Некоторые задачи анализа социальных сетей

Славнов Константин

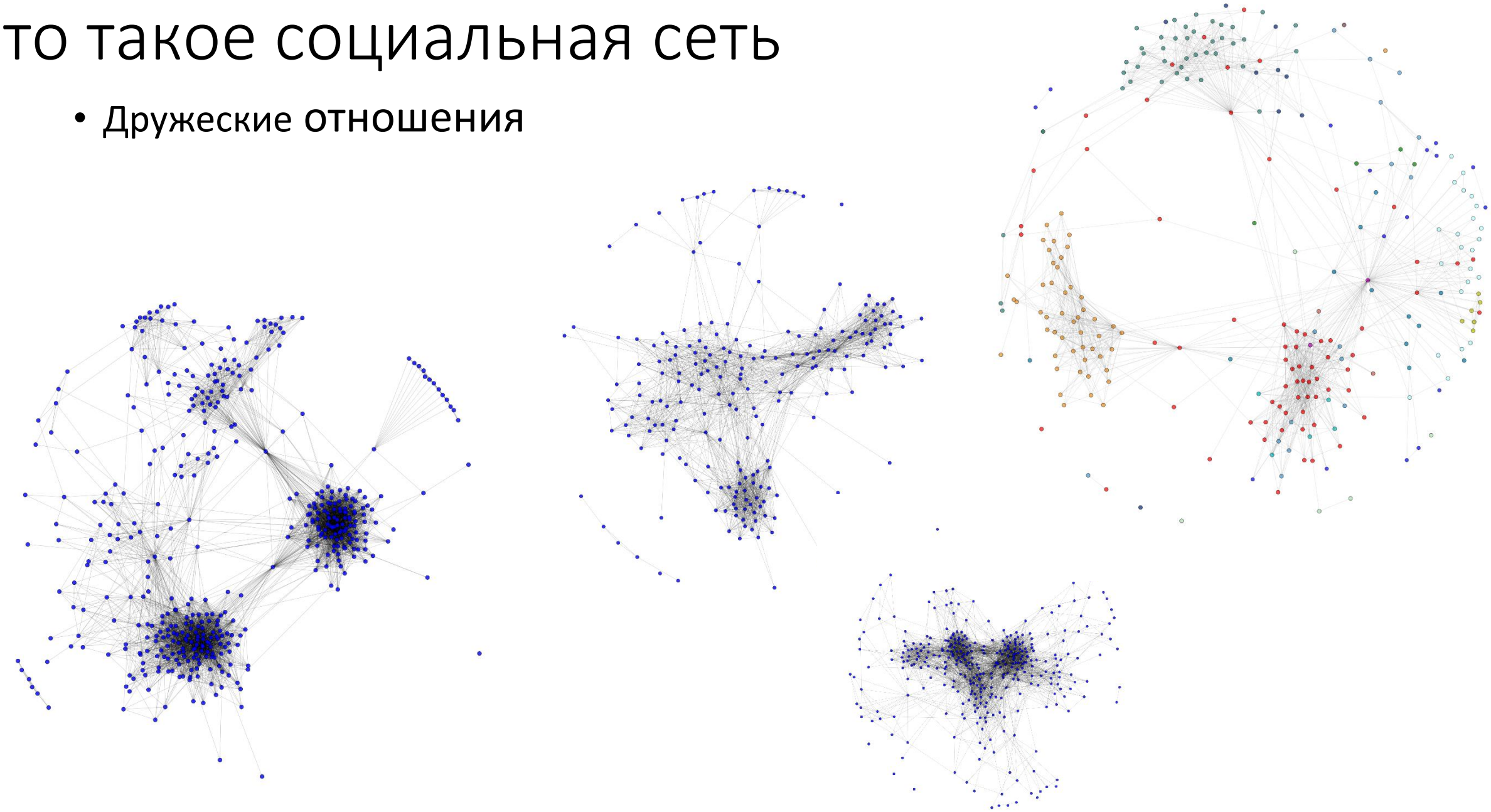
15.10.2014

План

- Что это такое. Примеры. Особенности
- Что можно посчитать
- Как можно моделировать
- Деанонимизация

