Прикладные задачи анализа данных

АНАЛИЗ СОЦИАЛЬНЫХ СЕТЕЙ
SOCIAL NETWORK ANALYSIS
ЛЕКЦИЯ №1

Дьяконов А.Г.

Московский государственный университет имени М.В. Ломоносова (Москва, Россия)

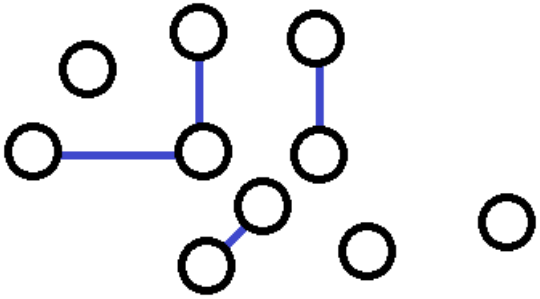
Исследование социальных сетей

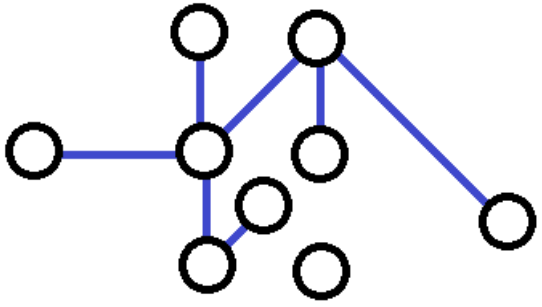
Социальная сеть – динамический граф (пример: мобильная сеть)

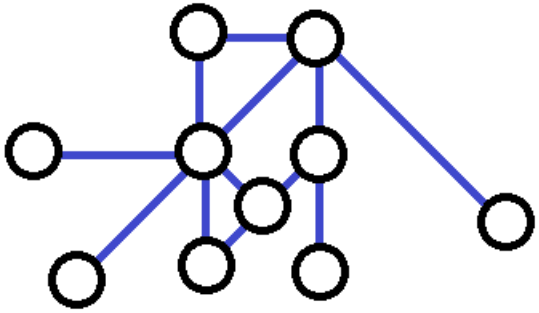
Вершины – пользователи (и группы)

Рёбра – дружба (членство) / связи, отношения

Кластеры – сообщества







Примеры соцсетей:

сети дружбы (Friendship Networks)

- «классические» (Facebook, vk, Одноклассники)

сети общения (Communication Networks)

- мобильные сети
- мессенджеры (Telegram, WhatsApp)
- микроблоги (Twitter)
- почтовые (связь по отправке писем)

информационные сети (Information Networks)

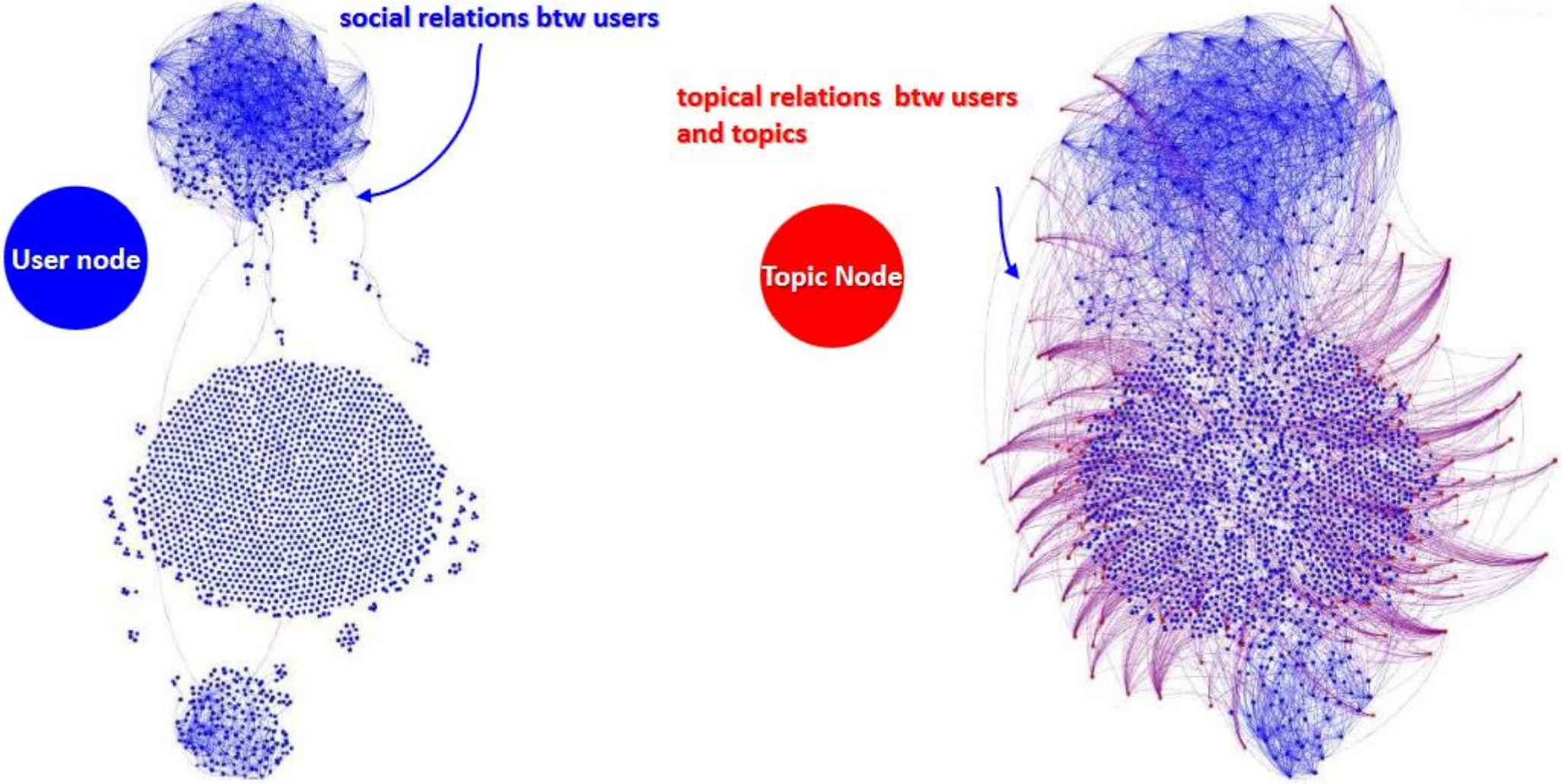
- сам интернет

- интернет-магазин (связь по одинаковым купленным товарам)
- научные сообщества (связь по публикациям)

Какие здесь графы?

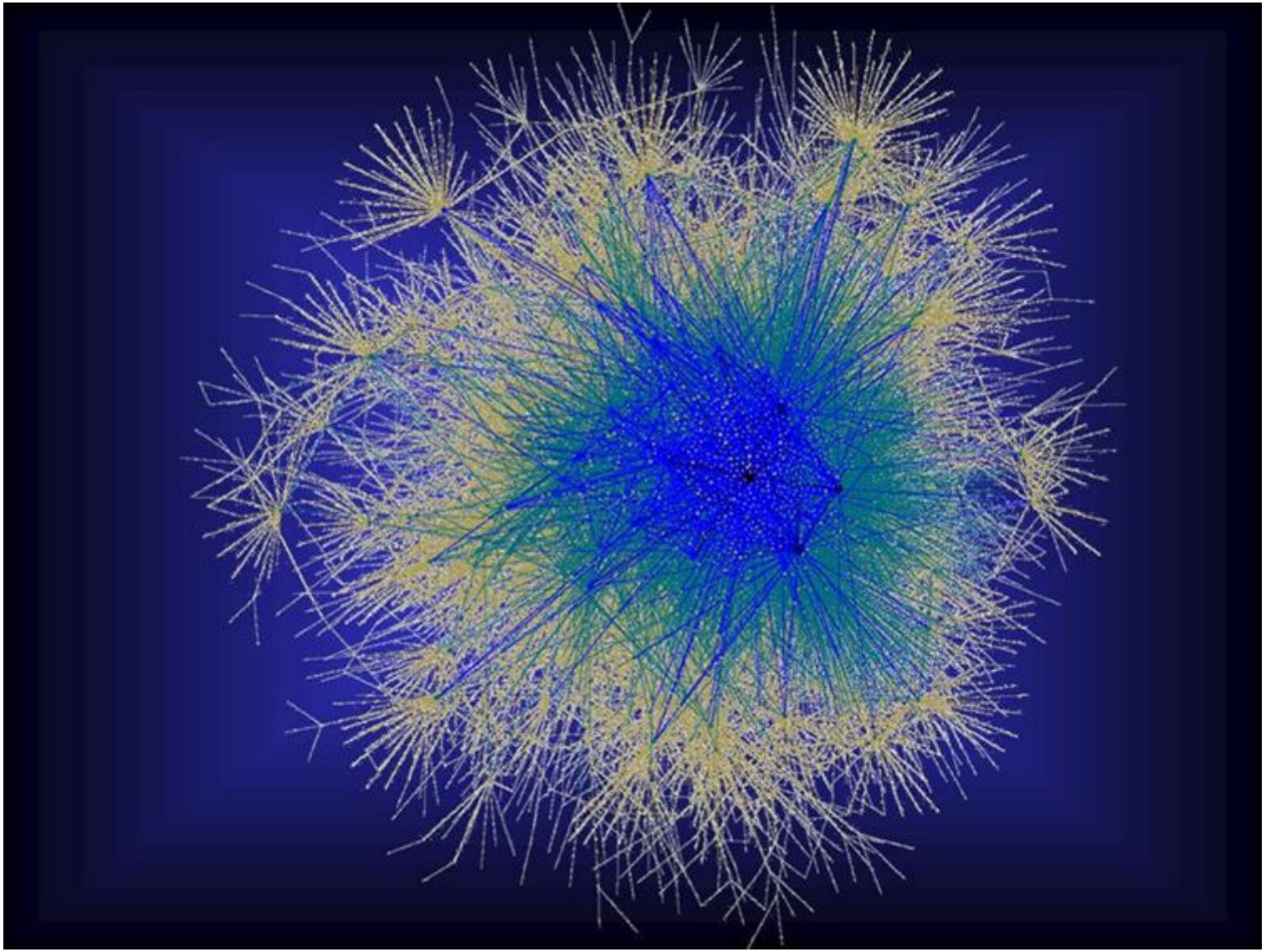
Какие задачи здесь актуальны (возможны)?

Вершины не обязательно пользователи



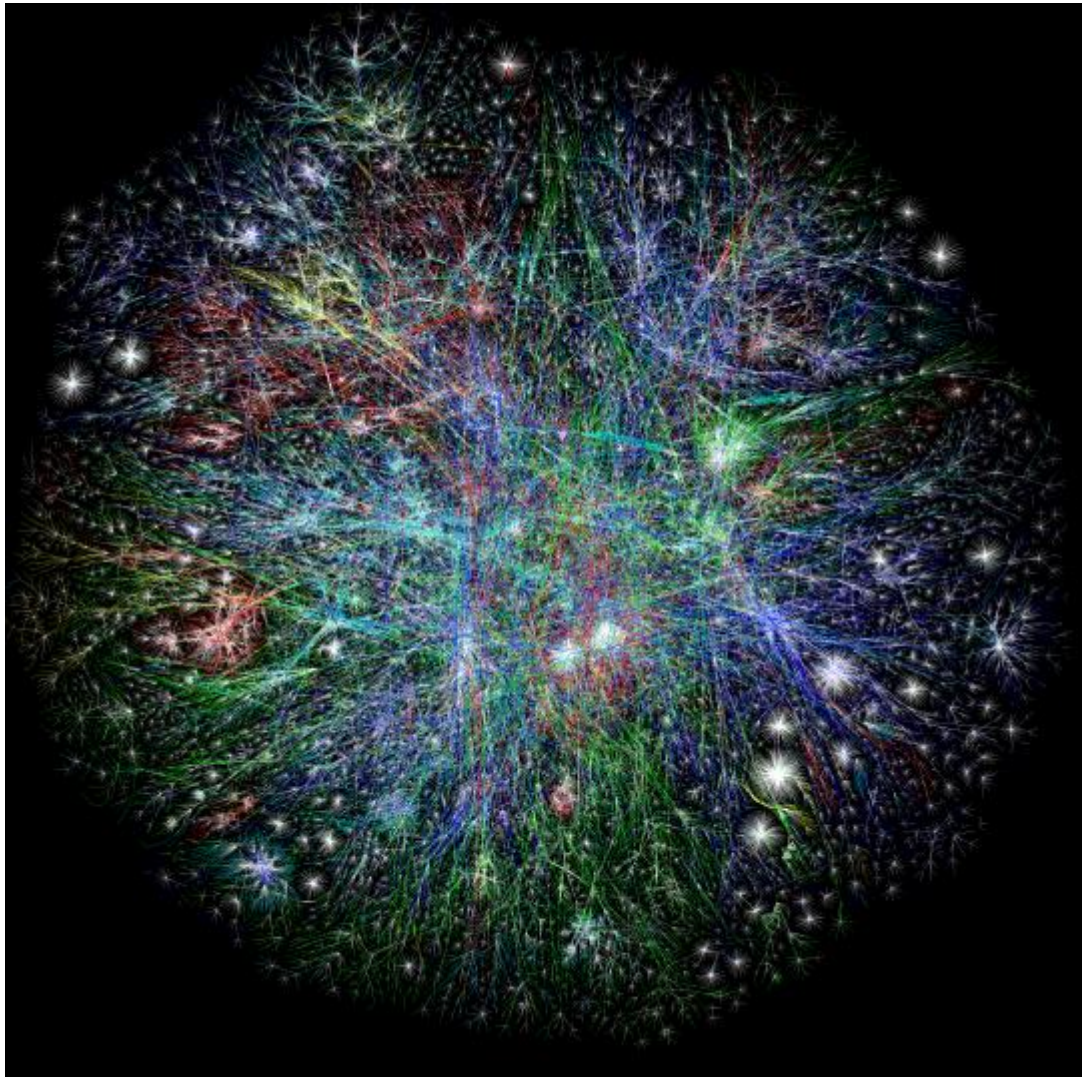
http://legacydirs.umiacs.umd.edu/~hadi/cmsc498j/slides/lec-1.pdf

Картинки с графами



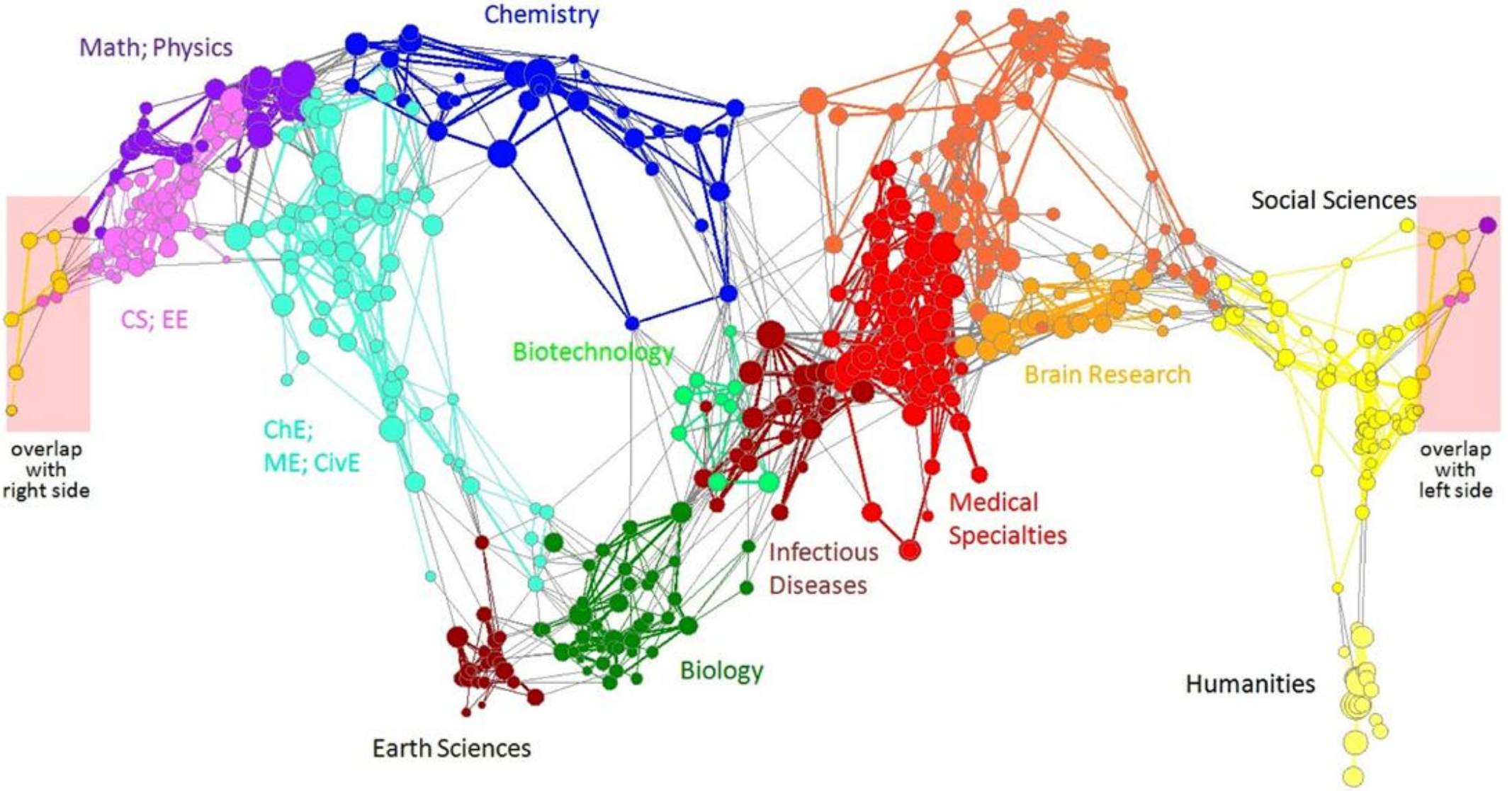
graph of the BGP (Gateway Protocol) web graph,
consisting of major Internet routers (6400 вершин, 13000 рёбер)
Ross Richardson, Fan Chung Graham

Web-граф



Примеры графов

Граф цитирований



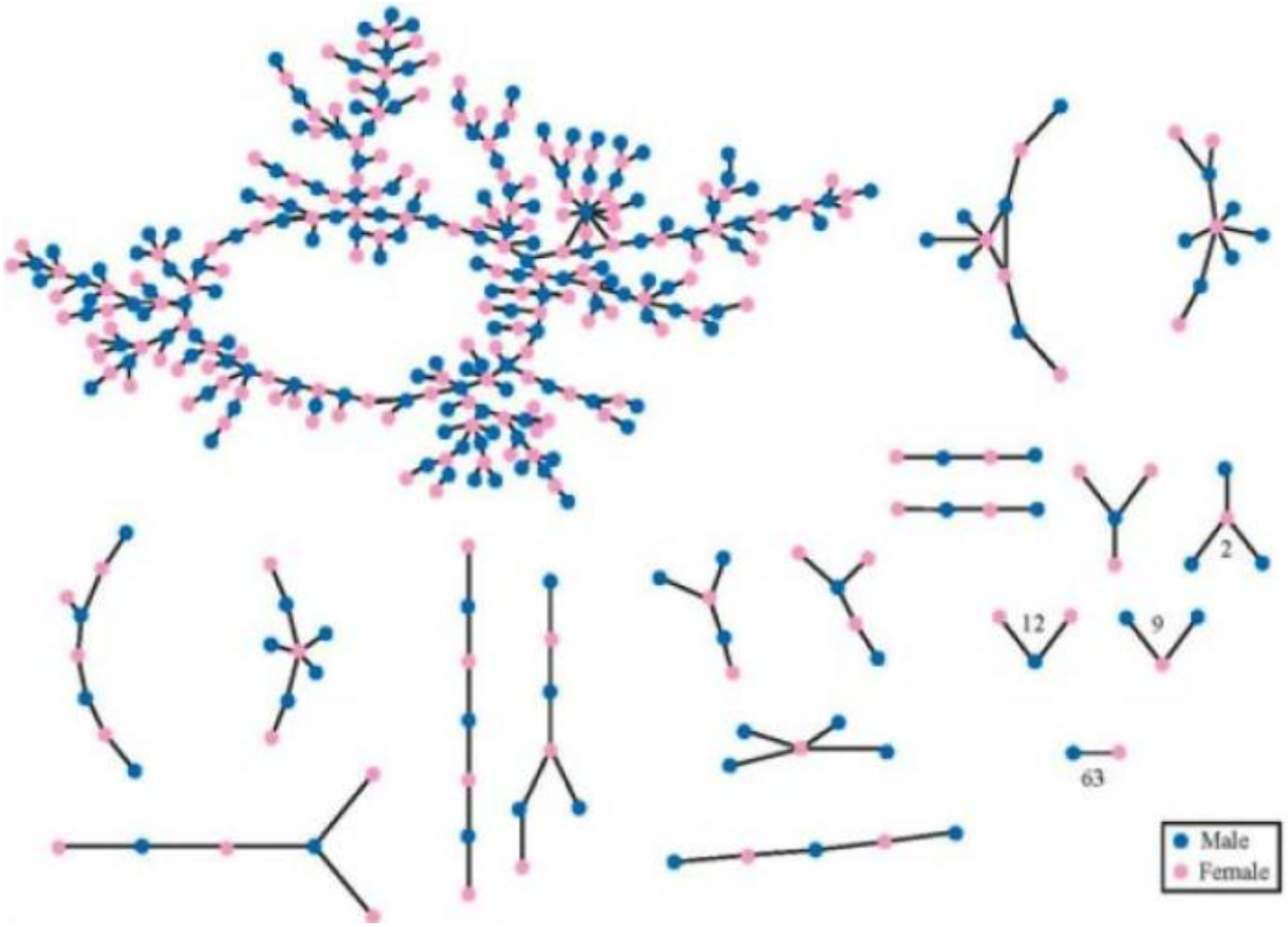
Börner и др.

