дальше можно рассматривать двудольный граф как обычный граф с одним типом узлов, а в конце разделить узлы по цвету.

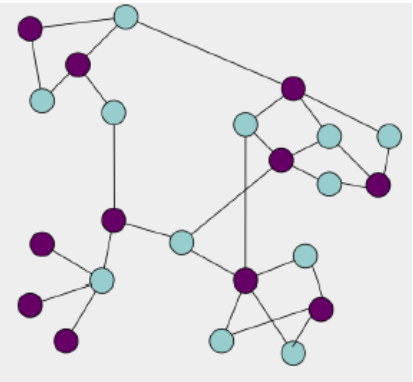


Рис. 63. Двудольный граф

Двудольный граф на рис. 63 отвечает той матрице, которая была введена выше, и описывает связь ключевых слов и рекламодателей. Если требуется найти сообщества на этом графе (группы людей и слов вместе), то можно рассматривать целиком граф с помощью матрицы смежности и использовать любой способ нахождения сообществ в графе. И только на последнем этапе происходит возврат от матрицы смежности к матрице инциденций.

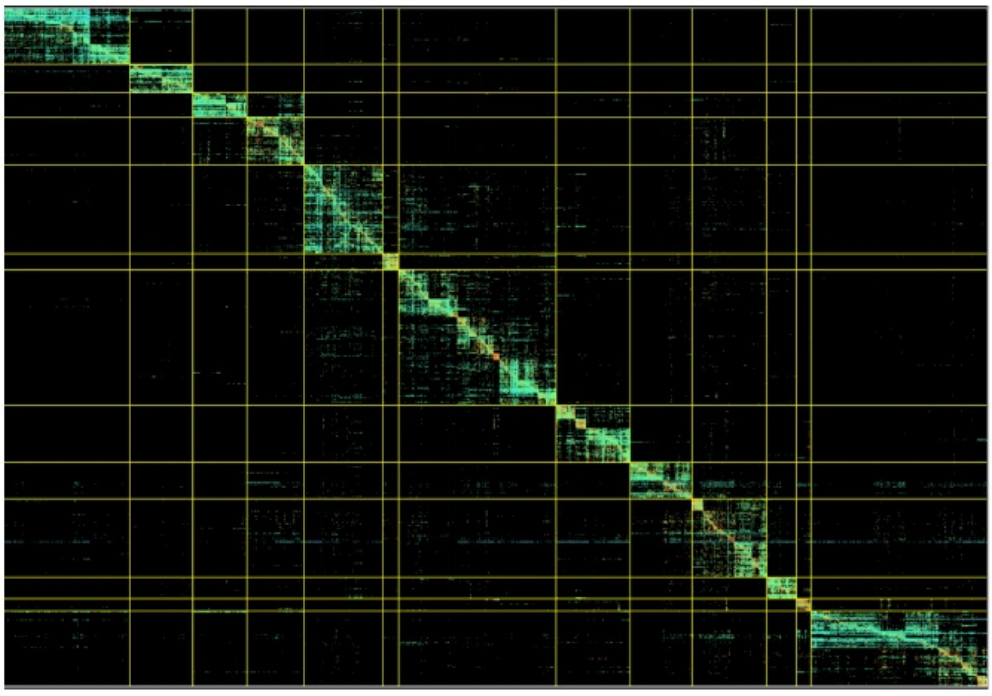


Рис. 64. Кластеризованные данные

На рис. 64 эти же данные уже кластеризованы. Для этого двудольный граф был представлен в виде обычного графа и проведена кластеризация. После этого узлы были перенумерованы таким образом, что узлы, принадлежащие одному и тому же кластеру, получили последовательные номера. В результате этого колонки и ряды матрицы оказались переставленными в соответствии с новой нумерацией.

Кластерная структура матрицы представлена здесь в виде квадратиков, где каждый квадратик – это группа рекламодателей и группа рекламных слов, в которых существует тесное соотношение между ними. С точки зрения двудольного графа – это «маленькие рынки».

8.2. СЕТИ АФФИЛИИРОВАННОСТИ

Еще одно интересное свойство двудольного графа состоит в том, что из него можно создать две проекции на однодольные графы.