Что такое социальная сеть

- Дружеские отношения



Что такое социальная сеть

- Сеть протеиновых взаимодействий

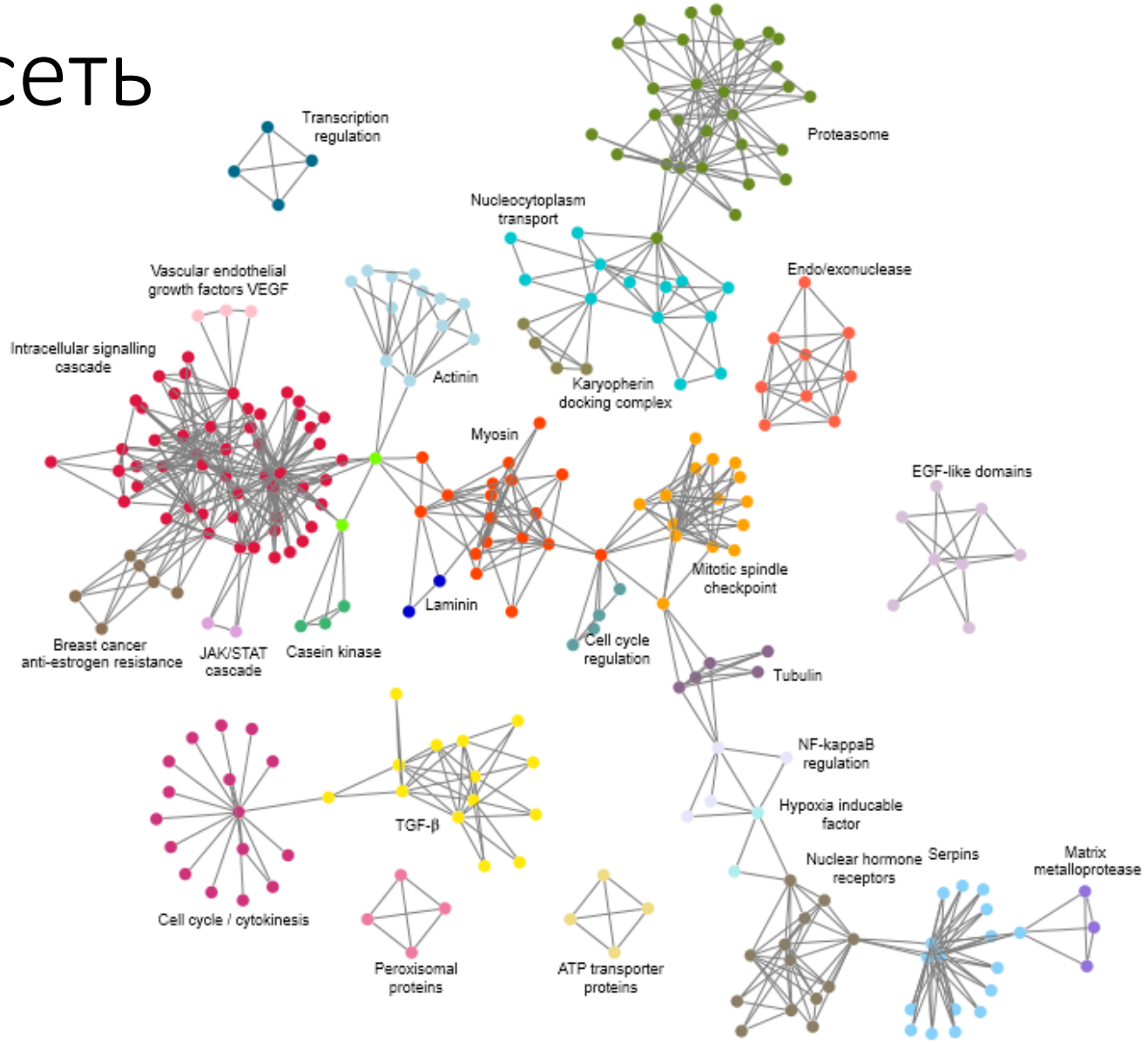


FIG. 3 Community structure in protein-protein interaction networks. The graph pictures the interactions between proteins in cancerous cells of a rat. Communities, labeled by colors, were detected with the Clique Percolation Method by Palla et al. (Section XI.A). Reprinted figure with permission from Ref. (Jonsson *et al.*, 2006). ©2006 by PubMed Central.

Что такое социальная сеть

- Граф цитирования

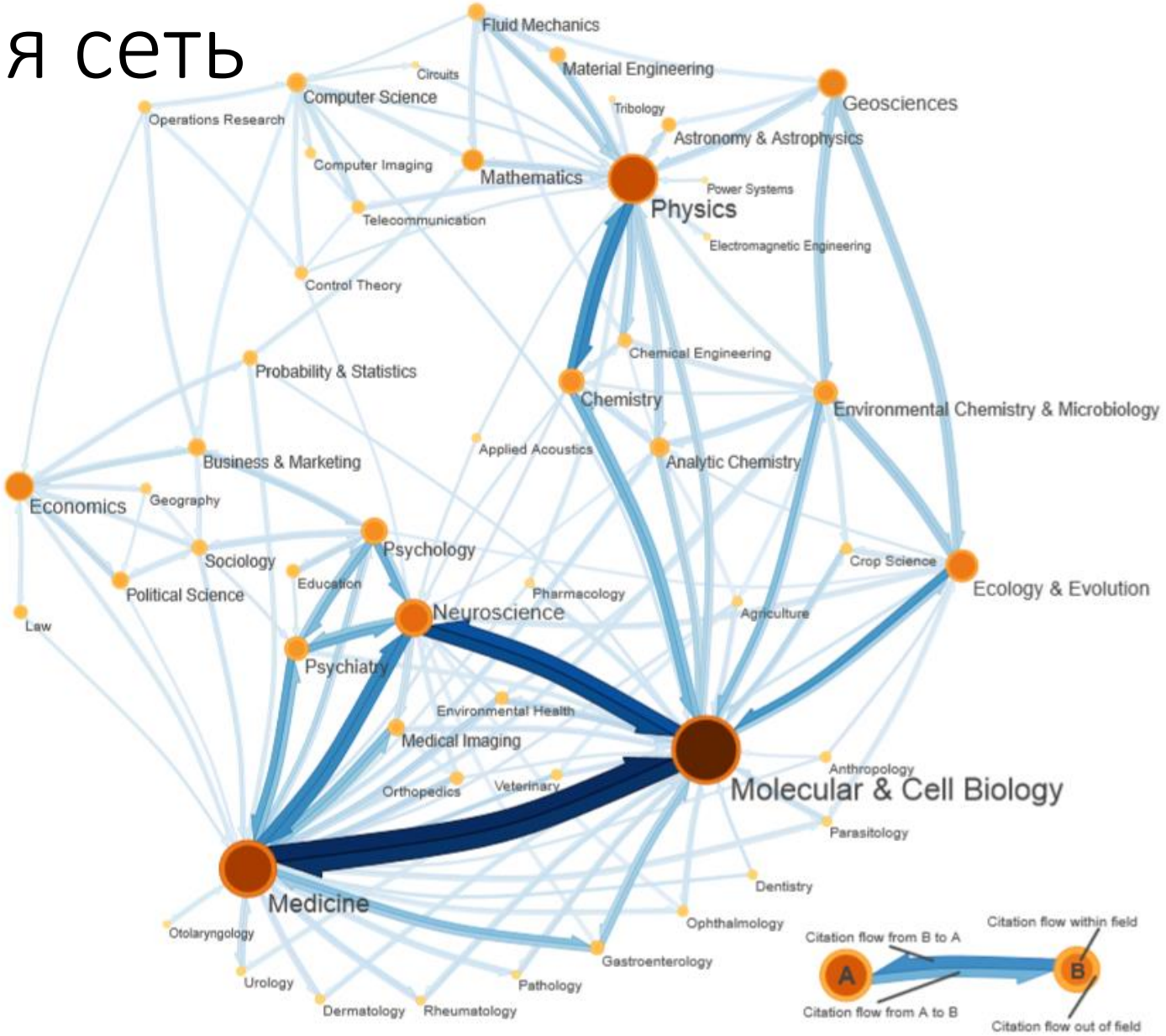


FIG. 40 Map of science derived from a clustering analysis of a citation network comprising more than 6000 journals. Reprinted figure with permission from Ref. (Rosvall and Bergstrom, 2008). ©2008 by the National Academy of Science of the USA.