Примеры графов

граф метро

Примеры графов

Граф романтических отношений



Задачи с социальными сетями

- **Анализ поведения пользователей**
  - выявление аккаунтов-дубликатов
  - пользователей нарушающих, склонных нарушать правила, не похожих на других
- **Прогнозирование**
  - поведения пользователей (когда будет пользоваться услугами, в какую группу вступит, с кем подружится)
  - предсказание и предотвращение ухода пользователей
  - предсказание трафика (в каком объёме будет скачивать/закачивать)
- **Рекомендация**
  - предсказание эффективности действия рекламы для конкретного пользователя
  - формирование таргетированных предложений (рекламы, по вступлению в группы, заполнению профиля и т.п.)

Задачи с социальными сетями

- **Кластеризация**
  - разбиение пользователей на группы (для более корректного A/B-тестирования, разработки стратегий под группы, более тщательного анализа аудитории)
  - выявление «кругов общения пользователей» (друзей, которых объединяет некоторая сущность, например «друзья по вузу»)
  - выделение сообществ
  - выделение базисов источников информации в блогосфере
- **Взаимодействие с другими соцсетями/ресурсами**
  - матчинг сетей/графов (установление соответствия между пользователями одной сети и другой)
  - использование данных соцсети для решения задач других заказчиков
    - скоринг (оценка заёмщика) - в банках
    - персональные рекомендации - в интернет-магазинах
    - таргетированная реклама - в рекламе, СМИ (таргетированные новости)

Задачи с социальными сетями

- **Анализ текстов**
  - обнаружение недопустимых текстов (оскорблений, рекламы, нарушения закона и т.п.)
  - анализ общественного мнения по постам
  - анализ лояльности к брендам по постам
- **Визуализация**
  - поиск закономерностей в данных соцсети и их представление анализ общественного мнения по постам
  - научные исследования графов соцсетей

Основные понятия теории графов

Граф

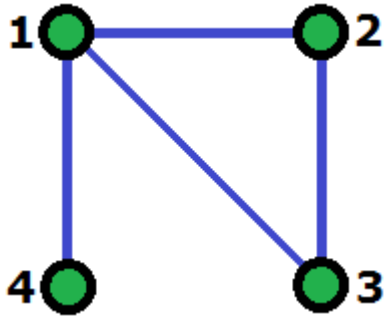
$$(V, E)$$

$$i, j \in V$$

$$\{i, j\} \in E$$

ребро / дуга

смежные вершины / соседи / друзья



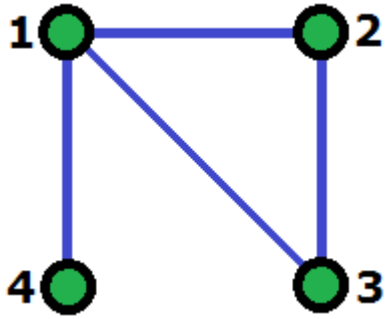
Вершины 1 и 2 смежны

Рёбра (1, 2) и (1, 3) смежны

Вершины 1 и ребро (1, 2) инцидентны

Основные понятия теории графов

Граф



матрица сопряжённости (Adjacency Matrix)

	1	2	3	4
1		1	1	1
2	1		1	
3	1	1		
4	1			

как правило разреженная

диагональная матрица степеней

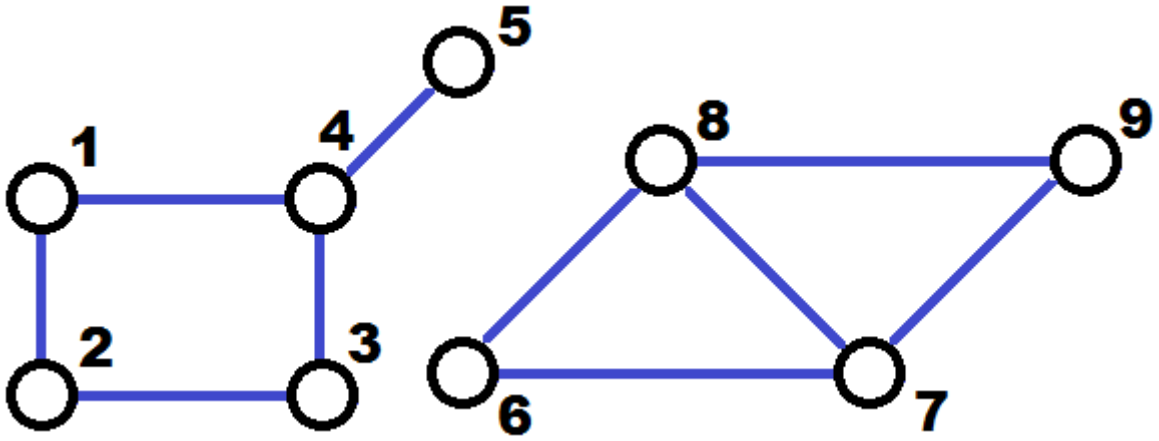
	1	2	3	4
1	3			
2		2		
3			2	
4				1

матрица Лапласа

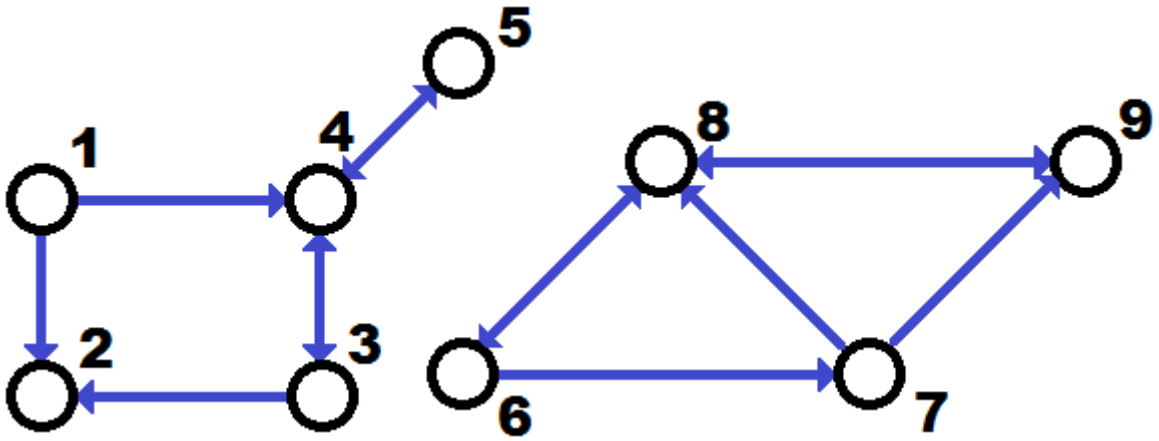
	1	2	3	4
1	3	-1	-1	-1
2	-1	2	-1	
3	-1	-1	2	
4	-1			1

Основные понятия теории графов

Неориентированные



Ориентированные (digraphs) (где?)



соседство

окрестности

степень

входящая/исходящая степень (indegree/outdegree)

связные компоненты

клика

максимальная клика

кратчайший путь

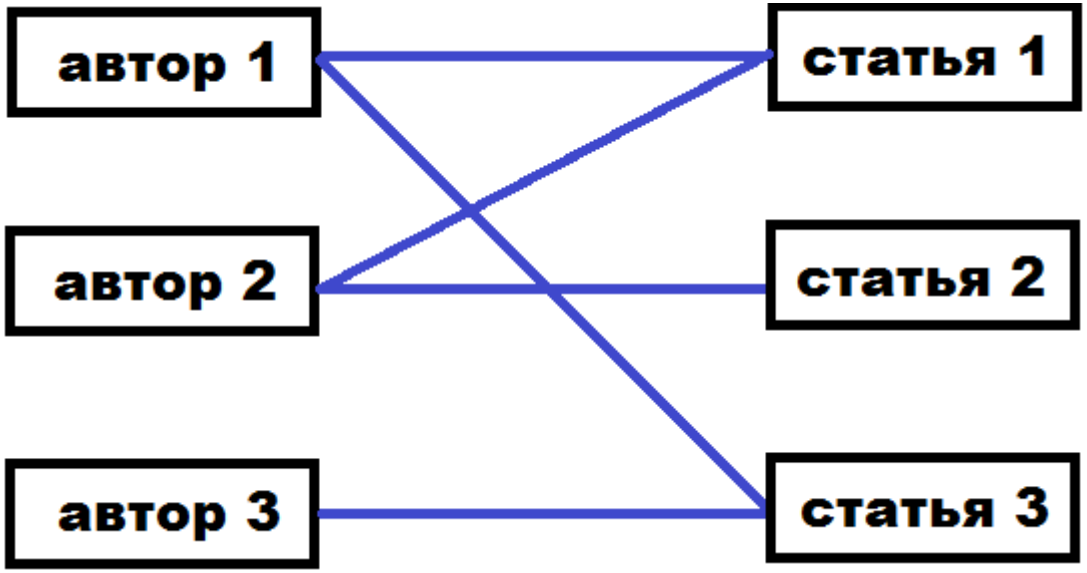
поток

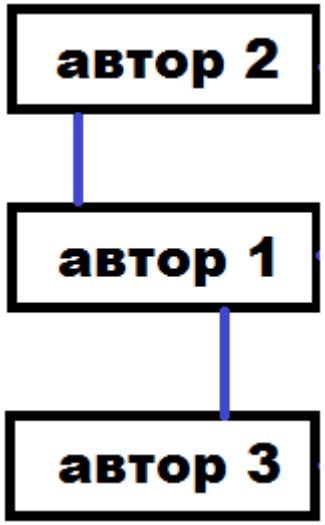
диаметр

+ взвешенные графы

Основные понятия теории графов

Двудольные графы (bipartite)





Ещё: фильмы – актёры

Научные сообщества

Граф цитирования (ориентированный)

Граф соавторства (неориентированный/двудольный)

Граф сходства статей (с весами)

Основные понятия теории графов

Плотность графа (Graph Density)

$$\frac{2|E|}{|V|(|V|-1)}$$

Расстояние между вершинами – длина кратчайшего пути между ними

Диаметр – максимальное расстояние (по всем парам вершин графа)

Маршрут в графе — это чередующаяся последовательность вершин и рёбер графа вида

$$v_0,(v_0,v_1),v_1,\ldots,(v_{k-1},v_k),v_k$$

любые два соседние элемента (вершина и ребро) инцидентны

Маршрут замкнут (closed), если $v_0 = v_k$

Основные понятия теории графов

Путь (Walk) — последовательность рёбер (в неориентированном графе) и/или дуг (в ориентированном графе), такая, что конец одной дуги (ребра) является началом другой дуги (ребра). Или последовательность вершин и дуг (рёбер), в которой каждый элемент инцидентен предыдущему и последующему.

Простой путь (Trail) — путь, все рёбра которого попарно различны

Path – **A walk with distinct nodes & edges.**
A closed trail is a circuit l
A cycle is a closed walk with no repeated nodes except v 0 = vl

Длина пути – число рёбер в нём

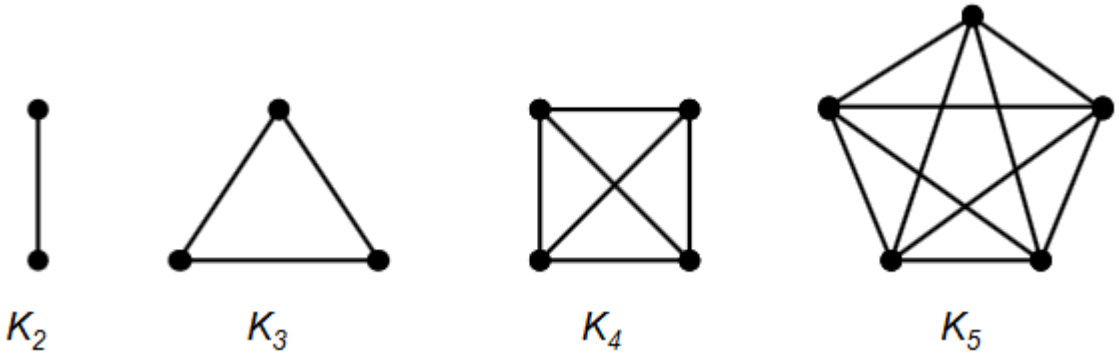
Основные понятия теории графов

Сильно связный (strongly connected) – если для любой пары (u, v) вершин, u достижима из v и наоборот

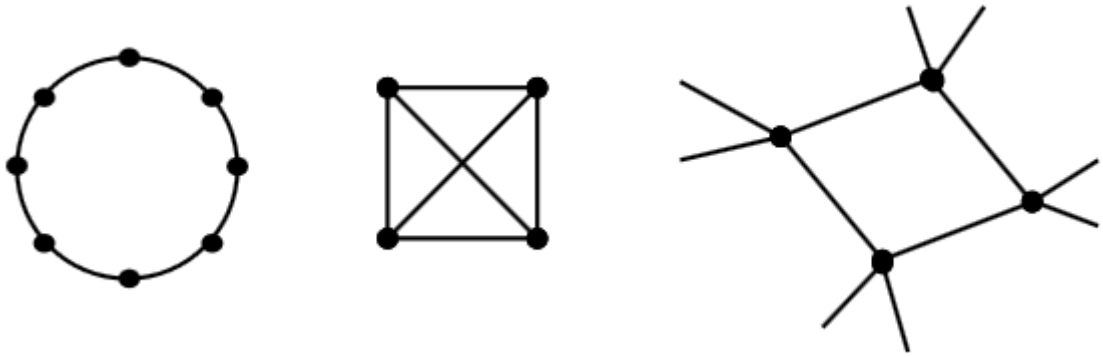
Слабо связный (weakly connected) – если сильно связный после устранения ориентации рёбер

Основные понятия теории графов

Полные графы (complete graph)



d-регулярные (d-regular)



Основные понятия теории графов

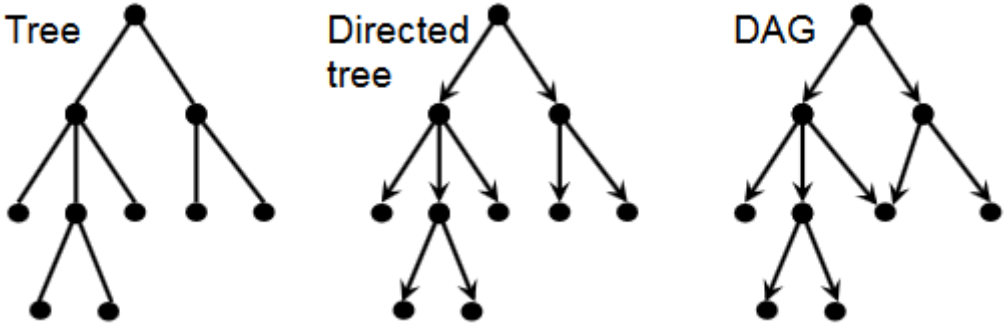
Дерево – связный граф без циклов

Лес – граф без циклов

родитель (parent), ребёнок (child), лист (leaf)

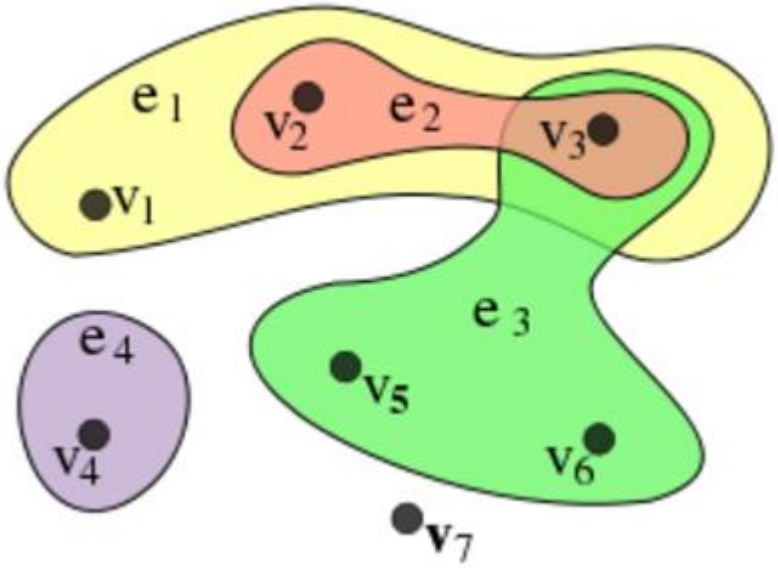
Направленное дерево (Directed tree)

Направленный ациклический граф (DAG = Directed Acyclic Graph)