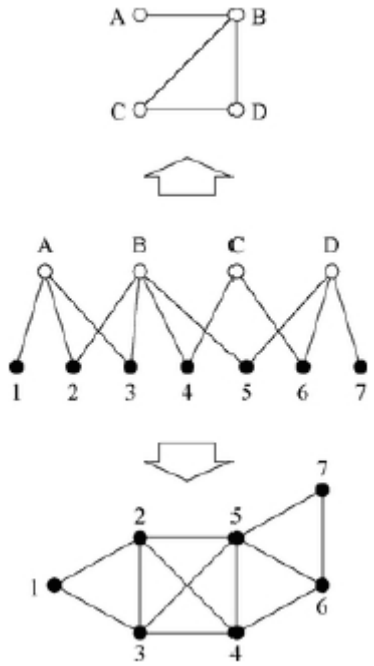


Рис. 65. Проекции на однодольные графы

Если A, B, C, D на рис. 65 – это объекты, то можно составить граф связи между объектами, основываясь на самих связях в двудольном графе. Например, из объекта A можно попасть в объект B двумя способами. Построим граф только из одних объектов A, B, C, D. Тогда объекты A и B будут связаны ребром в этом графе. Вес ребра, т.е. количество способов, по которым можно добраться из узла в узел, равен 2. Объекты A и C не будут соединены ребром, потому что нельзя за два шага добраться из A в C, т.е. прямая связь между ними отсутствует. Прямая связь в этом графе имеет место только в том случае, если можно из одного объекта в другой дойти за два шага.

Можно построить проекцию на 1, 2, 3, 4, 5, 6, 7 (пусть это будут люди), а также построить связь между людьми. Здесь два человека связаны, если они, например, принадлежат одной и той же группе. Например 2 и 3 принадлежат группе A и группе B, а это значит, что на нижнем графе у ребра между 2 и 3 вес будет равен двум. 1 и 7 не имеют связи.

Двудольный граф на однодольный можно спроецировать двумя способами:

– проекция на пользователей

$P' = BB^T$,

– проекция на группы

$P'' = B^T B$.

Эти матрицы – почти матрицы смежности, но имеющие циклы от узлов, приходящих на самих себя.

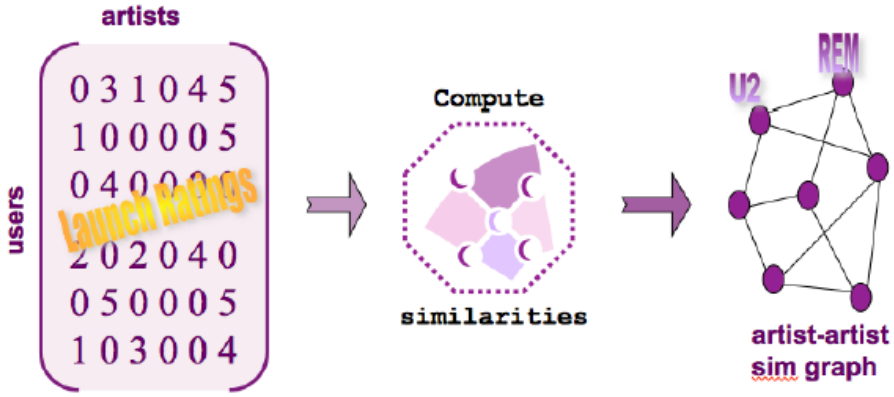


Рис. 66. Матрица рейтингов

Пример использования проекций на однодольные графы – задача нахождения групп схожих артистов (рис. 8).

Пусть есть какая-то система ранжирования, где пользователь ставит свои оценки музыкантов, и набор пользователей, которые выставляют оценки музыкантам. Тем самым формируется двудольный граф, который представлен матрицей рейтингов.

Используя эту матрицу, можно найти группы музыкантов, которые, по мнению пользователей, похожи между собой. Идея расчета состоит в следующем: два музыканта похожи между собой, если примерно одинаковое количество пользователей оценило их одинаково.