Что такое социальная сеть

- История почтовой переписки внутри корпорации

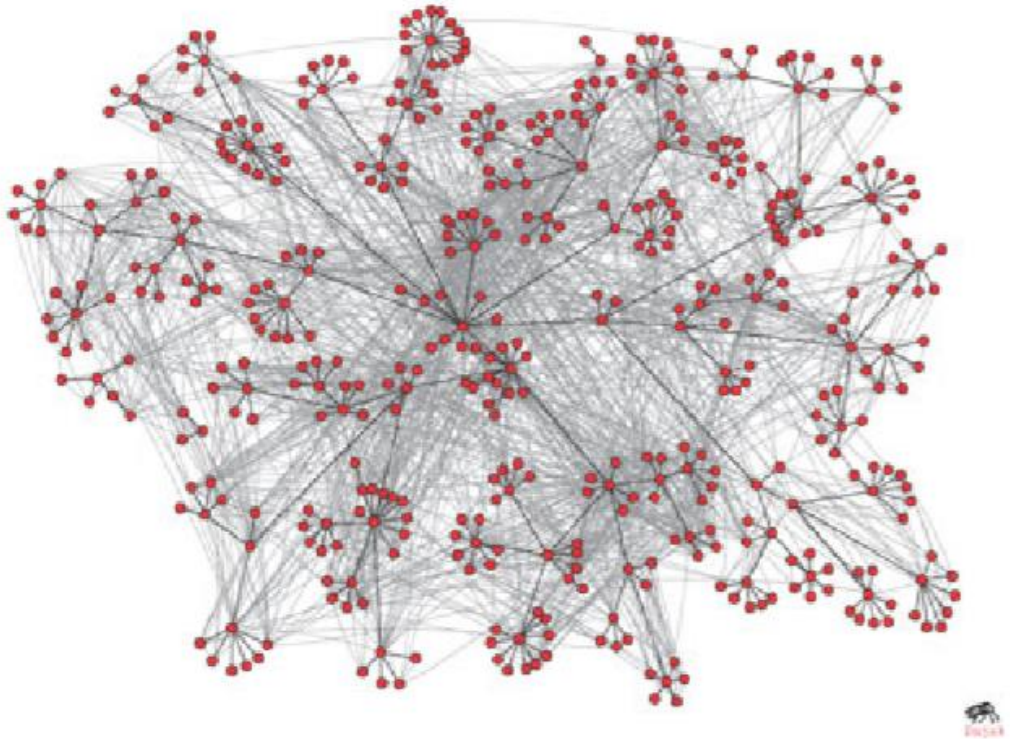
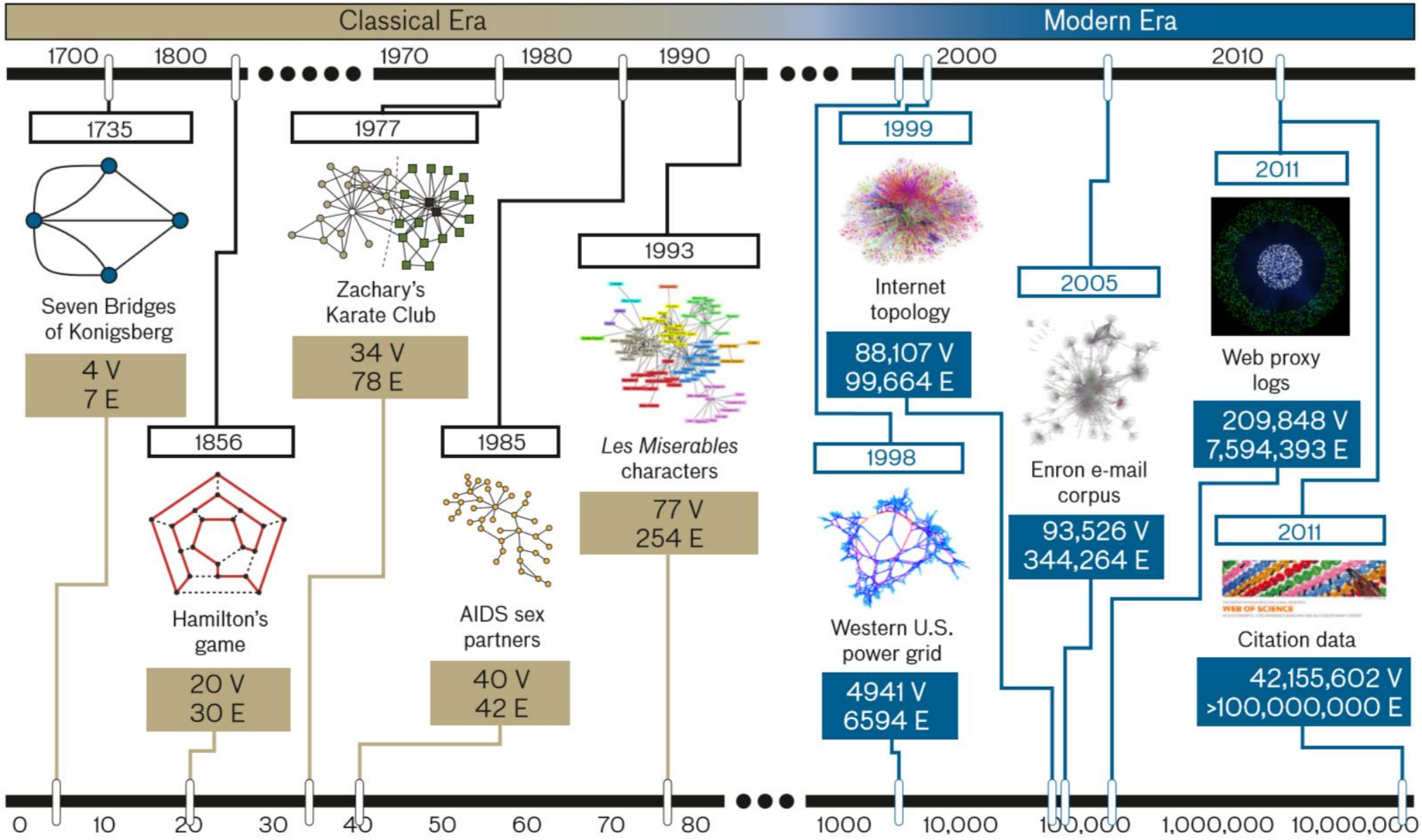


Figure 1.2: Social networks based on communication and interaction can also be constructed from the traces left by on-line data. In this case, the pattern of e-mail communication among 436 employees of Hewlett Packard Research Lab is superimposed on the official organizational hierarchy [6]. (Image from http://www-personal.umich.edu/ ladamic/img/hplabsemailhierarchy.jpg)

Размеры

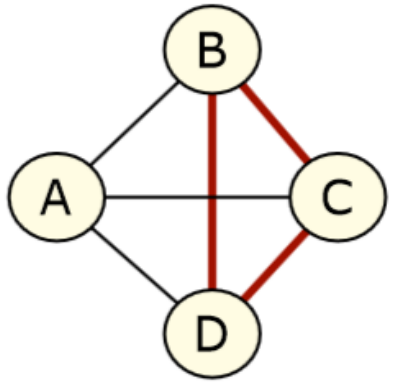

Classical Era
Modern Era
1700
1800
1970
1980
1990
2000
2010
1735
Seven Bridges of Konigsberg
4 V
7 E
1856
Hamilton's game
20 V
30 E
1977
Zachary's Karate Club
34 V
78 E
1985
AIDS sex partners
40 V
42 E
1993
Les Miserables characters
77 V
254 E
1999
Internet topology
88,107 V
99,664 E
1998
Western U.S. power grid
4941 V
6594 E
2005
Enron e-mail corpus
93,526 V
344,264 E
2011
Web proxy logs
209,848 V
7,594,393 E
2011
Citation data
42,155,602 V
>100,000,000 E
0
10
20
30
40
50
60
70
80
1000
10,000
100,000
1,000,000
10,000,000

Числовые характеристики

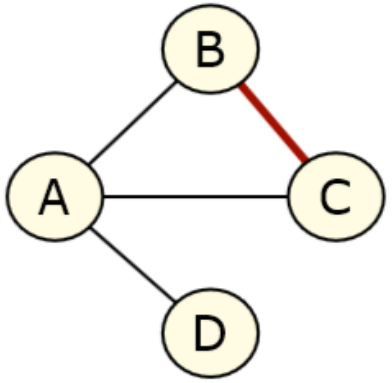
- Среднее расстояние между двумя вершинами
- Диаметр графа
- Степень вершин и ее распределение
- Центральность узла
- Коэффициент кластеризации
- Коэффициент ассортативности
- Доля треугольников
- …