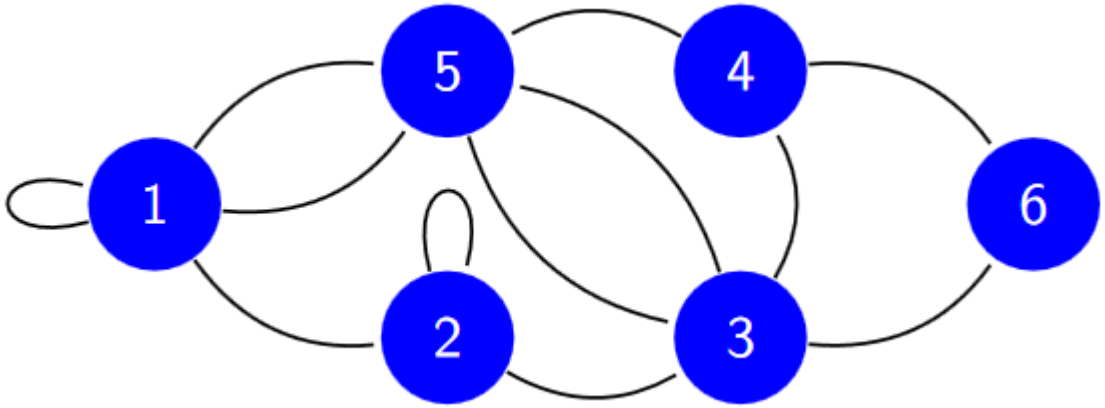


Основные понятия теории графов

Гиперграф

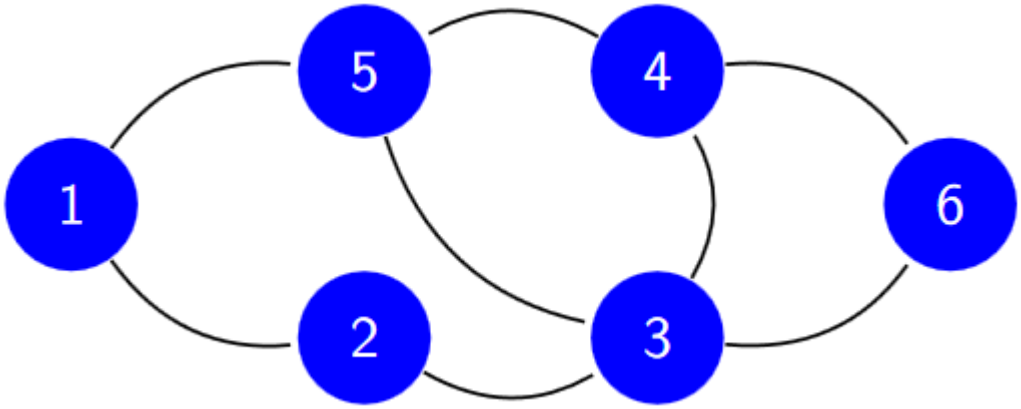


Мультиграф



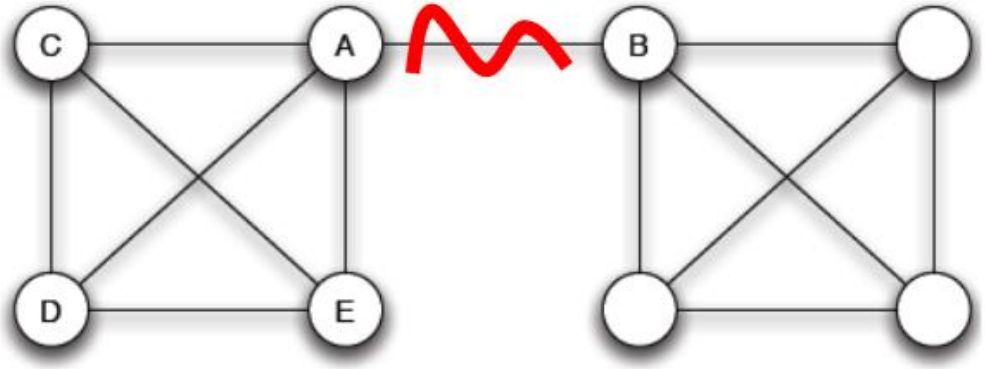
м.б. петли и кратные рёбра

Простой граф – без кратных рёбер и петель

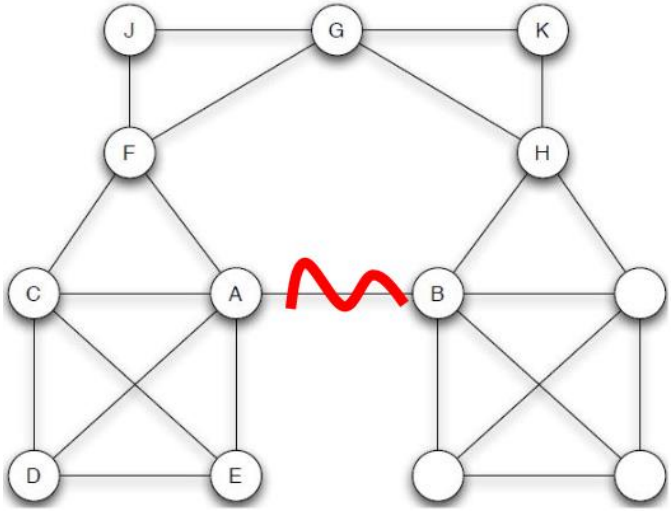


Основные понятия теории графов

Ребро (A, B) – мост, если удаление ребра увеличивает число связных компонент.



Ясно, что на практике определение очень строгое…



Основные понятия теории графов

Ребро (A, B) – локальный мост,
если вершины A и B не имеют общих друзей

~ если удаление увеличивает расстояние между вершинами, как минимум, на 2

пролёт моста (span of a local bridge) – расстояние между вершинами моста после его удаления

Чем полезно для нас?

Основные понятия теории графов

Ребро (A, B) – локальный мост,
если вершины A и B не имеют общих друзей

~ если удаление увеличивает расстояние между вершинами, как минимум, на 2

пролёт моста (span of a local bridge) – расстояние между вершинами моста после его удаления

Это неплохой признак!

Понятие сложной сети (Complex network)

1. Степенные законы распределения степеней вершин (Power law degree distribution)

2. Модель «малого мира» (малый диаметр и т.п.) («small world»)

3. Высокий коэффициент кластеризации (Hight clustering coefficient)

4. Разреженность (Sparcity)

5. Сильные и слабые связи, кластерная структура

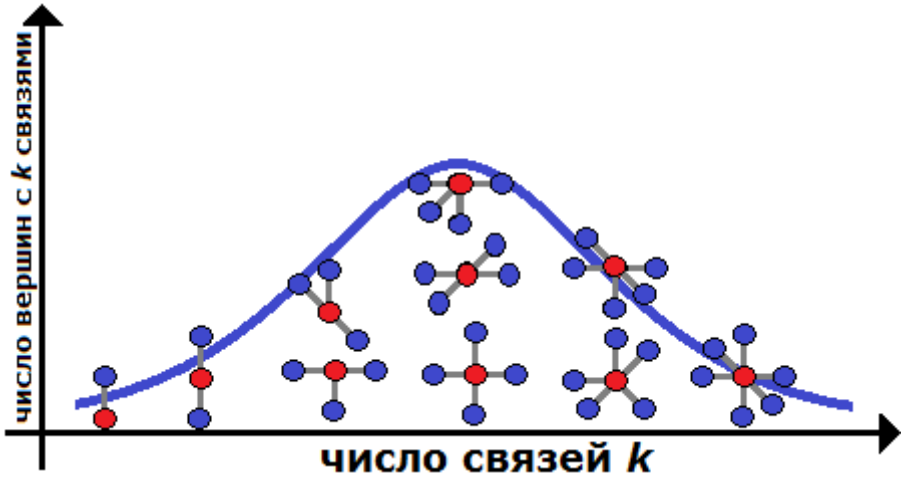
1. Распределение степеней вершин

Безмасштабные (scale-free) сети – сети, в которых степени вершин распределены по степенному закону:

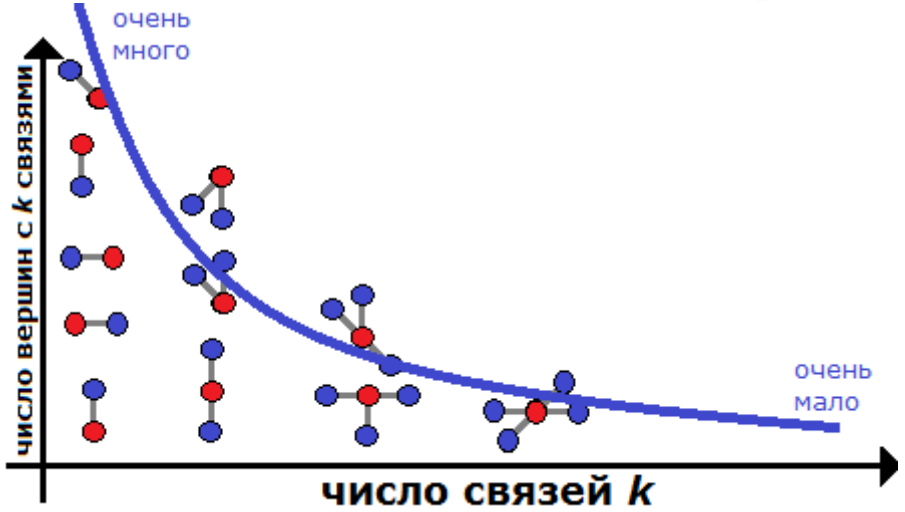
доля вершин с k связями ~ $k^{-\gamma}$,

обычно $2<\gamma<3$ и для k, начиная с некоторого

Случайный граф



Безмасштабный граф

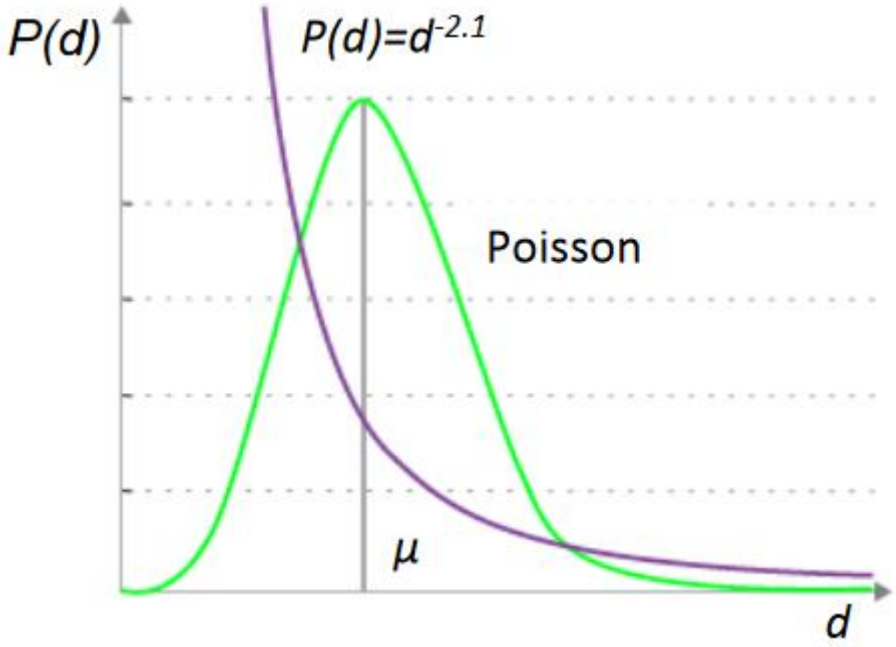


M.E.J. Newman «Power laws, Pareto distributions and Zipf's law» // Contemporary Physics, 2005, 46.5, pp. 323–351.

Безмасштабность (scale-free)

Функция $f(z)$ **безмасштабна, если** $f(\alpha z) = \beta f(z)$

как раз степенная…



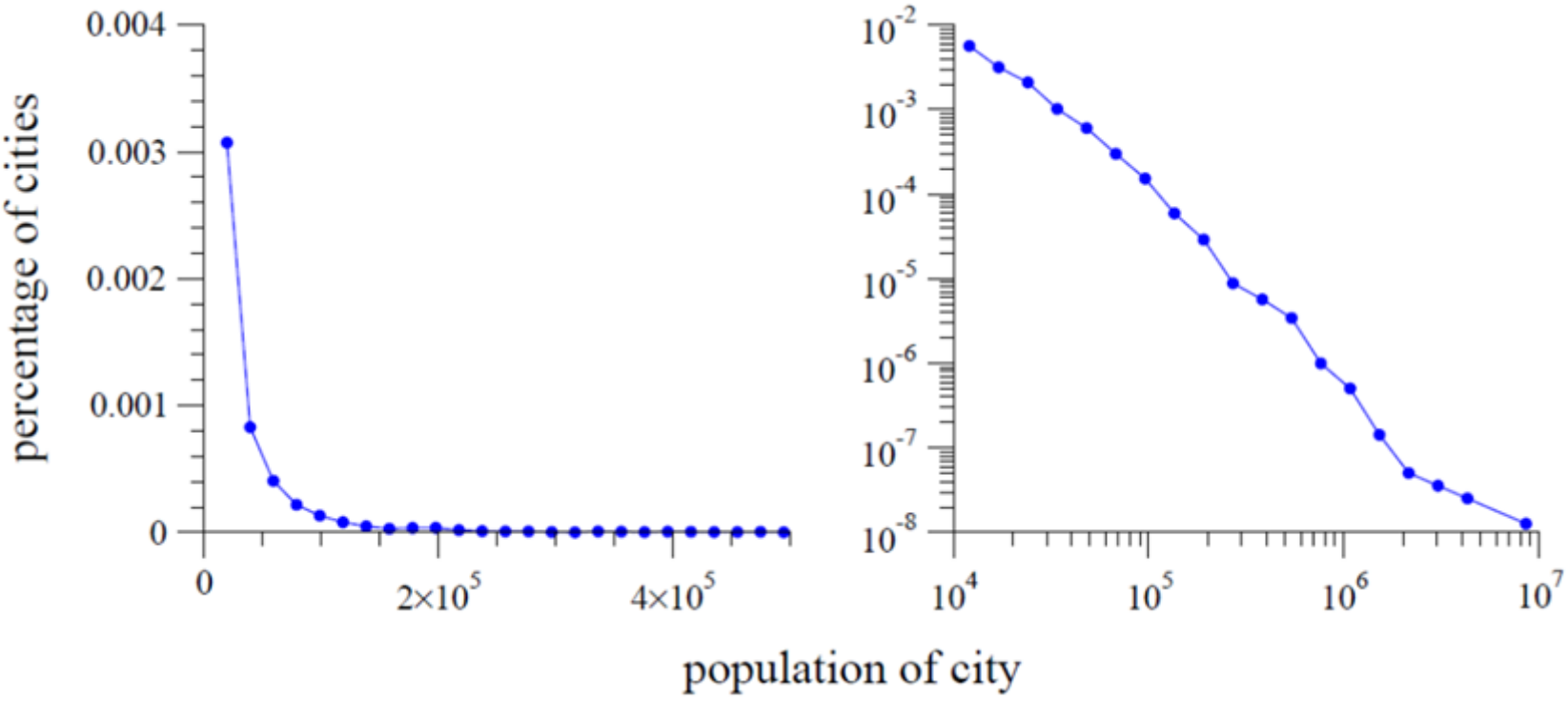
Случайный граф: степень случайной вершины $\mu \pm \sqrt{\mu}$

Безмасштабный: $\mu \pm \infty$

Степенной закон (power law)

Функция ~ $k^{-\gamma}$

доминирует, где измеряется «популярность»



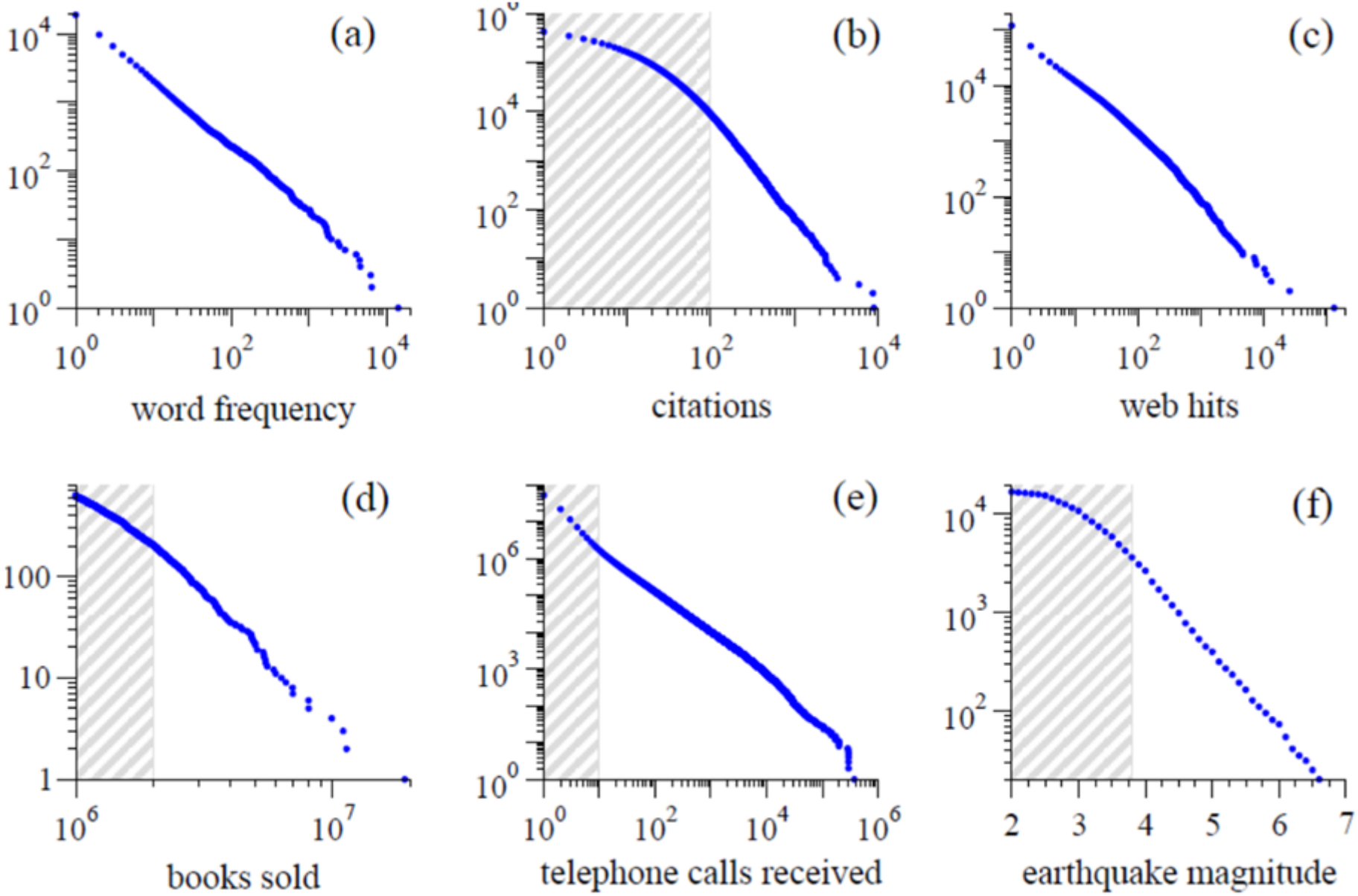
Надо смотреть в логарифмическом масштабе

(должна быть прямая линия)