Создадим граф связи между различными музыкантами. Два музыканта связаны, если достаточное количество пользователей оценило их одинаково. Вес на этом ребре как раз будет означать количество одинаковых оценок. Получается, что музыканты, которые нравятся одновременно некоторому количеству пользователей, попадают в один кластер. Если теперь сделать обратную проекцию – на группы пользователей, то получатся группы пользователей с более-менее одинаковым вкусом.

Приведем другие примеры сетей с помеченными ребрами:

– социальные сети: дружба – положительное ребро, враждебность – отрицательное ребро;
– динамика дружбы, эволюция сетей.

Ребро в графе может нести некоторый смысл или иметь знак – положительный или отрицательный. Благодаря введению ребер со знаком локальное свойство (ребра со знаком) может привести к глобальным свойствам графа.

Например, рис. 67 иллюстрирует следующую ситуацию: есть муж, жена и общий друг. Все три человека – друзья между собой, и все три связи положительные. Далее по какой-то причине происходит развод, особой любви между бывшими мужем и женой не остается, и ребро оказывается негативным (отрицательным).

Что остается делать другу семьи? Происходит разрыв еще одной связи (с мужем или женой), т.е. имеет место динамика отношений.

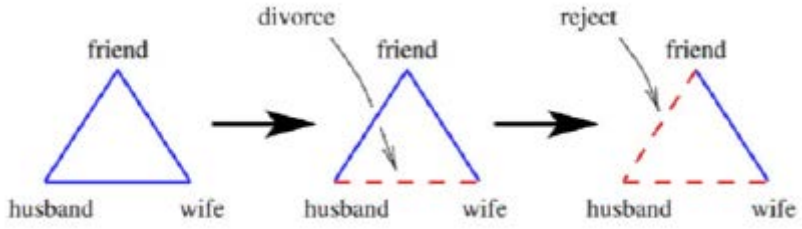


Рис. 67. Пример связи между узлами

Представим ситуацию:

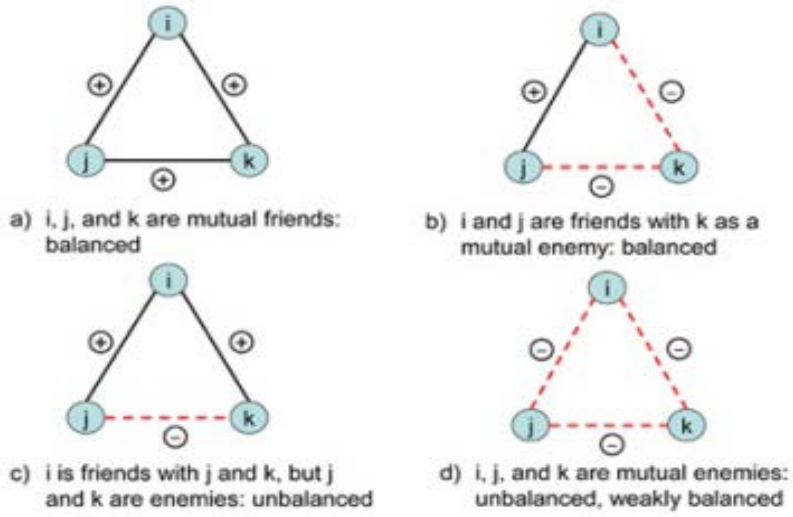


Рис. 68. Варианты связей

Этот подход был предложен социологами в теории социального баланса. Основная идея теории заключается в том, что если у нас есть два человека, и они связаны положительной связью, то у них есть тенденция схожести отношения к третьему узлу.

Если посмотреть на треугольники этого типа (рис. 68), то возможны 4 случая:

1) Все связи положительные. Такой треугольник сбалансирован.
2) Одна положительная связь и две отрицательные. В жизни это означает, что два человека «дружат против кого-то».

В двух других ситуациях возникает социальное напряжение.

3) Два положительных ребра и одно отрицательное – история о разведенных муже и жене и их общем друге. Тут либо надо изменить отношение, либо оборвать связь.
4) Все трое не любят друг друга. В такой ситуации просто будут порваны все связи.

Таким образом, можно говорить о сбалансированности или несбалансированности графа. Граф сбалансирован, если каждый треугольник в графе сбалансирован. Социологи рассматривали эту теорию для полных графов, когда каждый узел связан с каждым другим, так как социологи работали с небольшими группами людей, где каждый знает каждого.