Коэффициент кластеризации

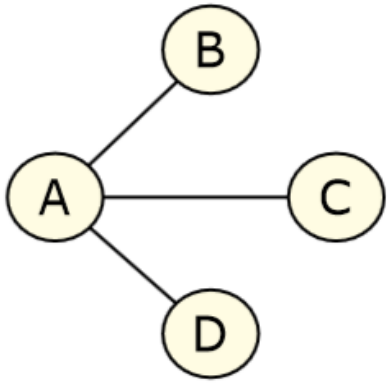
- Доля ребер, соединяющих соседние вершины друг с другом
- Для вершины A:



$$CC(A) = \frac{3}{3}$$



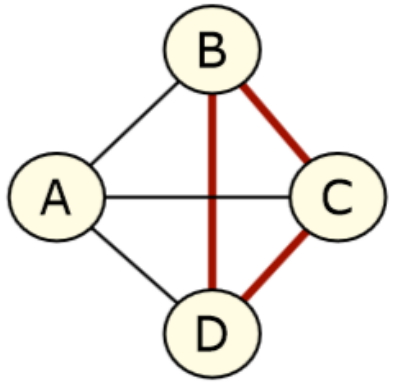
$$CC(A) = \frac{1}{3}$$

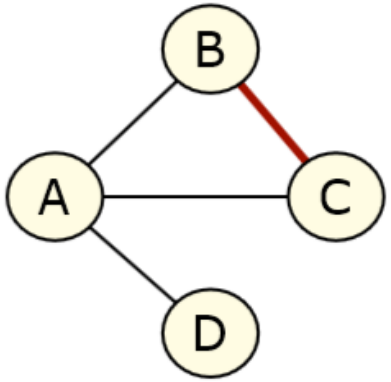


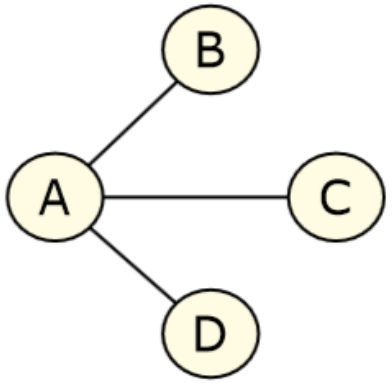
$$CC(A) = \frac{0}{3}$$

Коэффициент кластеризации

- Можно посчитать на всем графе
- На эго-подграфе
- Усреднить по всем эго-подграфам







$$CC(A) = \frac{3}{3}$$

$$CC(A) = \frac{1}{3}$$

$$CC(A) = \frac{0}{3}$$

Ассортативность

- Образование связей между вершинами в чем-то схожими друг с другом
- Вершина с большой степенью скорее будет связана с такой же вершиной

- Коэффициент ассортативности – коэф. Корреляции Пирсона между степенями соседних вершин.

Ассортативность: эмпирика

Таблица 2. Ассортативность социальных, технологических и биологических сетей. Reprinted with permission from Newman M. E. J. Mixing patterns in networks // Phys. Rev. E 67, 026126. © 2003 by the American Physical Society

	Сеть	Тип	Размер n	Ассортативность r
Социальные	соавторов по физике	неориентированная	52 909	0.363
	соавторов по биологии	неориентированная	1 520 251	0.127
	соавторов по математике	неориентированная	253 339	0.120
	сотрудничества актеров кино	неориентированная	449 913	0.208
	директоров компаний	неориентированная	7 673	0.276
	связей студентов	неориентированная	573	–0.029
	адресов электронной почты	ориентированная	16 881	0.092
Технологические	сеть электростанций	неориентированная	4 941	–0.003
	Интернет	неориентированная	10 697	–0.189
	«Всемирная паутина» (WWW)	ориентированная	269 504	–0.067
	взаимозависимости программного обеспечения	ориентированная	3 162	–0.016
Биологические	взаимодействий белков	неориентированная	2 115	–0.156
	метаболическая сеть	неориентированная	765	–0.240
	нейронная сеть	ориентированная	307	–0.226
	морская пищевая сеть	ориентированная	134	–0.263
	пресноводная пищевая сеть	ориентированная	92	–0.326