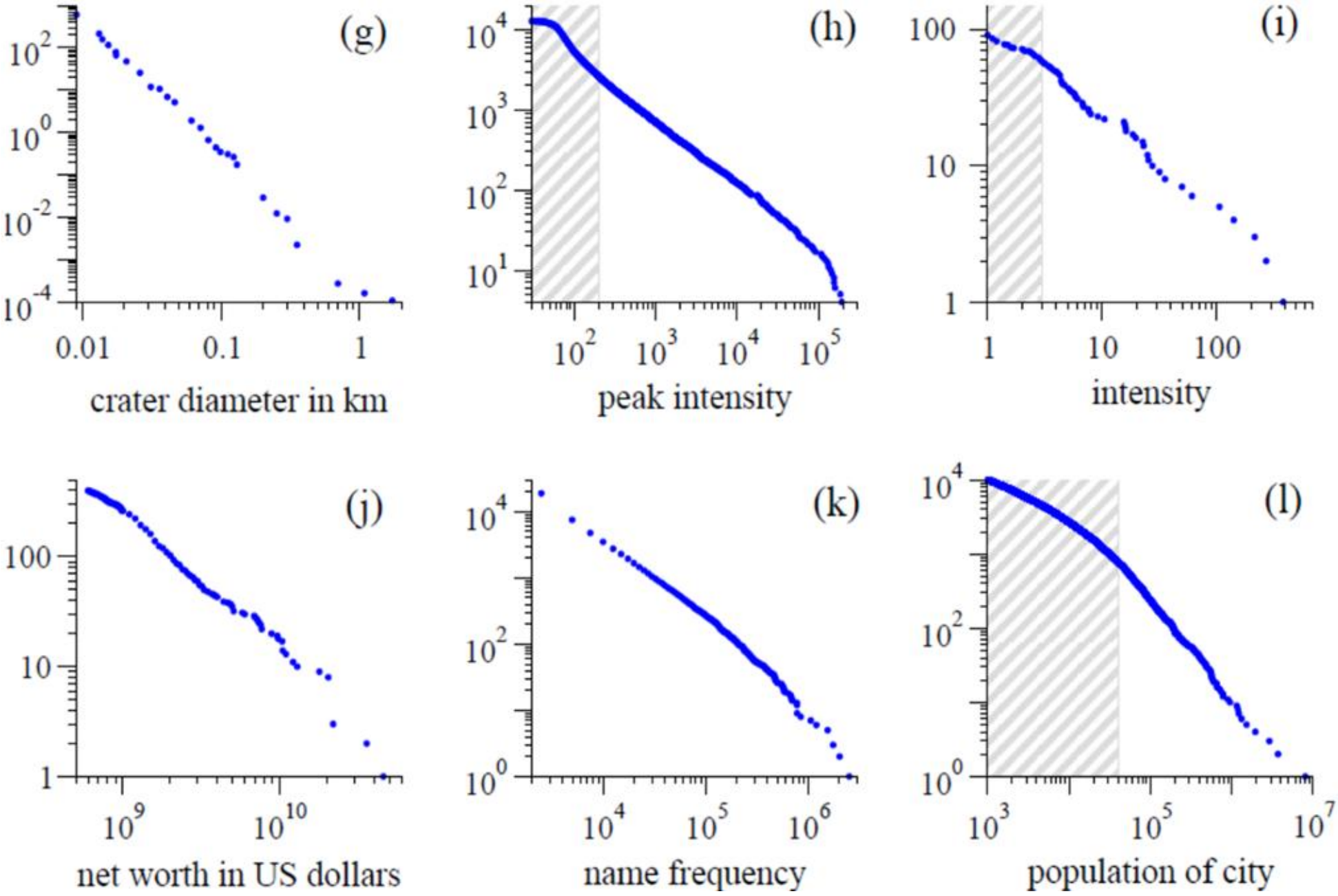
Закон Ципфа (Zipf's Law) Частота k-й по популярности буквы ~ 1/k

Степенной закон (power law)

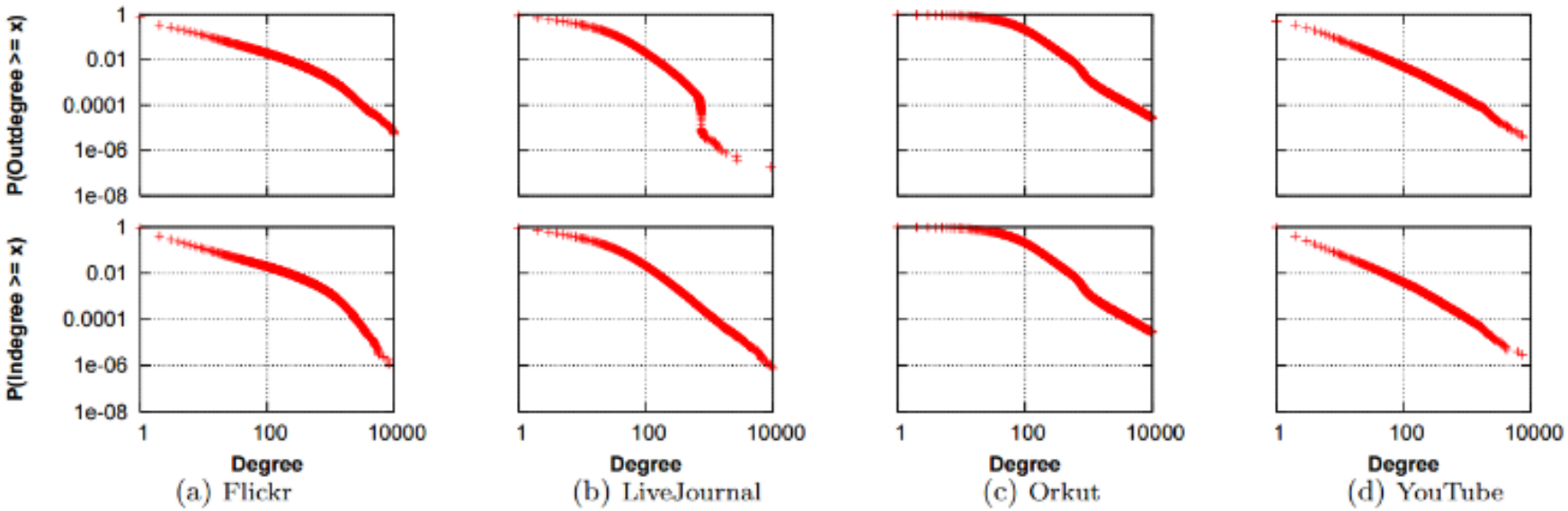
[Clauset et al ’07]



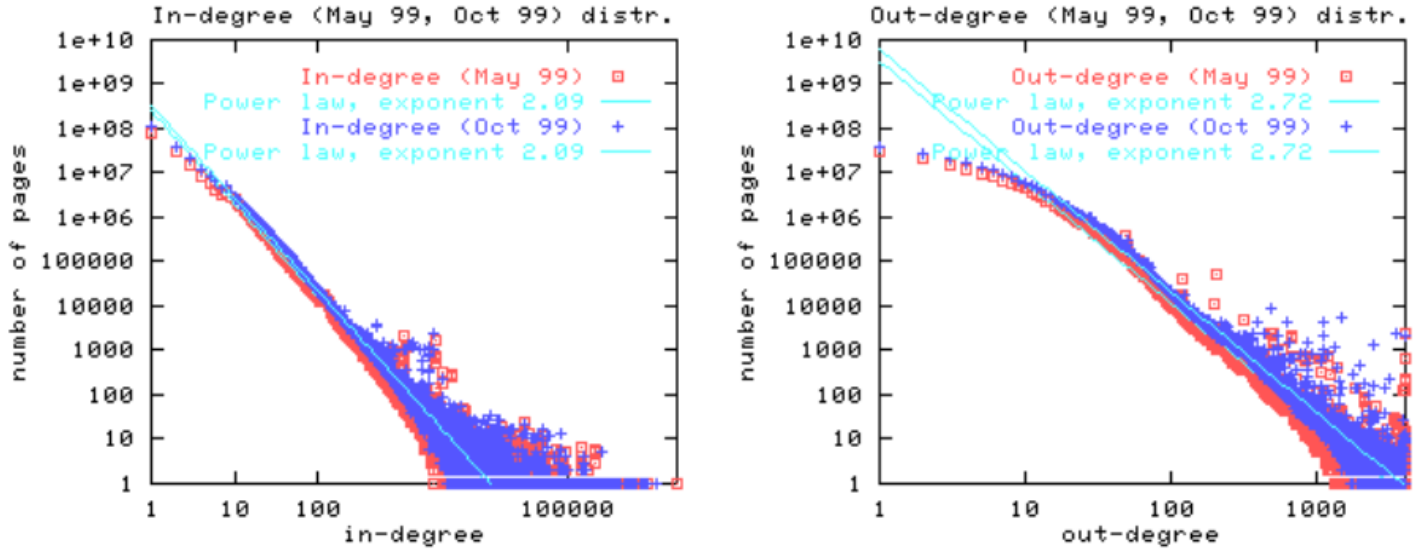
Степенной закон (power law)



1. Распределение степеней вершин



источник?

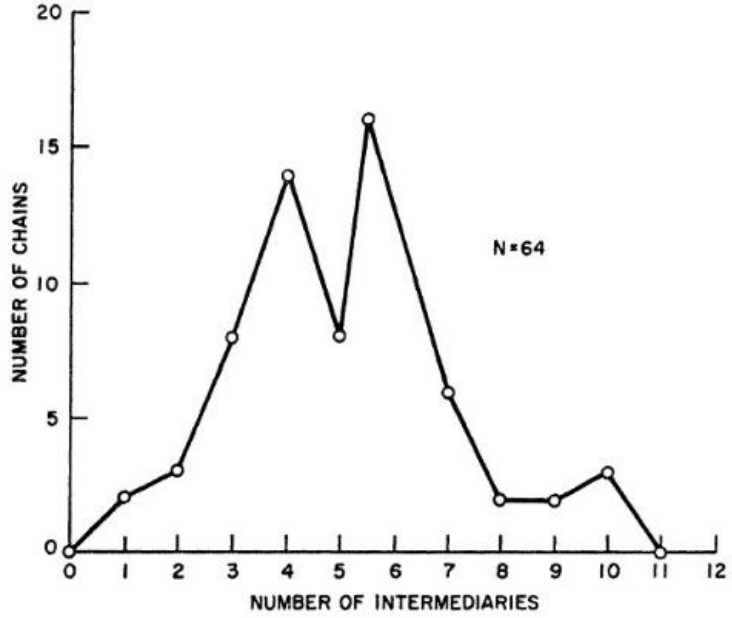


Degree distributions of the WWW analyzed in [Broder et al '00]

2. Модель малого мира

Stanley Milgram (1967)

Добровольцам задание – переслать письмо конкретному человеку:
Имя, адрес, род занятий
Но отправлять письмо можно только знакомому
дошло (64 из 296), медиана = 6, средний путь = 6.2

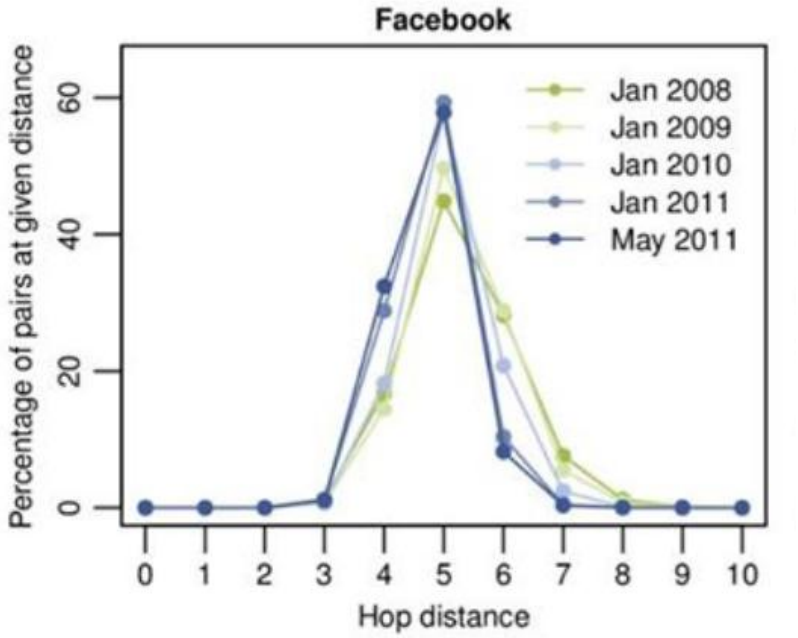


S. Milgram «The small-world problem» // Psychology Today, vol. 2, pp. 60-67, 1967

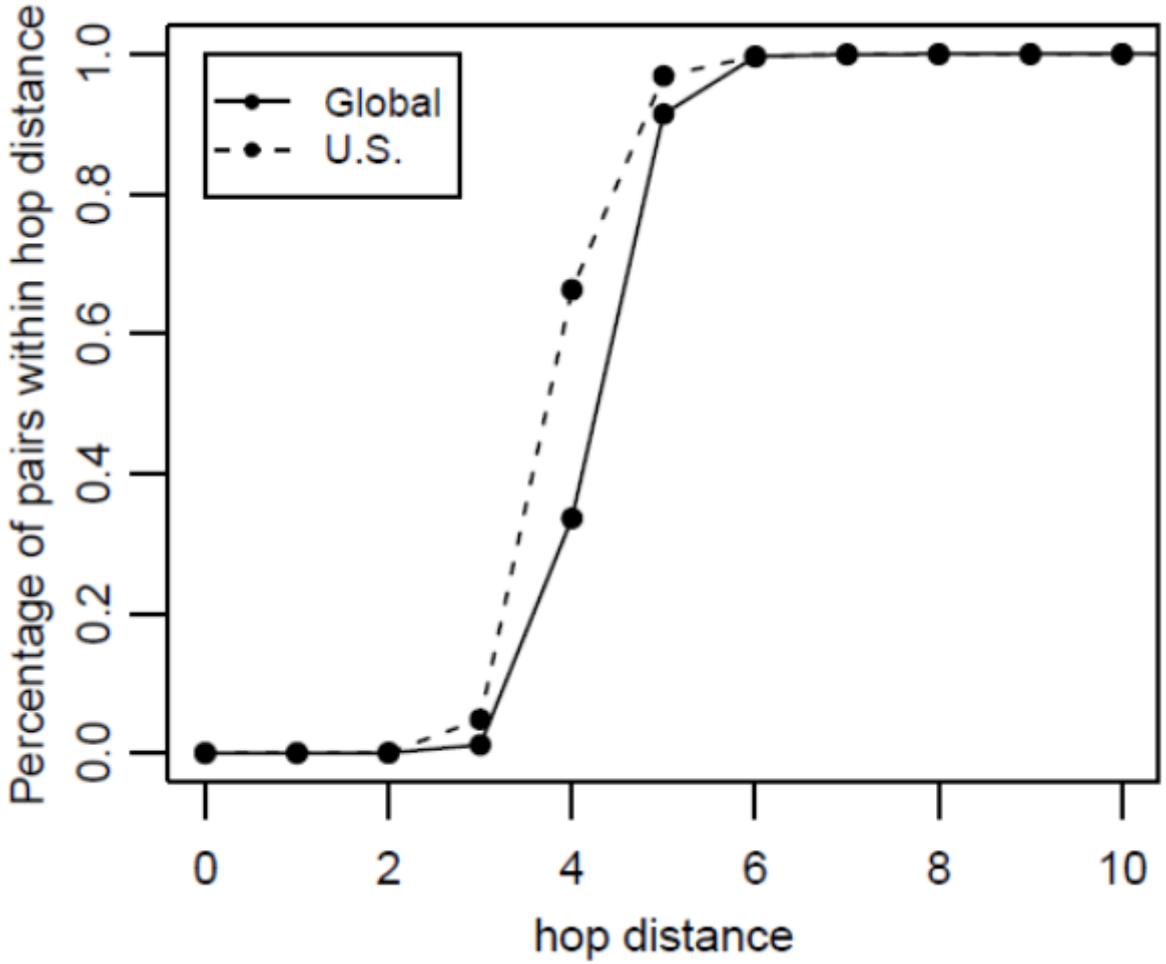
2. Модель малого мира

«Мир тесен» = «теория 6 рукопожатий»

Граф	Среднее расстояние между вершинами
Граф почтовых рассылок (D. Watts, 2001, 48000 вершин)	**6**
Граф сообщений в MSN Messenger (J. Lescovec и др. 2007, 240 млн. вершин)	**6.6**
Граф Фейсбука (L. Backstrom и др. 2012, 720 млн. вершин)	**4.74**



2. Модель малого мира



Вся сеть

92.0% на расстоянии <=5,

99.6% на расстоянии <=6.

США

96.0% на расстоянии <=5,

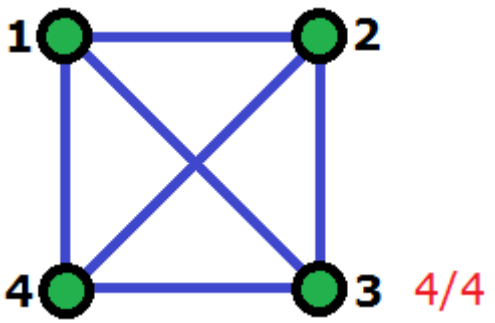
99.7% на расстоянии <=6.