Для полного графа доказана теорема баланса: если граф со знаками сбалансирован, тогда либо все люди в этом графе должны быть друзьями, либо людей из этого графа можно разделить на две подгруппы X и Y так, что каждая пара узлов внутри X – друзья, каждая пара внутри Y – тоже друзья, но между каждым узлом из пары X с каждым узлом из пары Y существует вражда.

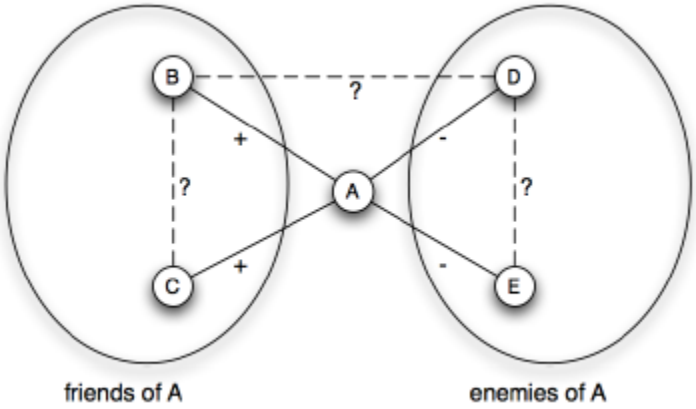


Рис. 69. К теореме баланса

Применение этой теоремы иллюстрирует рис. 69.

1. Выбираем один узел, например, A.
2. Все узлы по отношению к этому узлу либо друзья, либо враги
3. Собираем всех друзей в одну группу и всех врагов в одну группу.
4. Любые два узла из подмножества должны дружить.
5. Два узла из двух разных подмножеств дают отрицательную связь.

Неполная сеть (рис. 70) сбалансирована в том случае, если:

1. можно добавить ребра и расставить на них знаки так, чтобы полученный граф был сбалансирован;

2. можно разделить граф на два класса узлов так, что внутри каждого класса узлы дружат между собой, между классами – враждуют.

Граф с весами сбалансирован в том случае, если он не содержит циклов с нечетным количеством отрицательных ребер (рис. 12).

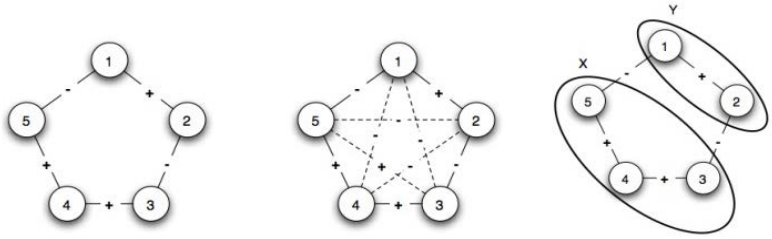


Рис. 70. Неполная сеть

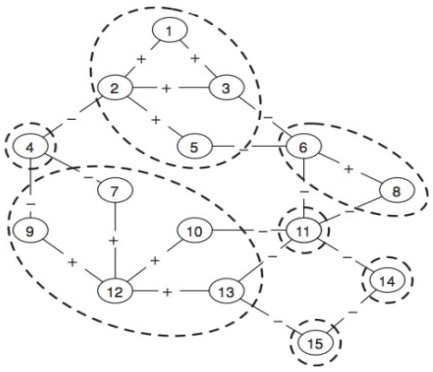


Рис. 71. Объединение положительных узлов в группы

Если граф разбивается на две группы так, что внутри группы люди дружат, а между группами – нет, это возможно только в том случае, если этот граф двудольный, т.е. не содержит циклов с нечетным количеством отрицательных ребер.

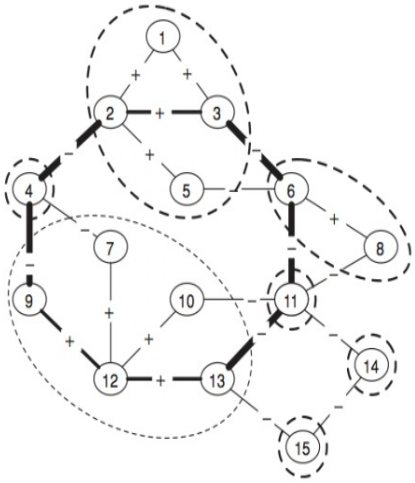


Рис. 72. Попытка раскраски графа

Например, граф на рис. 71 не может быть раскрашен двумя цветами, так как содержит цикл нечетной длины (рис. 72). Соответственно, такой граф не является структурно сбалансированным.