Теория 6 рукопожатий

180 миллионов людей
1.3 миллиардов ребер

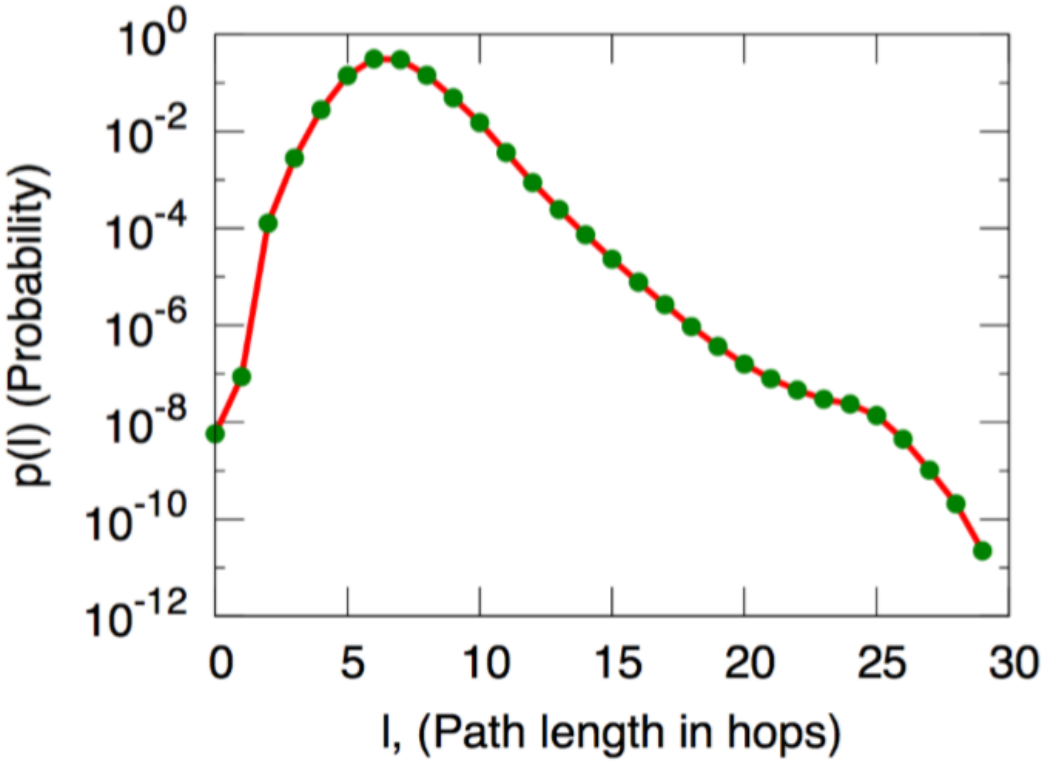


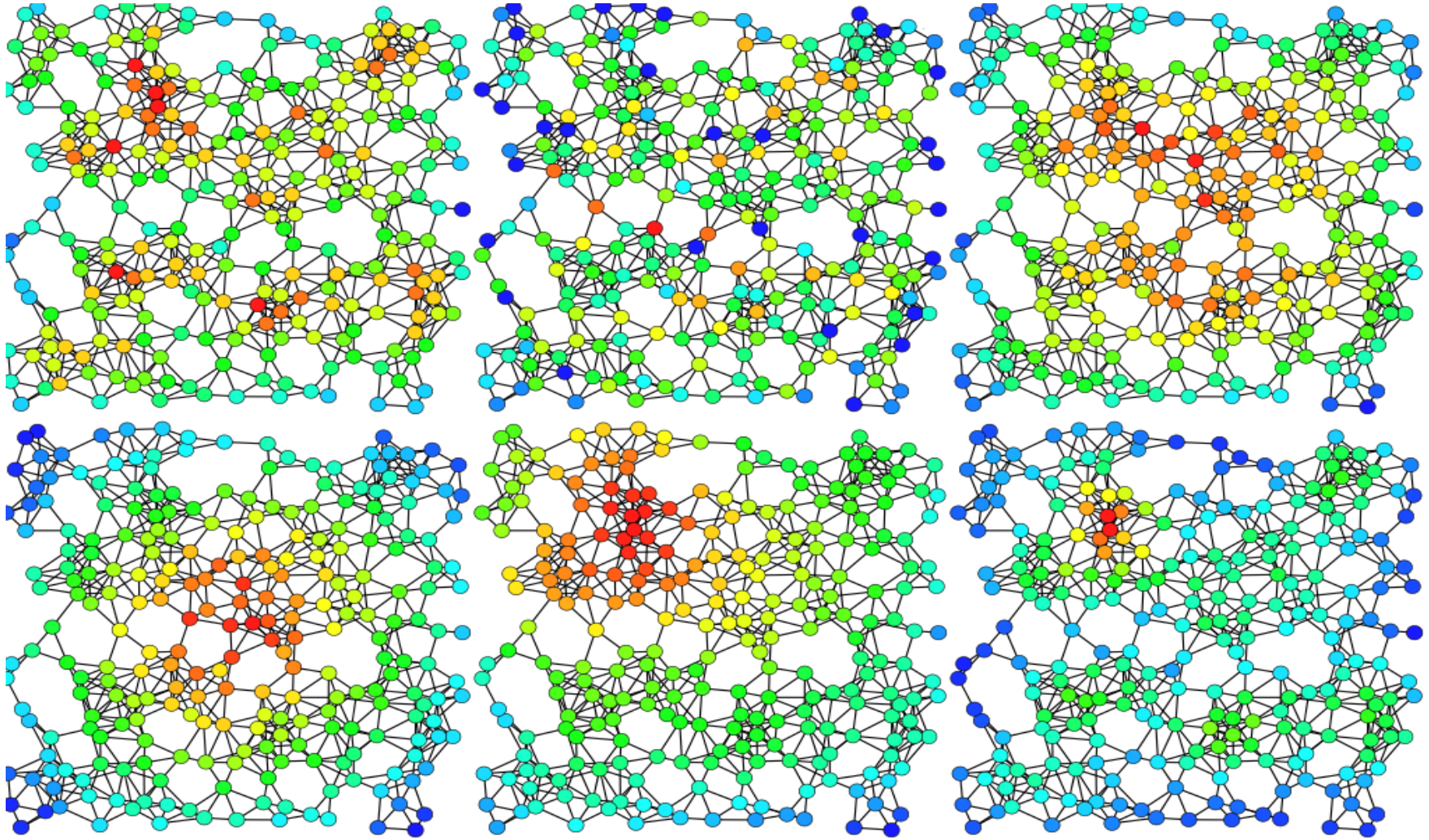
Figure 2.11: The distribution of distances in the graph of all active Microsoft Instant Messenger user accounts, with an edge joining two users if they communicated at least once during a month-long observation period [273].