3. Коэффициент кластеризации (полноты)

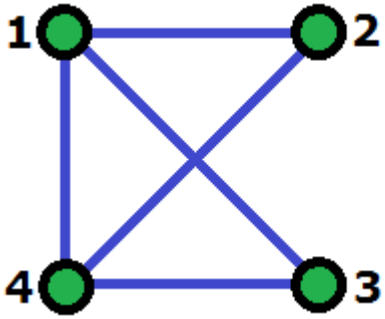
(CF = Clustering Coefficient)

1. Глобальный

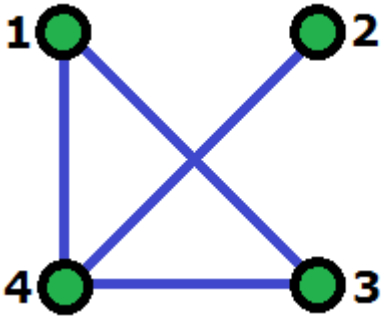
1.1. число треугольников / возможное число (число линий из трёх точек)



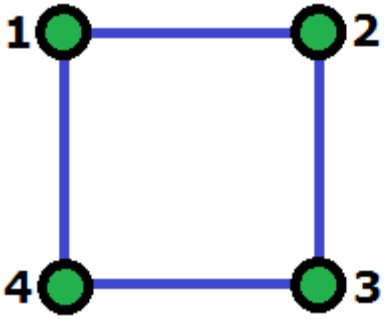
4/4



2/4



1/4



0/4

1.2. Среднее локальных коэффициентов

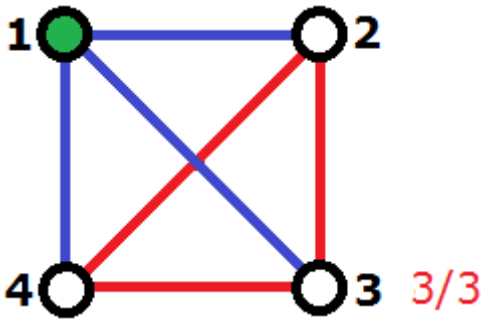
Внимание! Это признак;)

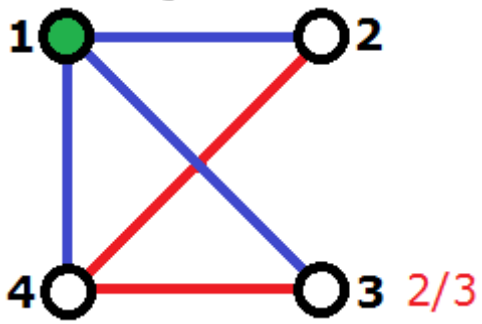
3. Коэффициент кластеризации (clustering coefficient)

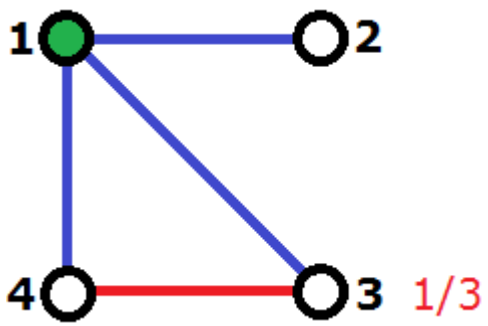
2. Локальный

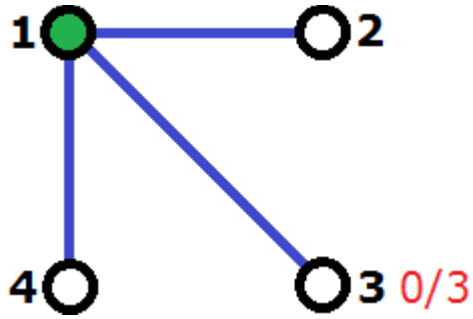
для вершины = насколько её соседи близки к образованию клики

число связей у соседей / число возможных связей







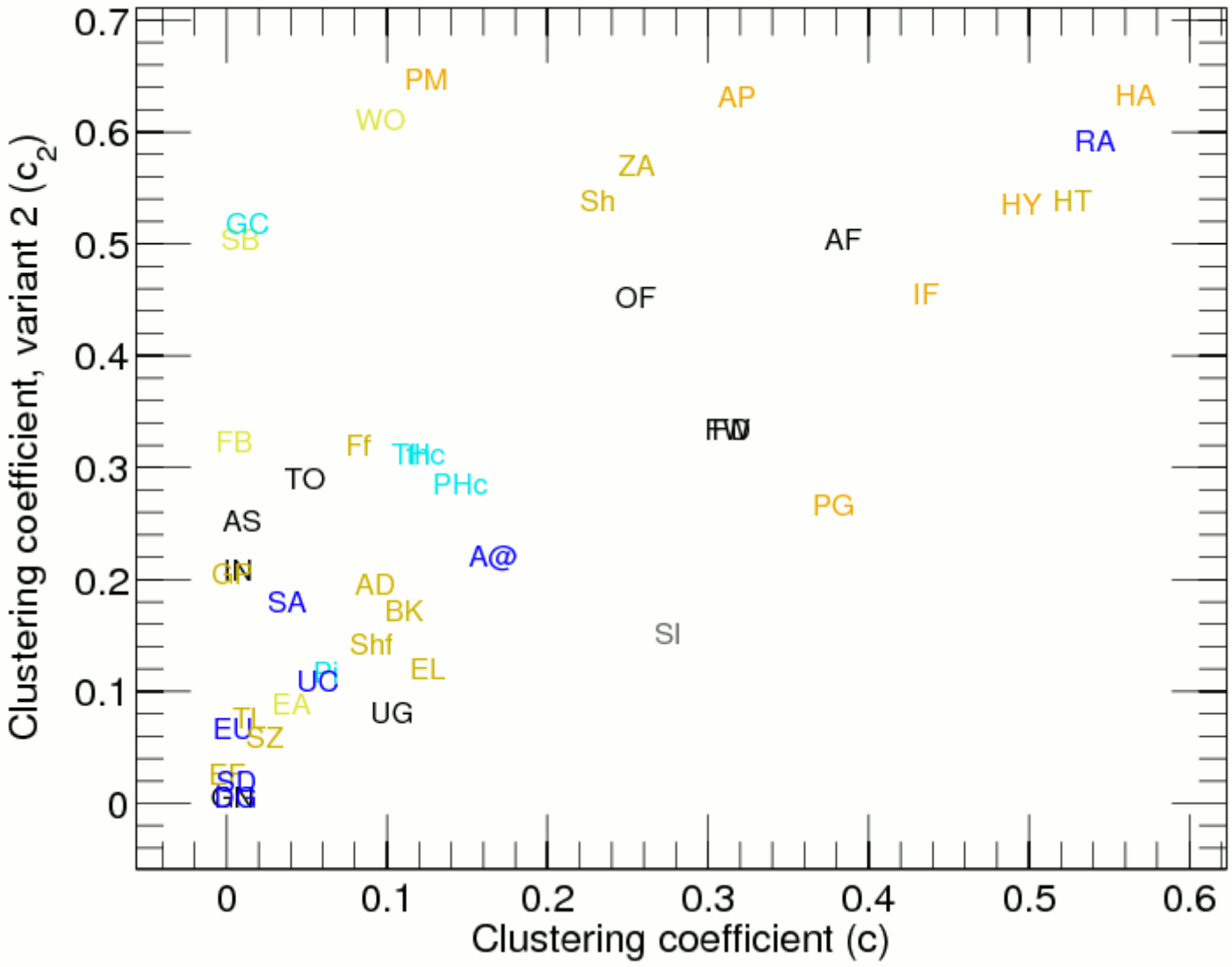


CF вершины A – вероятность дружбы двух случайных друзей A

Кстати, Bearman and Moody (2004)

Девочки-подростки с низким коэффициентом кластеризации более склонны к самоубийствам

3. Коэффициент кластеризации



Два способа определения коэффициента кластеризации

https://networkscience.wordpress.com/

4. Разреженность

Большинство реальных графов – разреженные (sparse).

Данные	Число вершин	Средняя степень
WWW (Stanford-Berkeley)	319,717	9.65
Social networks (LinkedIn)	6,946,668	8.87
Communication (MSN IM)	242,720,596	11.1
Coauthorships (DBLP)	317,080	6.62
Internet (AS-Skitter)	1,719,037	14.91
Roads (California)	1,957,027	2.82
Proteins (S. Cerevisiae)	1,870	2.39

из Leskovec et al., Internet Mathematics, 2009

5. Теория связей

Экспериметр Грановеттера
(Granovetter's Experiment) – 1960

«Сила слабых связей»
– мощный механизм социальной мобильности

Люди ищут работу через контакты,
но чаще через знакомых,
а не друзей

Getting A Job: A Study of Contacts and Careers. — Harvard University/ University of Chicago Press, 1974.

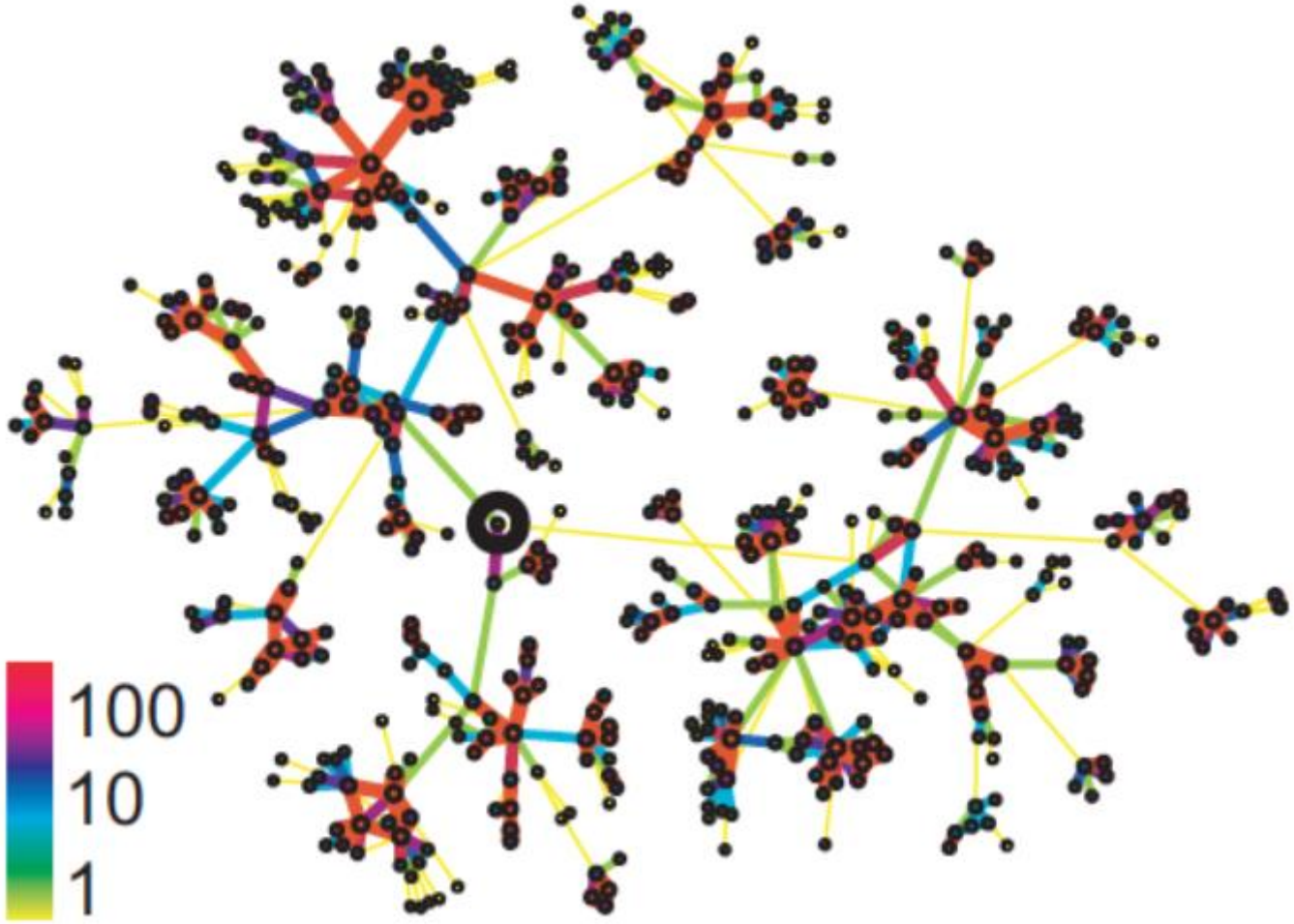
Сейчас есть возможность проверить…
как формализовать «друг» / «знакомый»

5. Теория связей

Сеть сотовой связи

(A, B), если были звонки A → B и B→ A за определённый период

«степень дружбы» – число звонков / средняя продолжительность / …



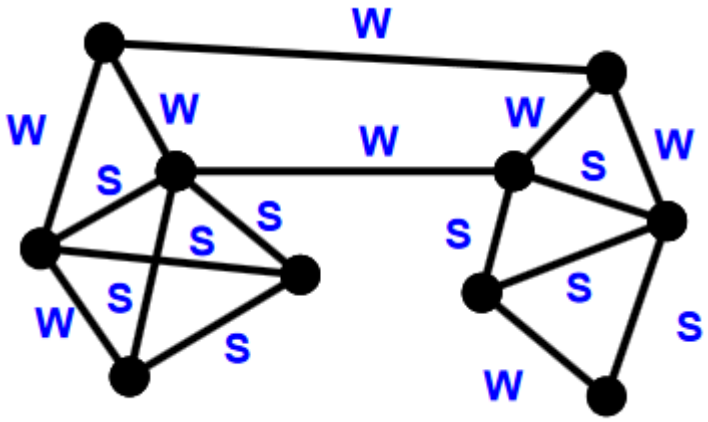
Картинка как при делении на сообщества;)

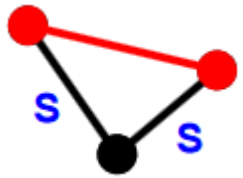
5. Теория связей

Сильные связи (Strong ties) больше отвечают за кластерную структуры

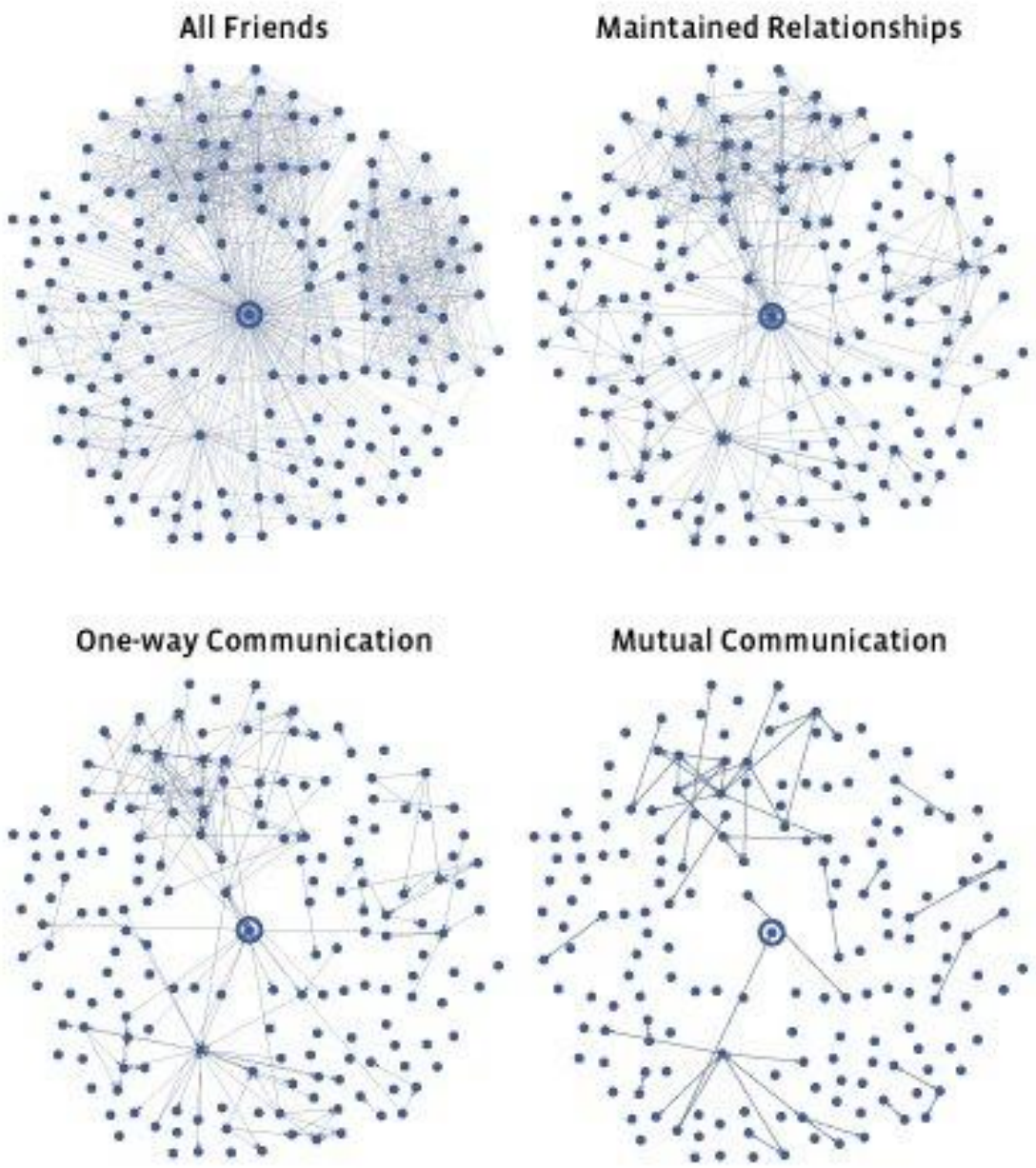
Слабые (Weak ties) – соединения сообществ

J. P. Onella et al., «Structure and tie strengths in mobile communication networks» PNAS, vol. 104, pp. 7332-7336, 2007





5. Теория связей: Maintained Relationships on Facebook



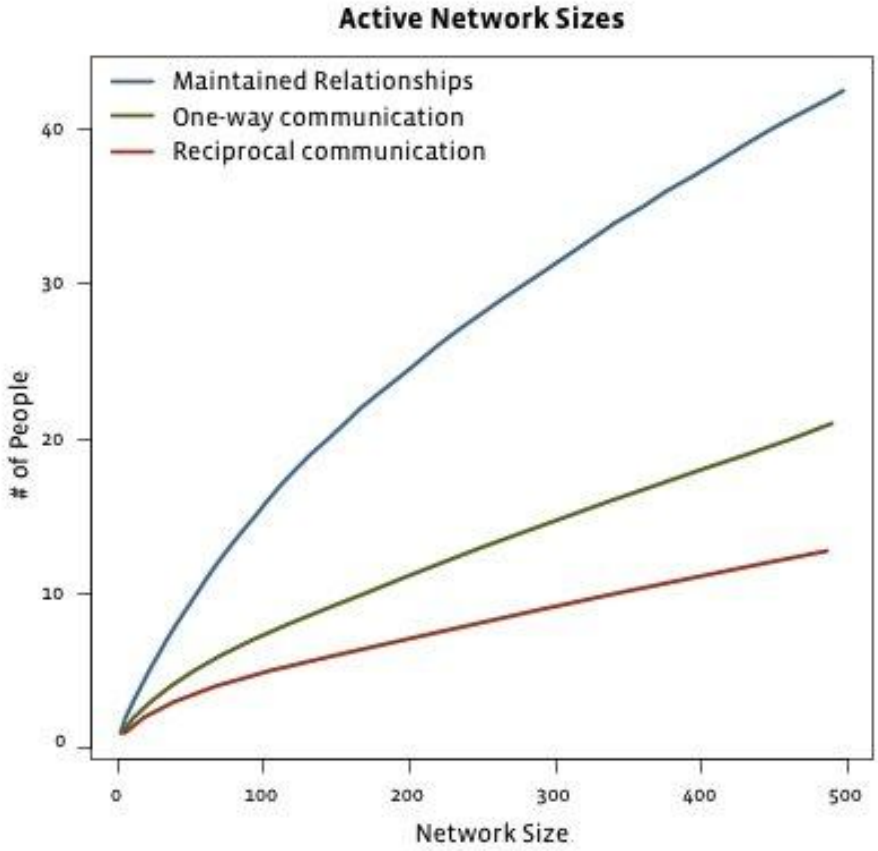
5. Теория связей: Maintained Relationships on Facebook

All Friends – все друзья

Reciprocal Communication – двусторонняя активная коммуникация

One-way Communication – односторонняя активная коммуникация

Maintained Relationships – «вовлечённость» – просматривал новости или профиль > 2 раз



5. Теория связей

Число Данбара

Робин Данбар (Robin Ian MacDonald Dunbar)