Литература

1. Barabasi A.L., Bonabeau E. Scale Free Networks // Scientific American, 2003, p. 50-59,
2. Newman M. The physics of networks. Physics Today, 2008
3. Newman M. Networks: An Introduction. Oxford University Press, 2010.
4. Newman M., Girvan M.. Finding and evaluating community structure in networks // Phys. Rev. E 69, 026113, 2004.
5. Newman M. Modularity and community structure in networks // PNAS Vol. 103, N 23, pp 8577-8582, 2006
6. Fortunato S. Community detection in graphs // Physics Reports, Vol. 486, pp. 75-174, 2010
7. Easley D., Kleinberg J.. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010. – Электронный ресурс. Режим доступа: http://www.cs.cornell.edu/home/kleinber/networks-book/
8. Социальные сети и виртуальные сетевые сообщества / отв. ред. Верченов Л.Н., Ефременко Д.В., Тищенко В.И. М: ИНИОН РАН, 2013. 360 с.
9. Ермолова Н. Продвижение бизнеса в социальных сетях Facebook, Twitter, Google+. М.: Альпина Паблишер, 2013. 357 с.
10. Гусарова Н.Ф. Интеллектуальные системы в управлении социальными процессами. – СПб: Университет ИТМО, 2015. 90 с.

УНИВЕРСИТЕТ ИТМО

Миссия университета – генерация передовых знаний, внедрение инновационных разработок и подготовка элитных кадров, способных действовать в условиях быстро меняющегося мира и обеспечивать опережающее развитие науки, технологий и других областей для содействия решению актуальных задач.

КАФЕДРА ИНТЕЛЛЕКТУАЛЬНЫХ ТЕХНОЛОГИЙ В ГУМАНИТАРНОЙ СФЕРЕ

Кафедра интеллектуальных технологий в гуманитарной сфере была организована в 1998 году и при образовании получила название «кафедра технологий профессионального обучения». Тогда же кафедру возглавил профессор Потеев Михаил Иванович, и с 2002 года кафедра стала выпускающей. В 2012 году кафедра была переименована в соответствие с основным направлением деятельности.

Центральной идеей образовательных программ, реализуемых кафедрой, является участие студентов в выполнении работ, связанных с возможными направлениями будущей деятельности, и с задачами, решаемыми университетом. Научные исследования, проводимые на кафедре, связаны с интеллектуальным анализом данных, математическим моделированием и проектированием информационных систем. В этих областях много интересных, сложных и нерешенных задач. На старших курсах студенты имеют возможность выбрать то направление в рамках профиля, которое им наиболее интересно.

Существующие международные программы

Сотрудниками кафедры интеллектуальных технологий в гуманитарной сфере ведутся переговоры о внедрении программ двойных/совместных дипломов и реализации международных программ академического студенческого обмена с вузами стран Шанхайской организации сотрудничества, а также Чехии, Англии и Финляндии.

Компании, в которых осуществляется производственная и преддипломная практика, а также компании, трудоустраивающие выпускников

Доктор Web, Государственный Русский музей, ООО «ИНТЕРФОРУМ», Ростелеком, Интерзет, ООО «Интермедиа», «ВКонтакте» и др.

http://www.ifmo.ru/ru/viewdepartment/13/kafedra_intellektualnyh_tehnologiy_v_gumanitarnoy_sfere.htm#ixzz3byzWK4D7

Гусарова Наталия Федоровна

АНАЛИЗ СОЦИАЛЬНЫХ СЕТЕЙ.
ОСНОВНЫЕ ПОНЯТИЯ И МЕТРИКИ

Учебное пособие

В авторской редакции

Редакционно-издательский отдел Университета ИТМО

Зав. РИО Н.Ф. Гусарова

Подписано к печати 18.11.2016

Заказ №

Тираж 50

Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49