Центральность

Degree

Betweenness

Katz



Closeness

Eigenvector

Alpha

Виды центральностей вершины

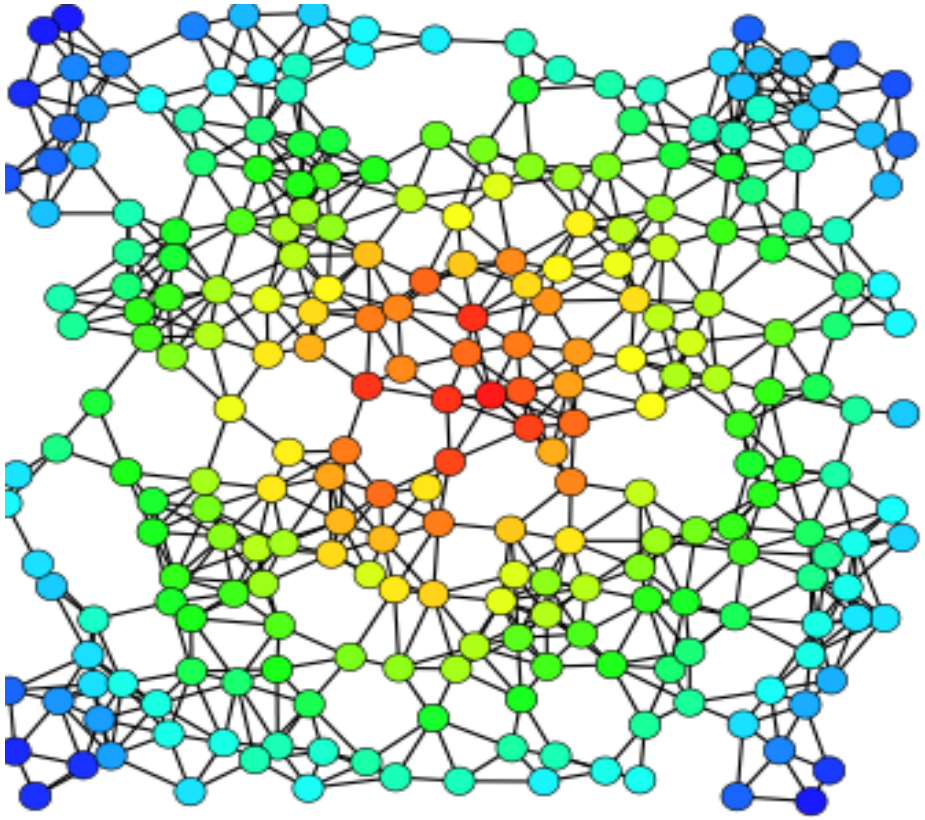
- Closeness

Близость – величина обратная среднему расстоянию от данной вершины до остальных

$$CL(v) = \frac{1}{\sum_{\substack{w \in V \\ w \neq v}} dist(v, w)}$$

Или

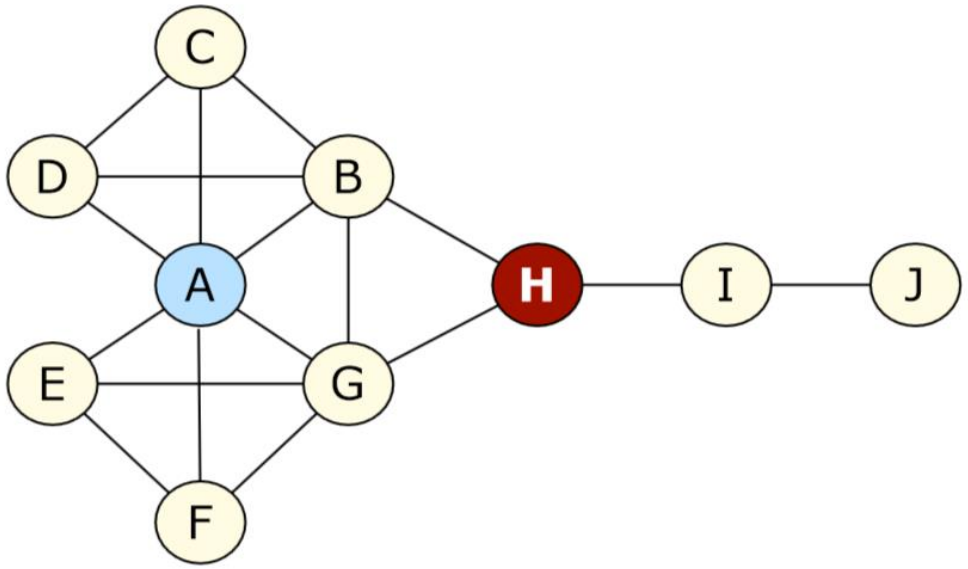
$$CL(v) = \sum_{\substack{w \in V \\ w \neq v}} 2^{-dist(v,w)}$$



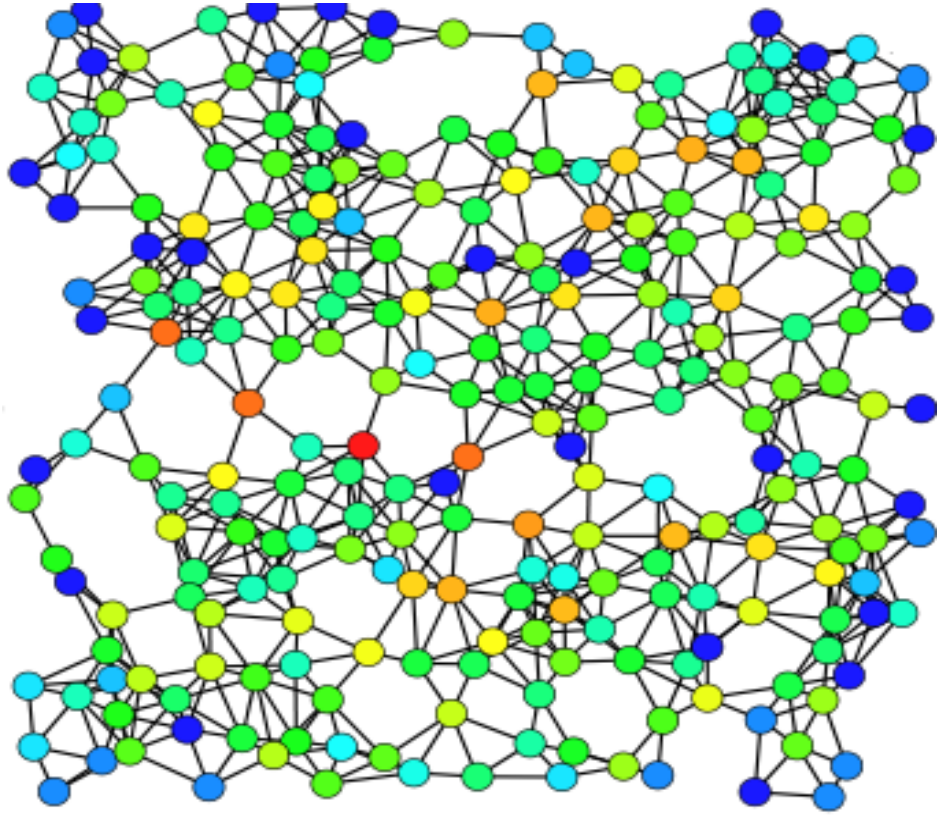
Виды центральностей вершины

- Betweenness

Доля кратчайших путей проходящих через данную вершину



H – 0.388
B, G – 0.236
I – 0.222
A – 0.166
C, E, D, F, J – 0.0



Виды центральностей вершины

- Eigenvector

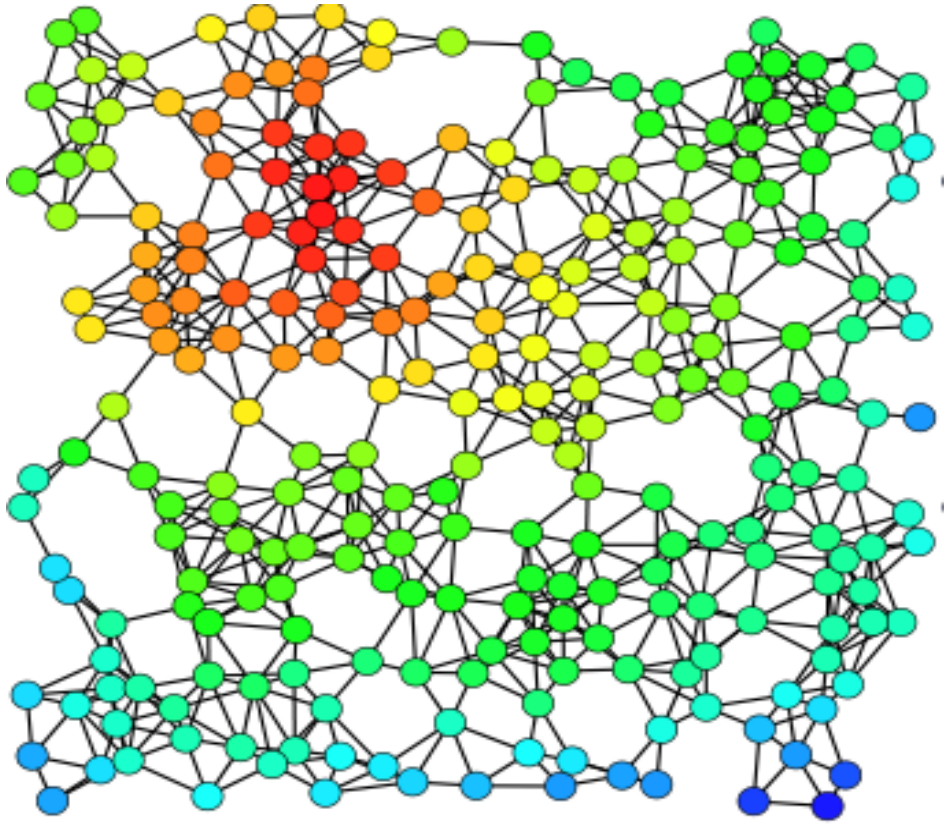
Пропорциональна сумме центральностей соседних вершин