Внутренний круг = 5

Симпатии = 12-15

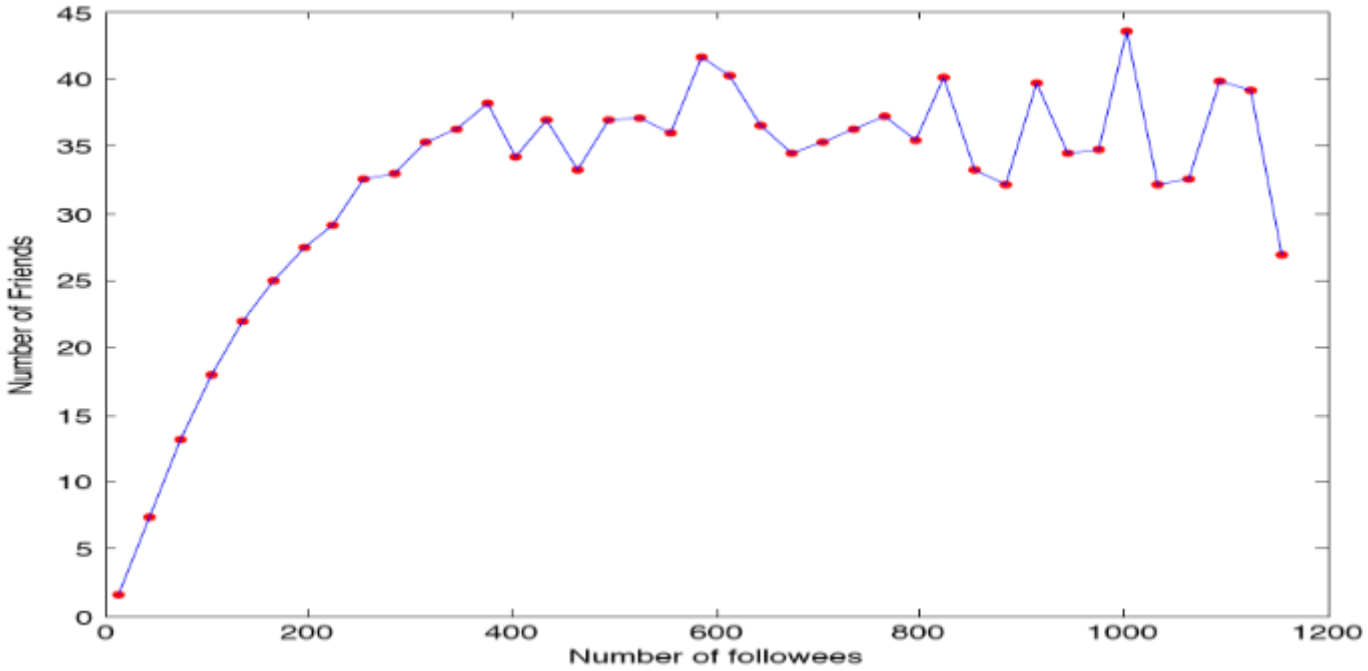
Полу-регулярная группа = 50

Стабильная социальная группа = 150 (число Данбара)

Друзья друзей (слабые связи) = 500

5. Теория связей

Сильные связи требуют времени и энергии для их поддержания

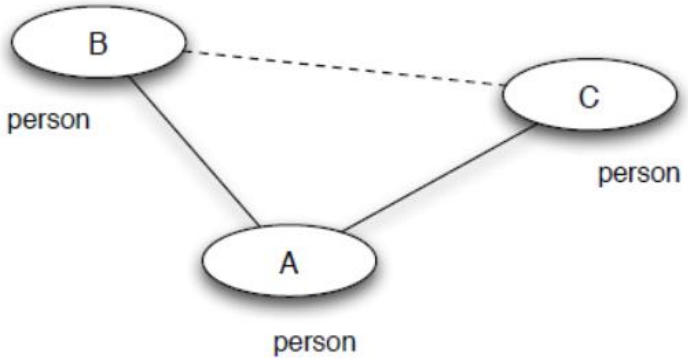


Twitter under the microscope.Huberman et al. 2008.
arxiv.org/pdf/0812.1045.pdf

5. Теория связей

Какие факторы доминирующие для создания связей

1. Triadic closure
«друг моего друга»



2. Homophily / assortative mixing

Гомофилия – принцип выбора друзей, по которому мы стараемся выбирать из себе подобных и быть похожими на друзей

«похожий на меня по интересам»
раса / возраст / хобби

3. Тяготение к важным вершинам

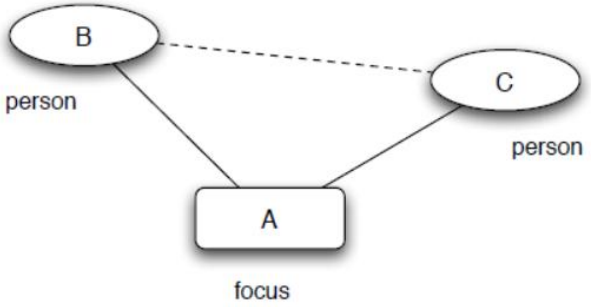
4. Случайные связи

5. Теория связей

Гомофилия

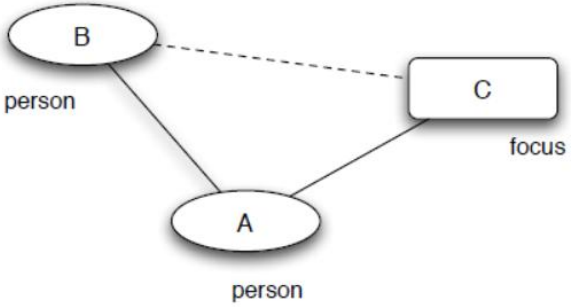
Селекция (Selection)

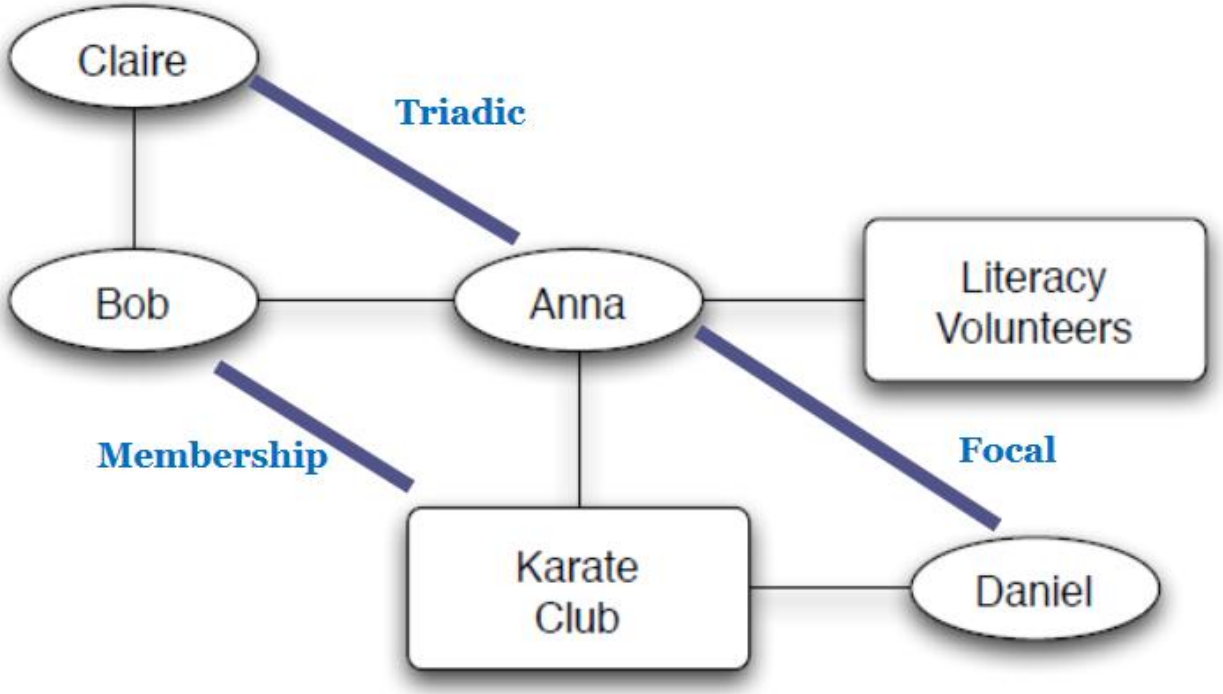
выбор людей «таких как я»



Социальное влияние

Сам адаптируюсь под других





5. Теория связей

Оценка гомофилии

Пусть в сообществе p **– вероятность быть мужчиной,**

$q = 1 - p$ **– женщиной.**

Если выбираем рёбра случайно, то вероятности

$$P(M,M) = p \cdot p$$
$$P(M,F) = p \cdot q$$
$$P(F,M) = q \cdot p$$
$$P(F,F) = q \cdot q$$

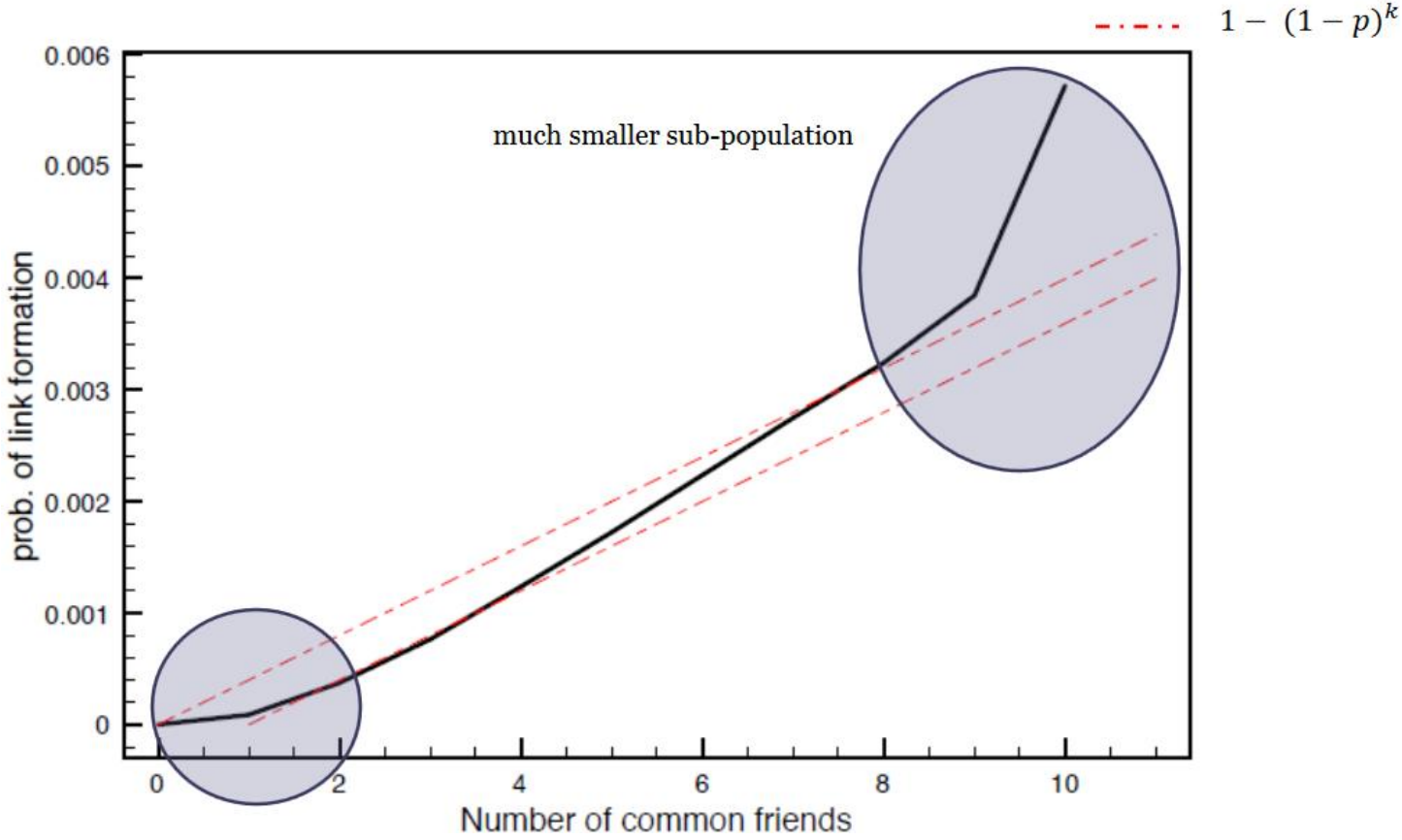
Можно сравнить процент дружбы разных полов с $2pq$

5. Теория связей

Оценка гомофилии

Пусть один общий друг порождает вероятность p подружиться,

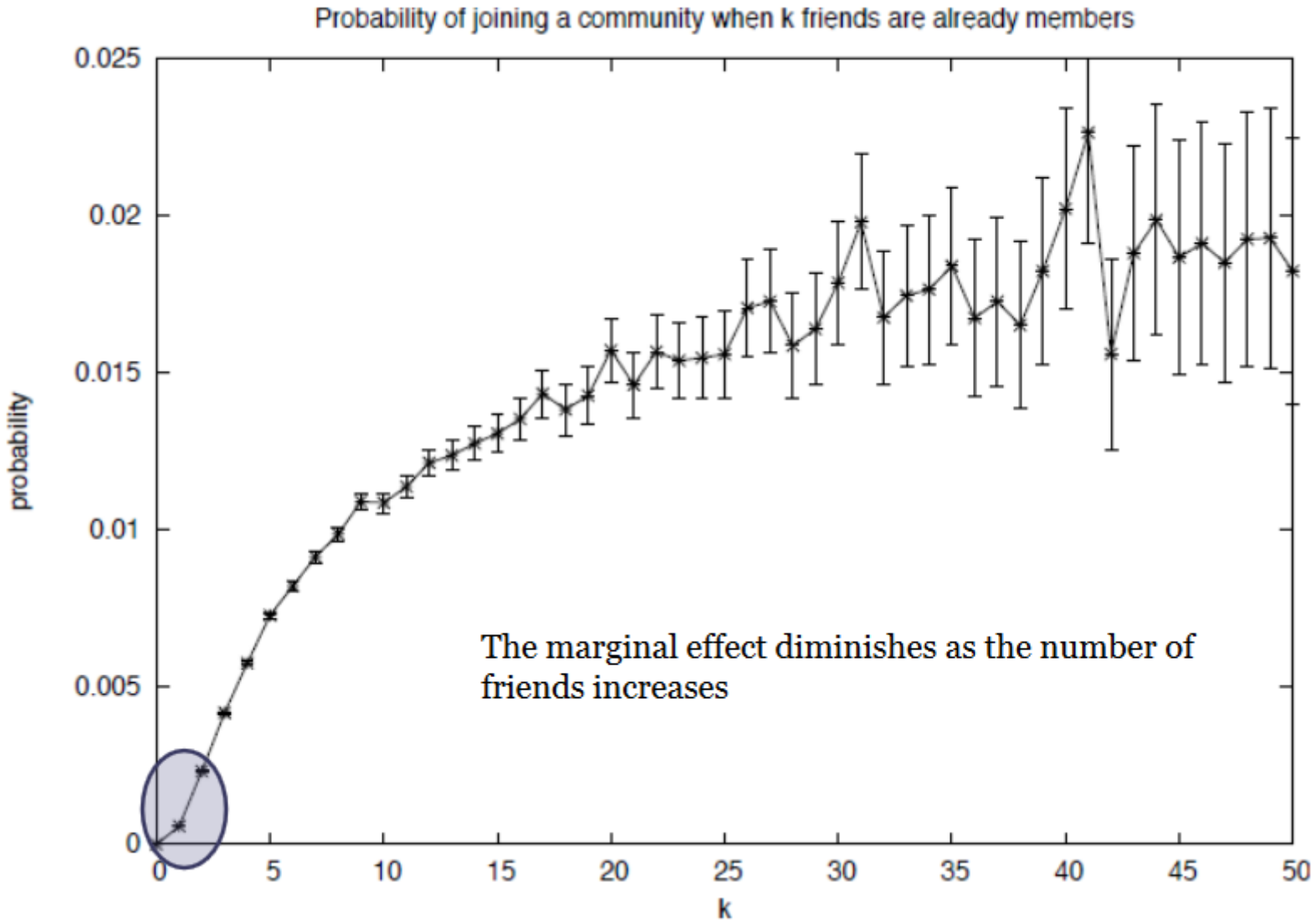
тогда k общих друзей – $1-(1-p)^k$



Gueorgi, and Watts «Empirical analysis of an evolving social network» // Science, 2006

5. Теория связей

Оценка социального влияния

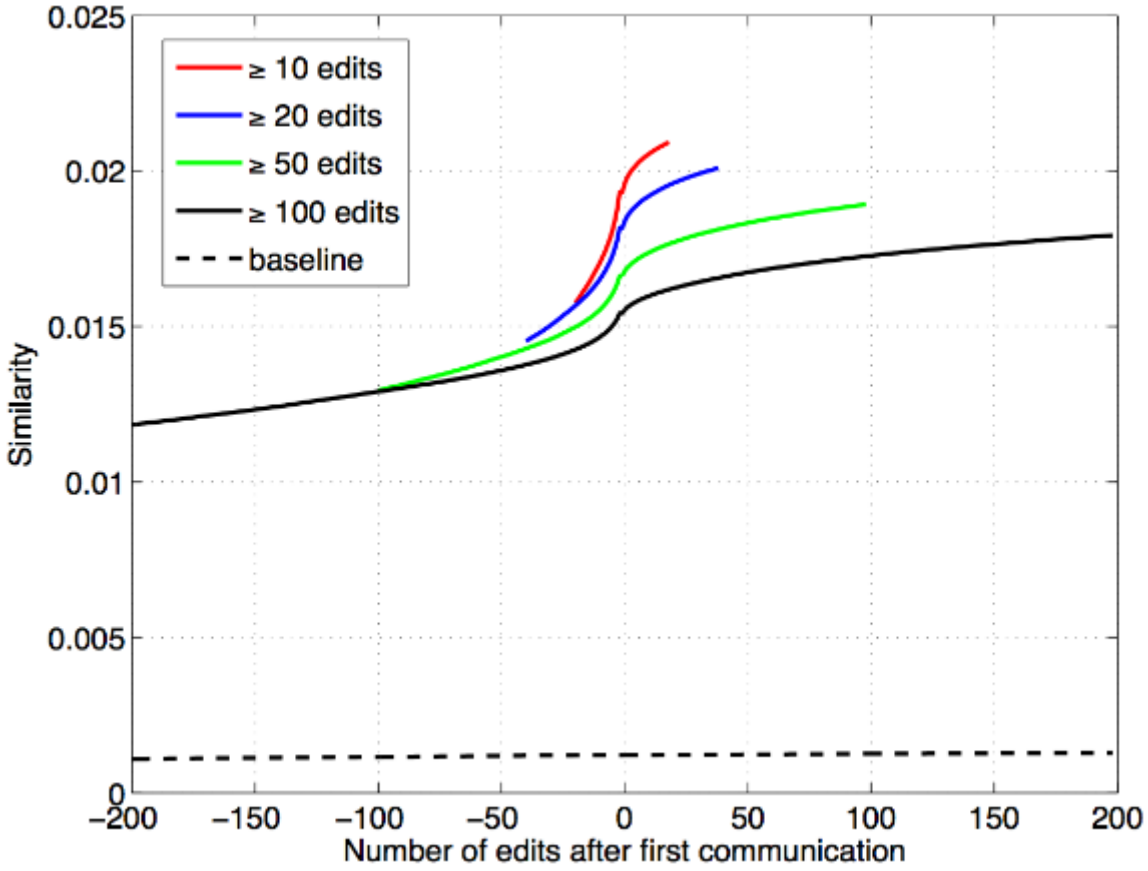


Backstrom, et al «Group formation in large social networks: Membership, growth, and evolution» // SIGKDD 2006