Узел важный если соединен с другими важными узлами

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

$$Ax = \lambda x \Rightarrow x_{max}$$

A – матрица смежности графа G



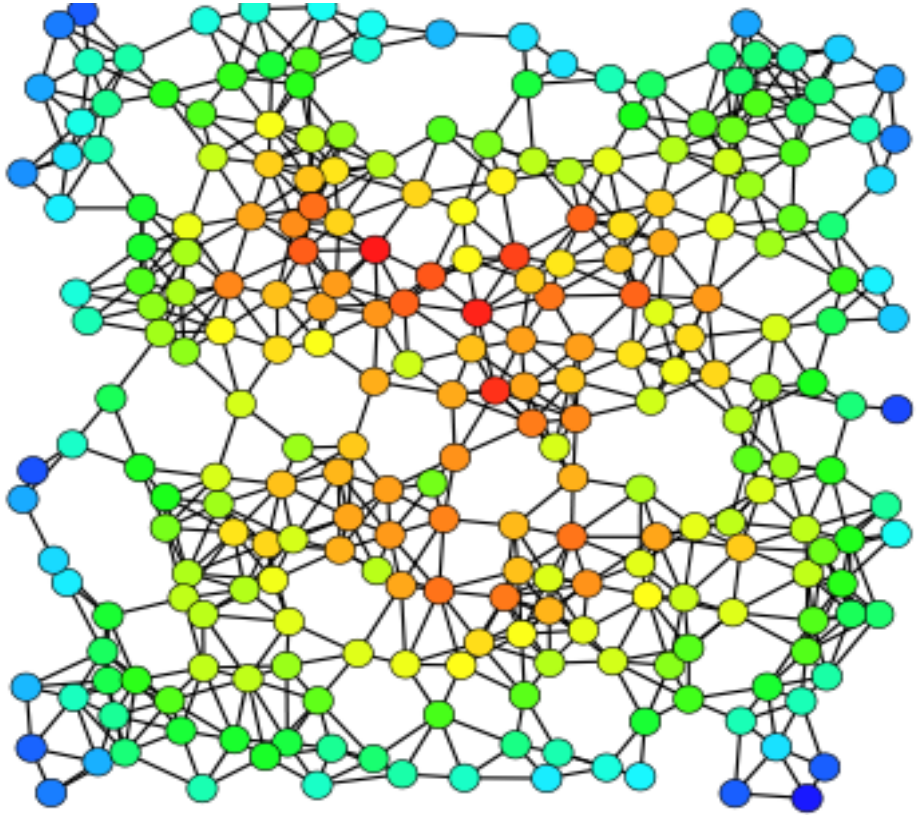
Виды центральностей вершины

- Katz

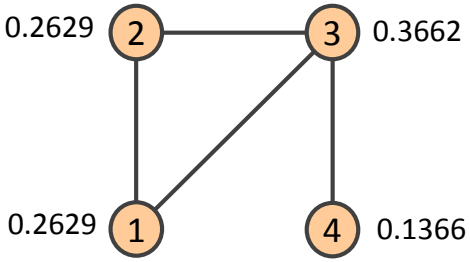
Степень влияния пользователя в социальной сети

$$C_{Katz}(i) = \sum_{k,j=1}^{\infty,n} \alpha^k \left(A^k\right)_{ji}, \qquad \alpha < \frac{1}{\lambda_{max}}$$