5. Теория связей

Оценка социального влияния

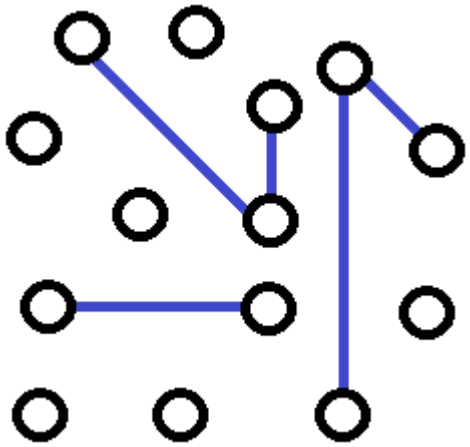
Редактирование Википедии



Feedback effects between similarity and social influence in online communities. Crandall, et al., SIGKDD 2008

Моделирование графов

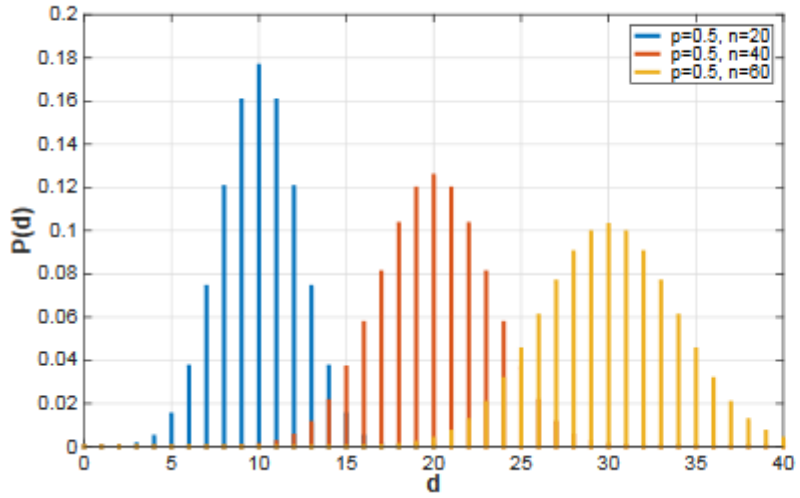
Генерация случайных графов: модель Эрдёша–Реньи (Erdös-Renyi)

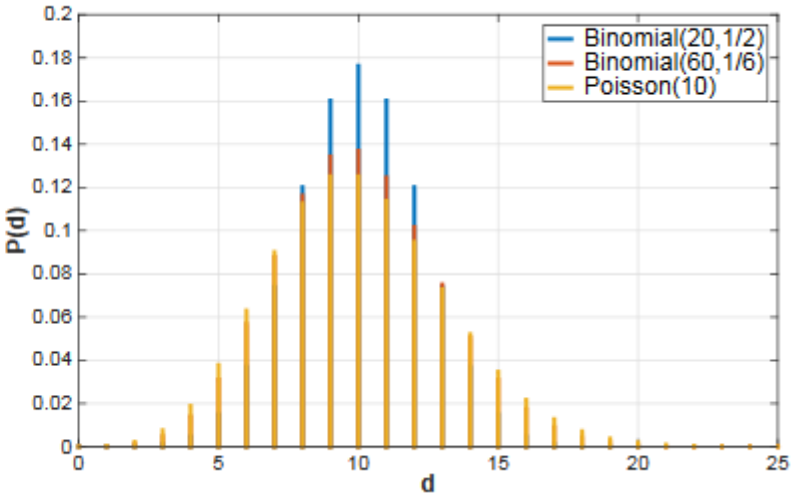


$$P(d^{\text{in}} = d) = C_{n-1}^{d} p^{d} (1-p)^{(n-1)-d}$$

~ сумма n-1 бернуллиевских величин, по ЦПТ

$$\text{norm}(np, np(1-p))$$





если $(n-1)p = \mu = \text{const}$**, то** $P(d) \xrightarrow[n\to\infty]{} e^{-\mu} \dfrac{\mu^{d}}{d!}$

Моделирование графов

модель Эрдёша–Реньи (Erdös-Renyi)

если $np > 1$**, то** $G_{n,p}$ **почти всегда имеет компоненту размера** $O(n)$

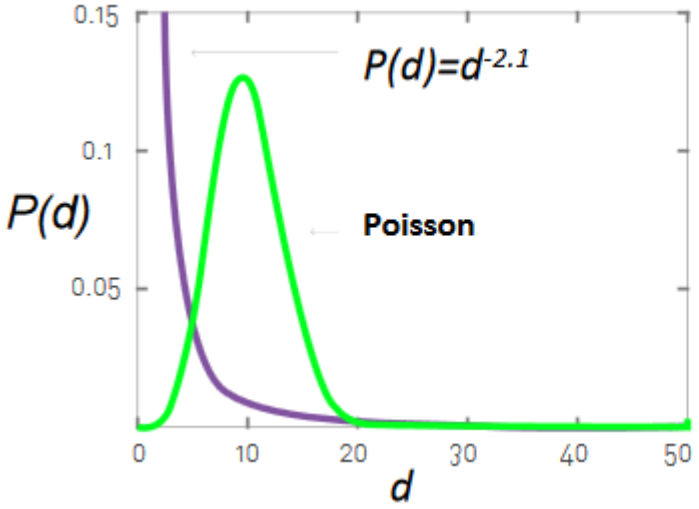
если $np < 1$**, то** $G_{n,p}$ **почти всегда размер компонент не выше** $O(\log n)$

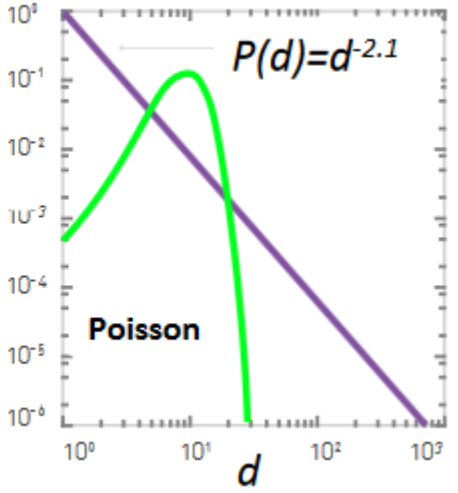
Маленький кластерный коэффициент $O(n^{-1})$

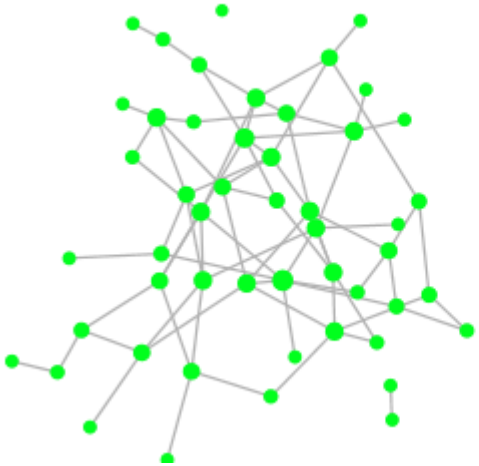
Диаметр $O(\log n)$

Моделирование графов

Что в реальной жизни…



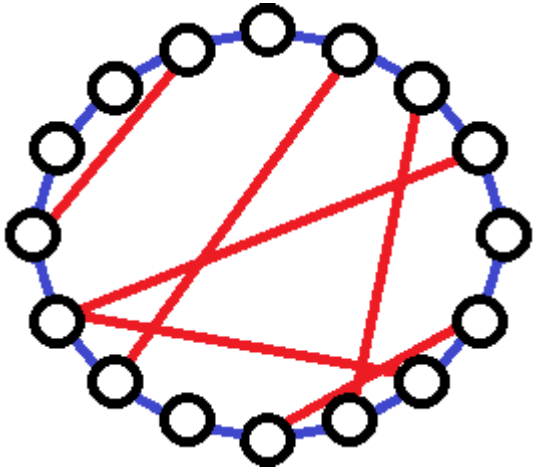




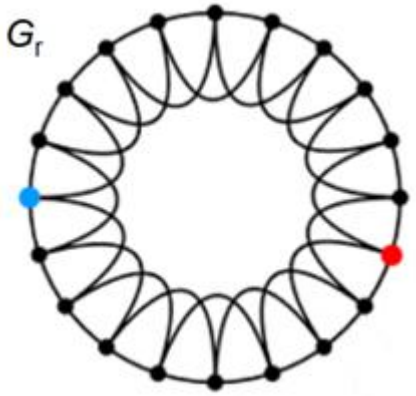
Моделирование графов

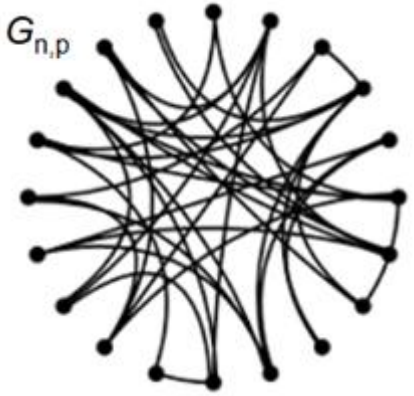
Watts–Strogatz model

Генерация случайных графов: модель малого мира



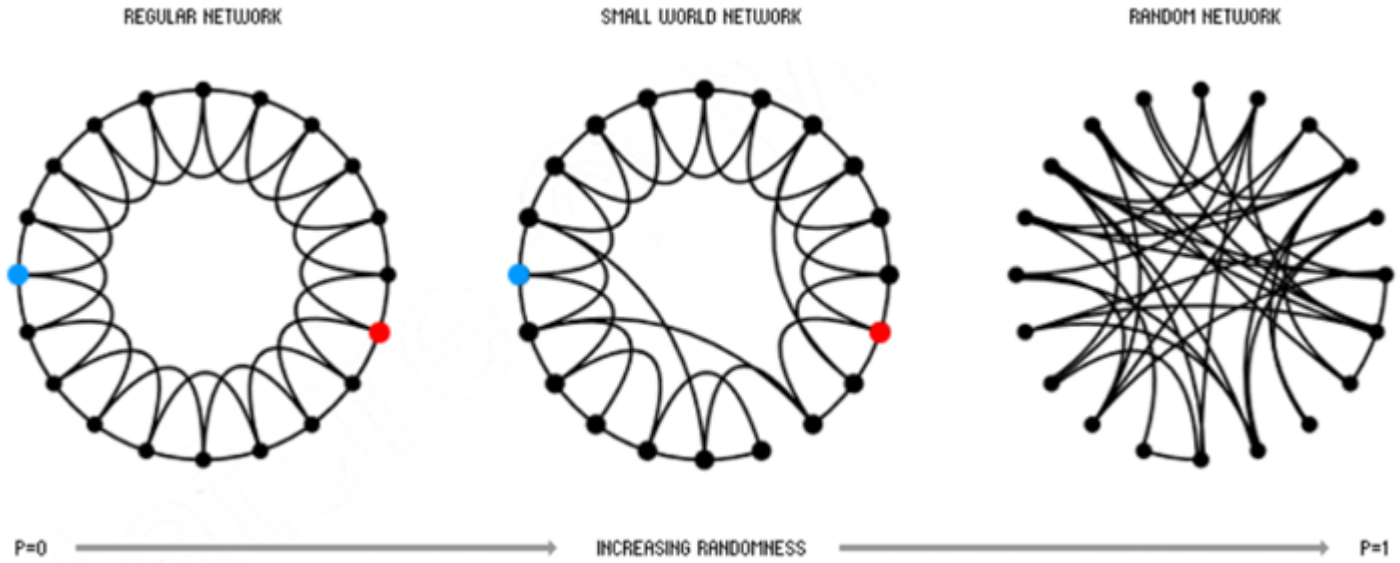
Моделирование графов





большая кластеризация и диаметр

малая кластеризация и диаметр



ДЗ Исследовать, когда «похож не реальный граф»

(малый диаметр и большой коэффициент кластеризации)

Моделирование графов

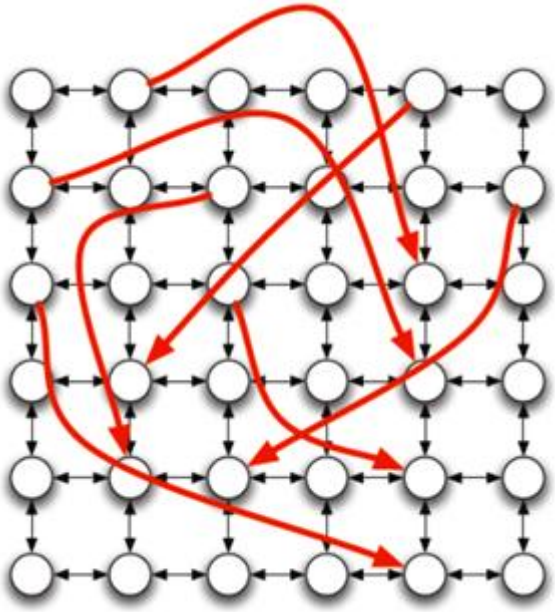
Причина малого мира Watts-Strogatz (1998):

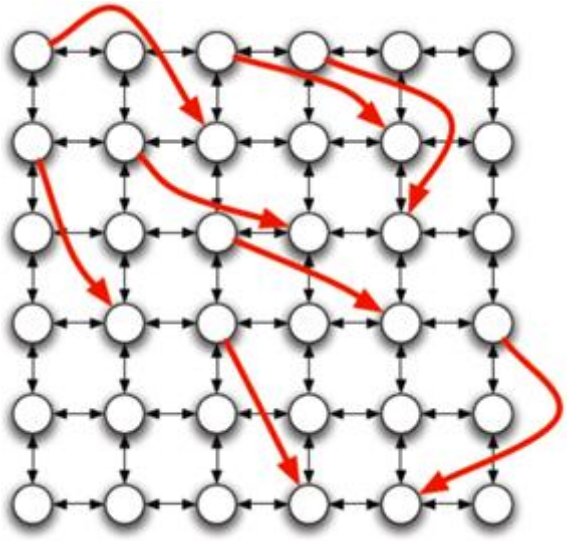
Гомофилия ⇒ сильно-связные графы (коэф. кластеризации)

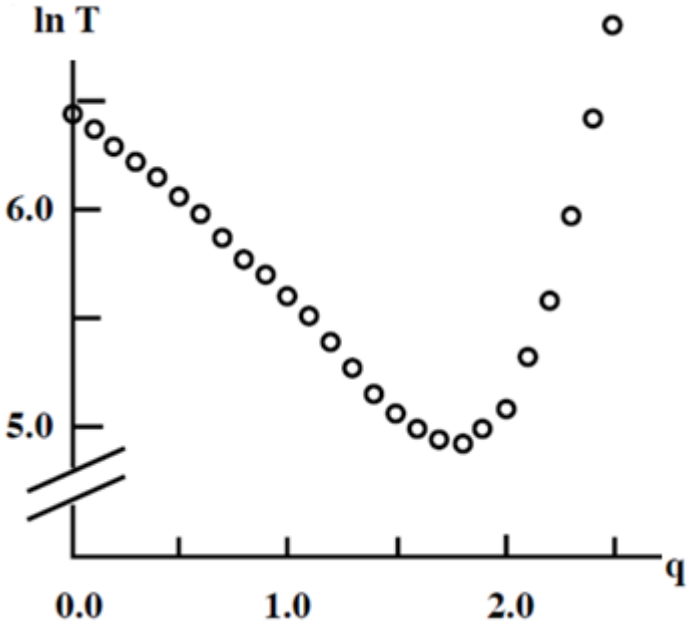
это локальная структура

Слабые связи ⇒ короткий путь в другие сообщества

это случайные связи







Связь с вероятностью $\propto d(i,j)^{-q}$ Минимальное ожидаемое время доставки, когда q ~ 2

Моделирование графов: Преимущественное присоединение

Создаём вершины $1, 2, \ldots$

Когда создана j**-я она соединяется с** $i : i < j$

с вероятностью p i**-я вершина выбирается случайно:**

$$P(j \to i) = \frac{1}{j-1}$$

с вероятностью $1-p$

$$P(j \to i) \propto \deg(i)$$

(есть вариант: дуга на вершину, на которую указывает случайная i)

Приводит к динамике «rich-gets-richer»

$$P(d^{\text{in}} = d) \propto d^{-\left(1+\frac{1}{1-p}\right)}$$

Преимущественное присоединение (preferential attachment) – к страницам, которые уже популярны

(такой реальный механизм популярности)

Преимущественное присоединение

Популярность – большая случайность

Если «всё переиграть» популярности поменяются

Но степенной закон остаётся

Эксперимент:

при заходе человек видит число скачиваний песен и нет

(много копий сайта)

Видит	Не видит
«популярные становятся популярнее»	значительно меньше популярность
причём по-разному!	нет её вариативности по копиям сайта

Salganik et. al. «Experimental study of inequality and unpredictability in an artificial cultural market. science 2006

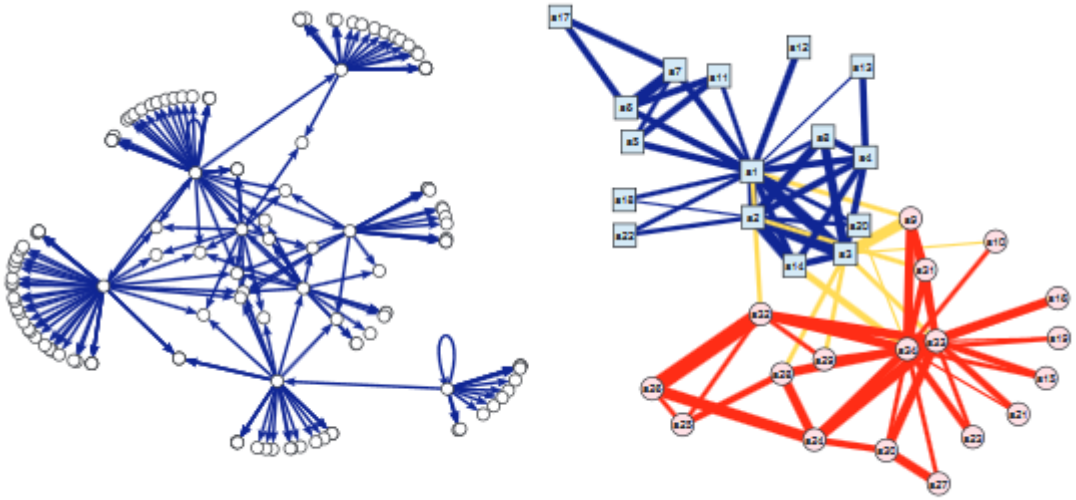
Визуализация графов

Spring-embedder methods

вершины – шарики, рёбра – стержни, надо добиться равновесия

Energy-placement methods

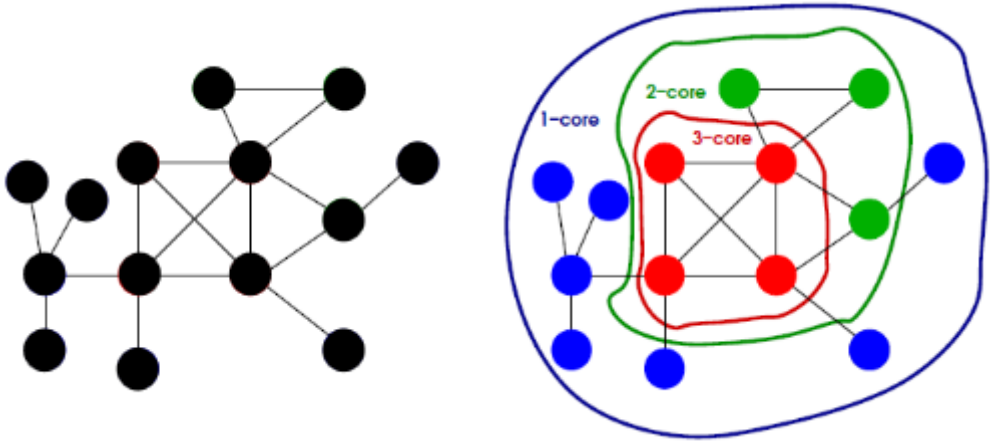
определяется функция энергии позиций вершин, минимизируется



Spring embedder Energy placement

B. Baingana and G. B. Giannakis «Centrality-constrained graph embedding» // ICASSP, 2013.

Визуализация графов



J. I. Alvarez-Hamelin et al, «Large scale networks fingerprinting and visualization using the k-core decomposition» // NIPS, 2005

Алгоритм: рекурсивно удалять рёбра степени меньше k, сложность

$$O(n_v + n_e)$$

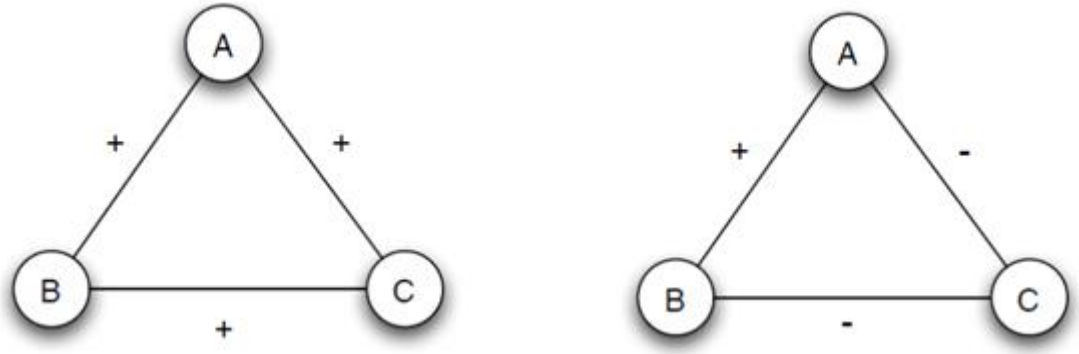
Сети с негативными связями

	вершин	рёбер	+ / –	связи
	119,217	841,200	85 / 15 %	Support / oppose
	82,144	549,202	77 / 23 %	Trust / Distrust
	7,118	103,747	79 / 21 %	Friends / Foe

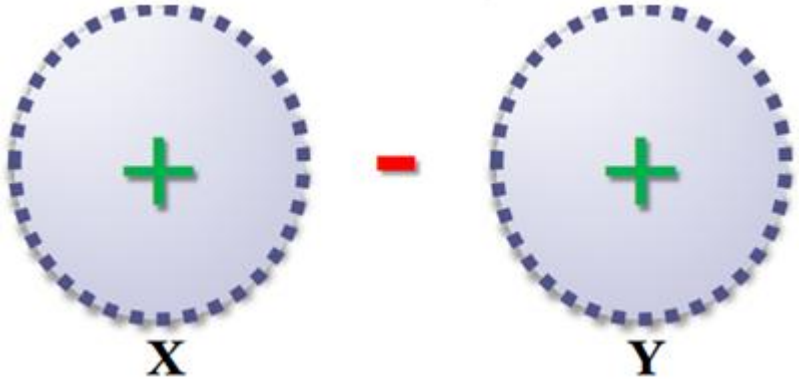
Guha, et. al. «Propagation of trust and distrust» WWW 2004

Сети с негативными связями

Размеченный полный граф сбалансированный (balanced), если каждый треугольник сбалансированный:



Размеченный полный граф сбалансированный ⇔ полностью положительный или разбивается на два сообщества (внутри сообщества все связи –, между сообществами +)



ДЗ – доказать

Сети с негативными связями

Первая мировая война

Fr: France

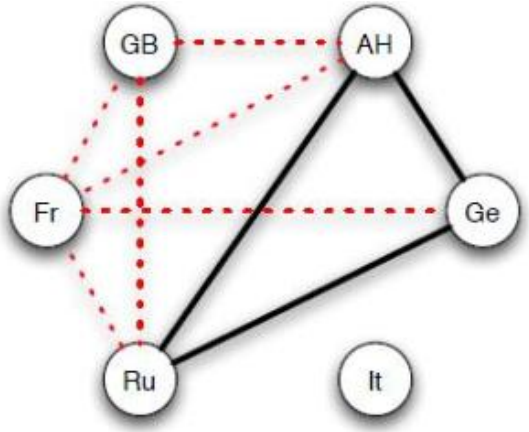
Ru: Russia

It: Italy

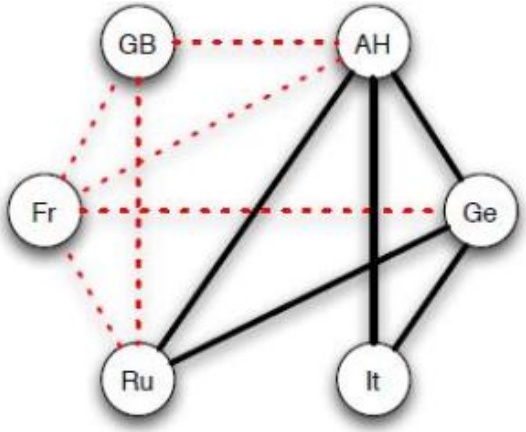
Ge: Germany

AH: Austria-Hungary

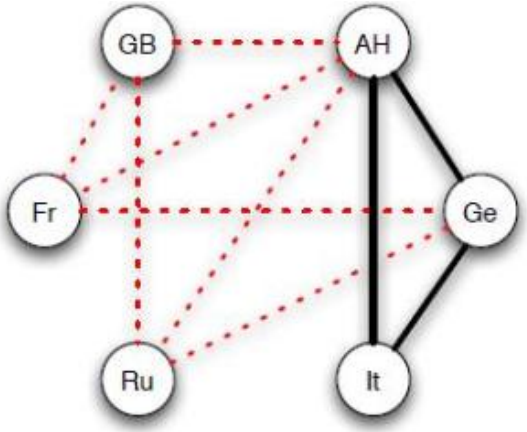
GB: Great Britain



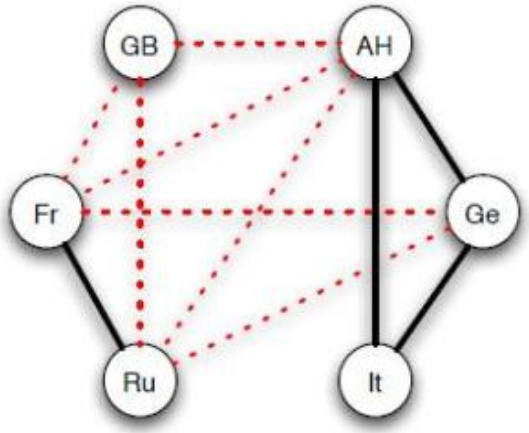
(a) *Three Emperors' League 1872–81*



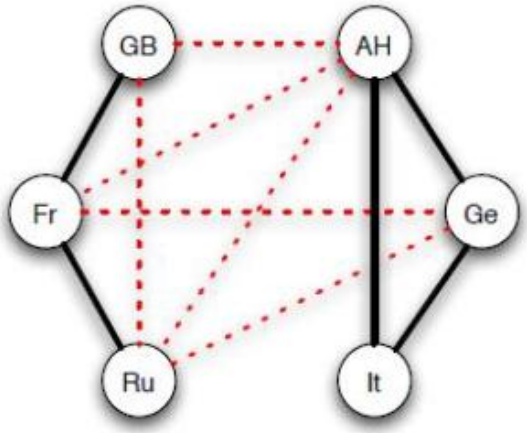
(b) *Triple Alliance 1882*



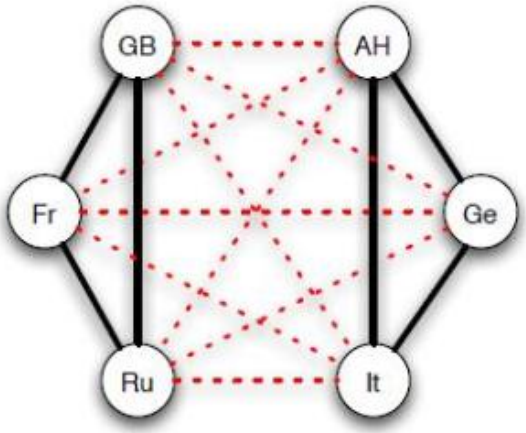
(c) *German-Russian Lapse 1890*



(d) *French-Russian Alliance 1891–94*



(e) *Entente Cordiale 1904*

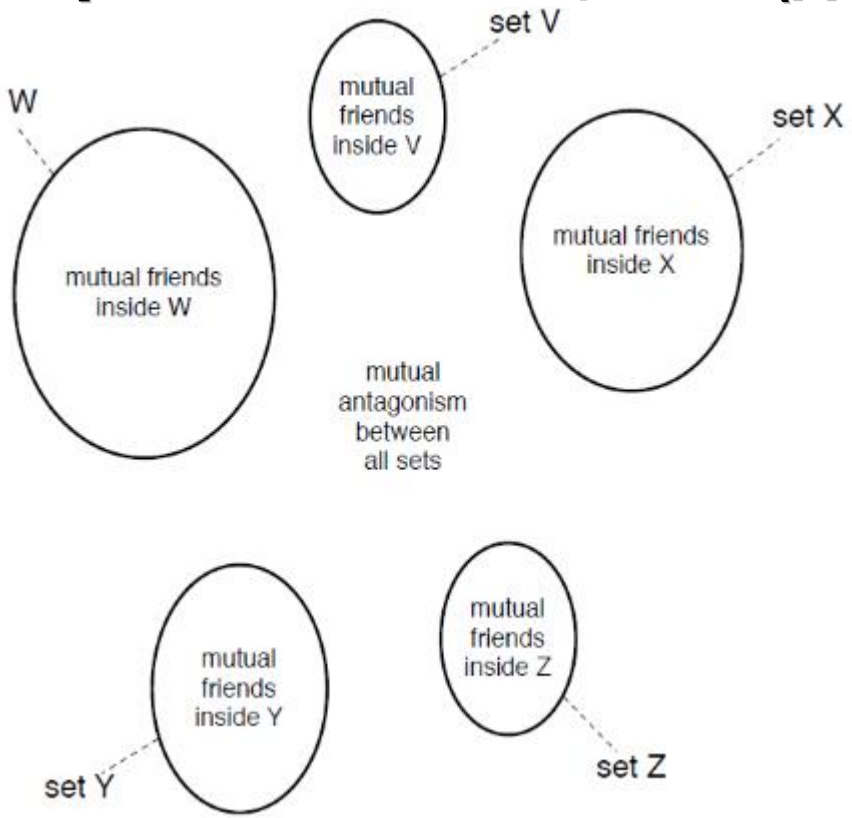


(f) *British Russian Alliance 1907*

Сети с негативными связями

Размеченный полный граф слабо сбалансирован (Weak Structural Balance Property) – нет треугольника с ровно двумя положительными рёбрами

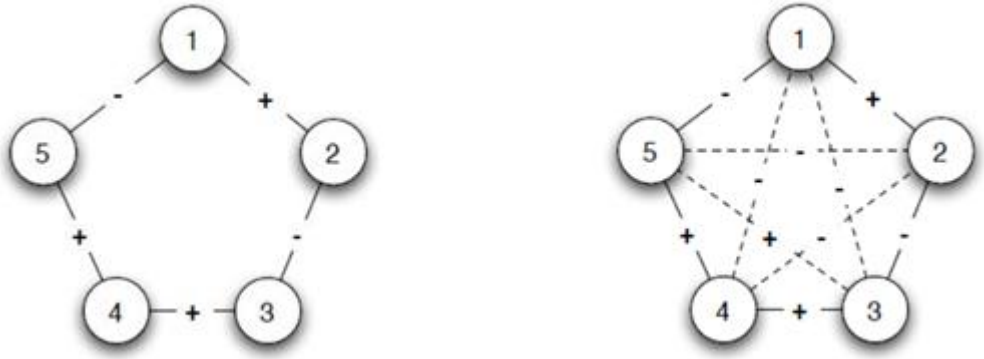
⇔ может быть разбит на сообщества (ДЗ доказать)



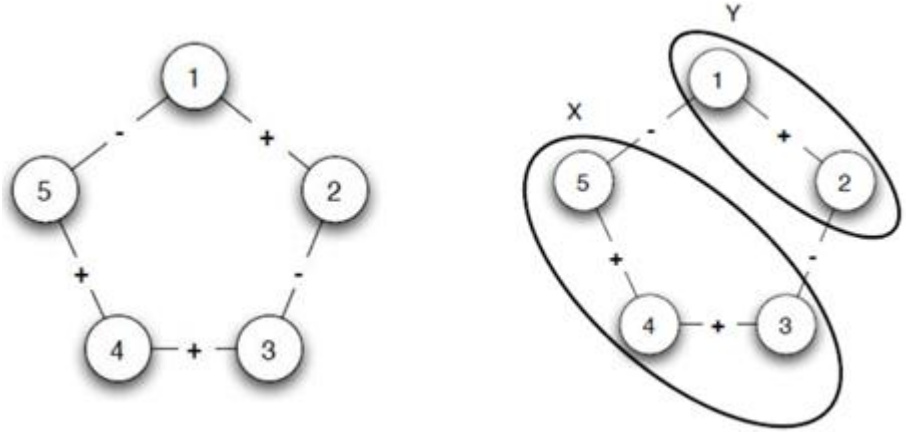
Сети с негативными связями

снимаем ограничение полноты

Граф сбалансирован, если его можно дополнить рёбрами до полного так, что получится полный сбалансированный граф



Результаты остаются: можно предложить метод, который находи разбиение или противоречие



⇔ нет циклов с чётным числом «–»

Сети с негативными связями

снимаем ограничение ВСЕ треугольники

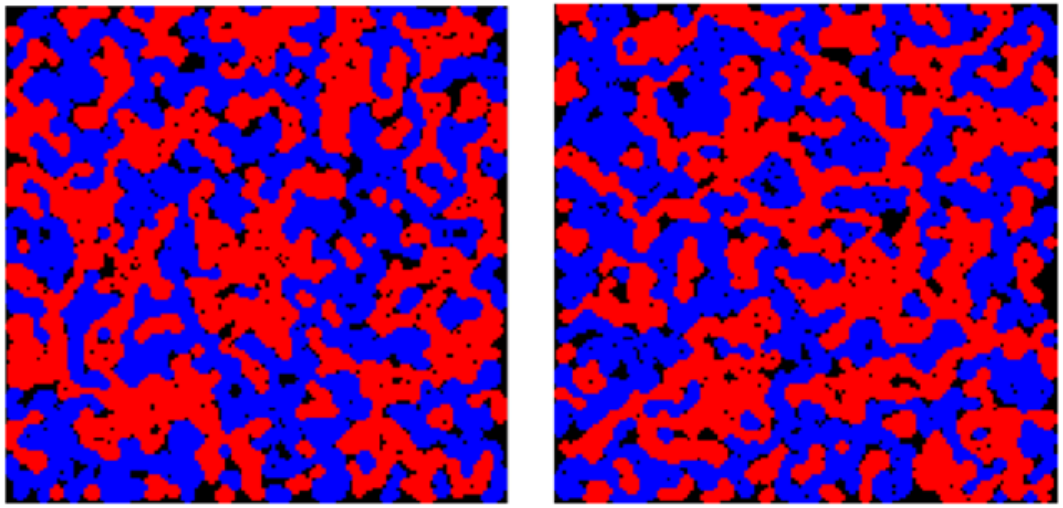
можно потребовать «почти все»

Модно предсказывать позитивность / негативность:

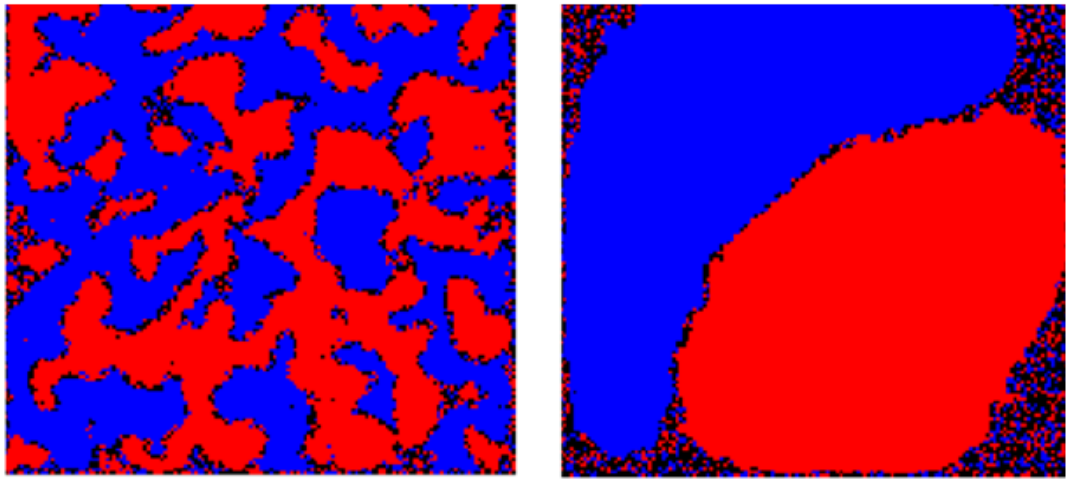
Leskovec et. al. «Predicting positive and negative links in online social networks» WWW 2010

«Игра в удовлетворённость»

локальные предпочтения приводят к глобальным патернам



2 запуска, 150×150, 10000 агентов, 8-соседство,
порог удовлетворения = 3, случайная инициализация



20 и 800 итераций, 150×150, 10000 агентов,
8-соседство, порог удовлетворения = 4, случайная инициализация

Что полезно: программирование

igraph – The network analysis package

http://igraph.org/

NetworkX: Python software for network analysis (v1.5)

http://networkx.lanl.gov

Gephi: Java interactive visualization platform and toolkit

http://gephi.org

Что полезно: курсы

Очень хороший

Hadi Amiri «Social Media Computing - CMSC 498J»

http://legacydirs.umiacs.umd.edu/~hadi/cmsc498j/syllabus.html

Очень хороший

Gonzalo Mateos «Network Science Analytics»

http://www2.ece.rochester.edu/~gmateosb/ECE442.html

Л.Жуков «Structural Analysis and Visualization of Networks» в ВШЭ

http://leonidzhukov.net/hse/2015/socialnetworks/

Что полезно: книги

David Easley, Jon Kleinberg «Networks, Crowds, and Markets: Reasoning About a Highly Connected World»
https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book.pdf

Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman «Mining of Massive Datasets»
http://infolab.stanford.edu/~ullman/mmds/book.pdf

Eric D. Kolaczyk «Statistical Analysis of Network Data: Methods and Models»

M. E. J. Newman «Networks: An Introduction» Oxford U. Press