$$C_{Katz}(i) = \left((I - \alpha A^T)^{-1} - I\right)e$$



Число матрицы A^k отвечает за количество путей длинны k между вершинами



$\boldsymbol{A}$

0	1	1	0
1	0	1	0
1	1	0	1
0	0	1	0

$\boldsymbol{A^2}$

2	1	1	1
1	2	1	1
1	1	3	0
1	1	0	1

$\boldsymbol{A^3}$

2	3	4	1
3	2	4	1
4	4	2	3
1	1	3	0

$\boldsymbol{A^4}$

7	6	6	4
6	7	6	4
6	6	11	2
4	4	2	3

Виды центральностей вершины

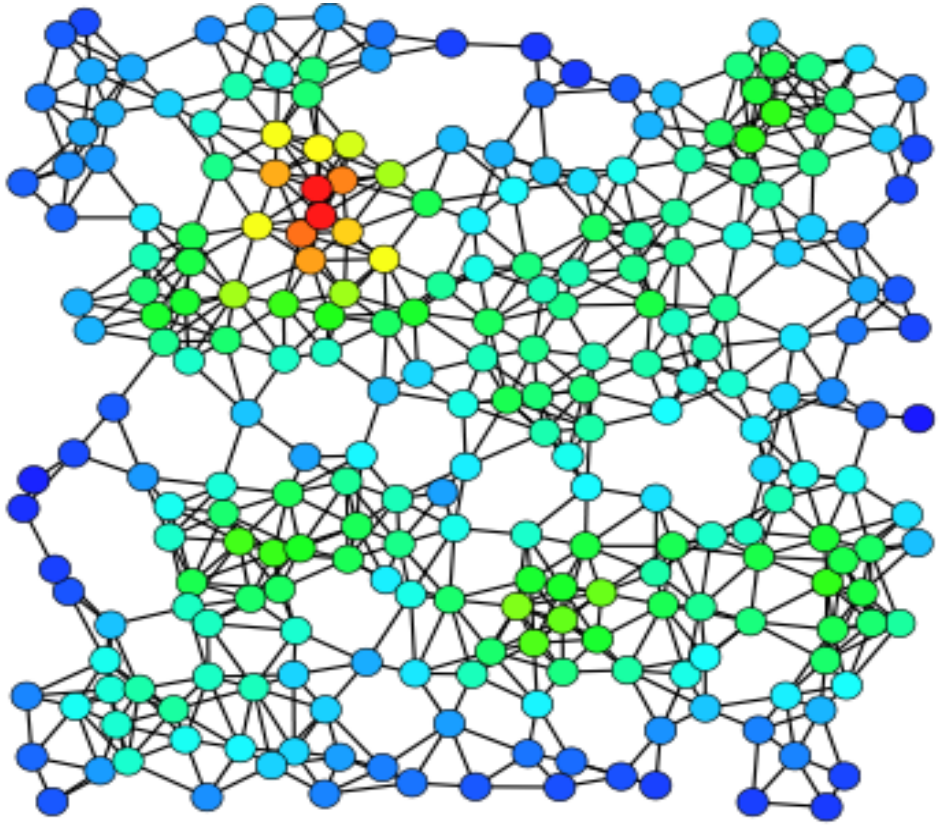
- Alpha

Учитывает внешнюю информацию

$$x_v = \frac{1}{\lambda}\sum_{t \in G} a_{v,t} x_t \ + e_i$$

$$C_{Alpha}(i) = (I - \alpha A^T)^{-1} e$$

e – внешняя информация о важности узлов

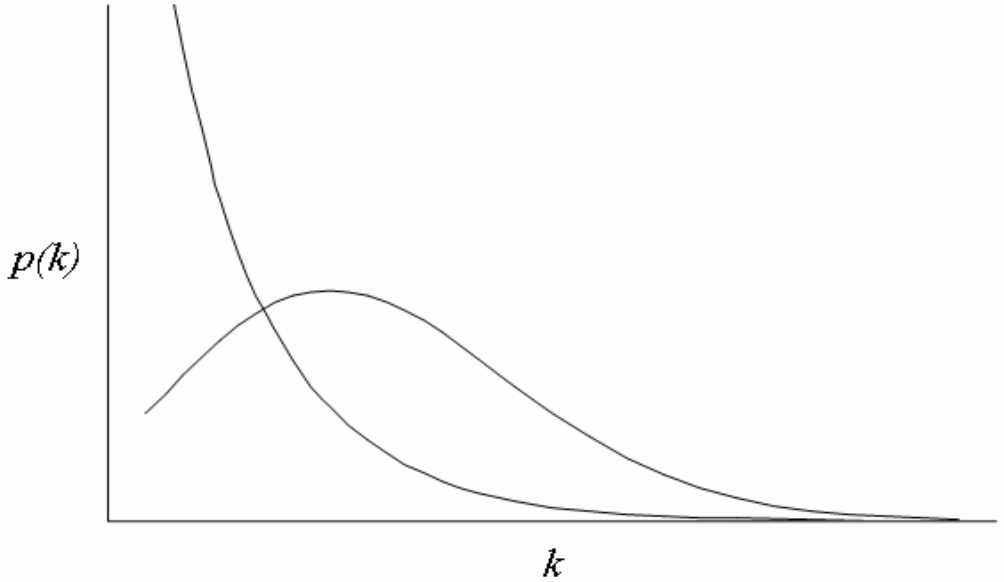


Распределение на степенях вершин

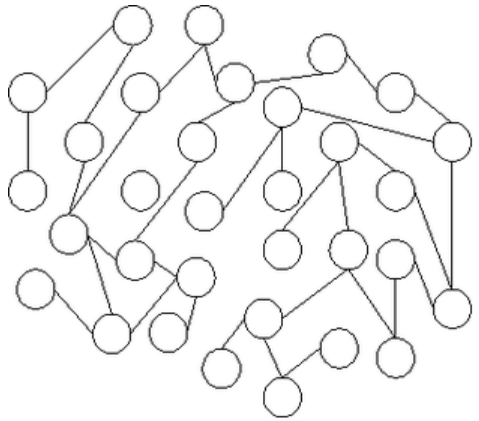
Безмасштабные сети(Scale-free network)

$$P(k) \sim k^{-\gamma}$$

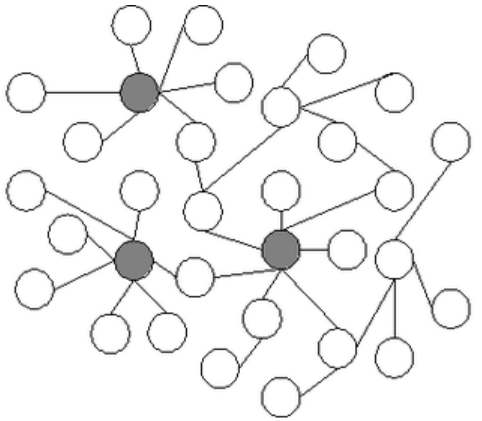
$$2 < \gamma < 3$$



Распределение степеней вершин
у случайных и безмасштабных графов



Случайный граф



Безмасштабный граф

Распределение на степенях вершин

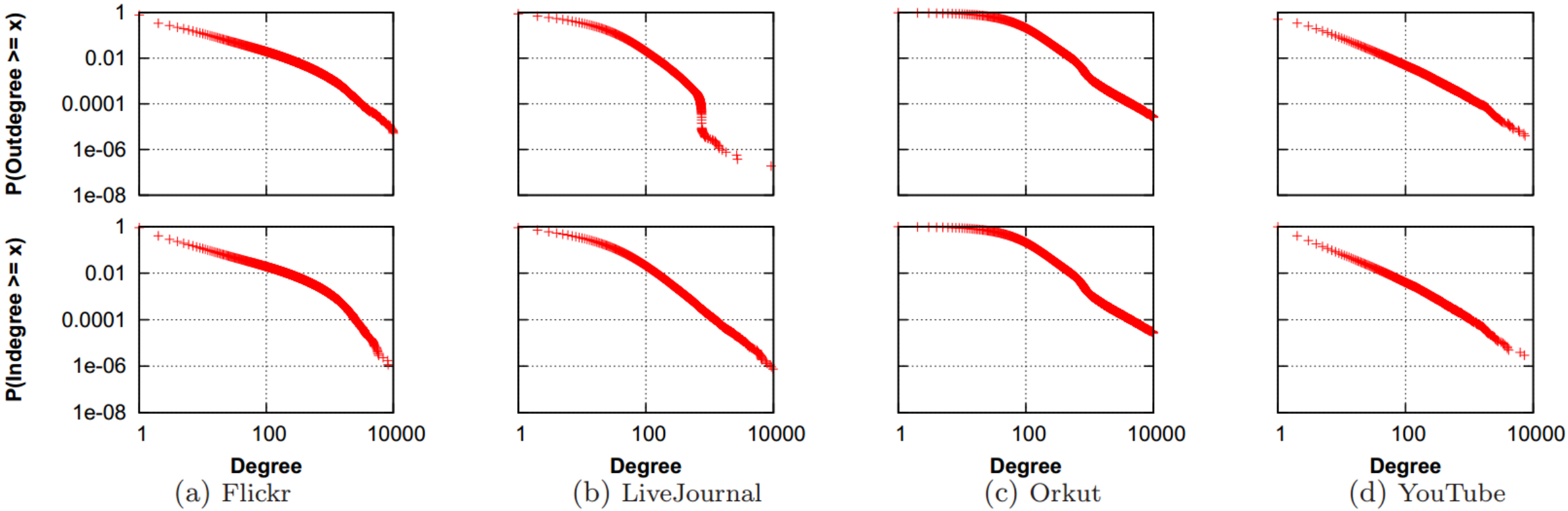


Figure 2: Log-log plot of outdegree (top) and indegree (bottom) complementary cumulative distribution functions (CCDF). All social networks show properties consistent with power-law networks.

Источник: Measurement and Analysis of Online Social Networks; Alan Mislove

Генерация графов с заданными свойствами

Задача:

Научиться генерировать графы, по свойствам похожие на социальные.

Распределение степеней вершин

Диаметр

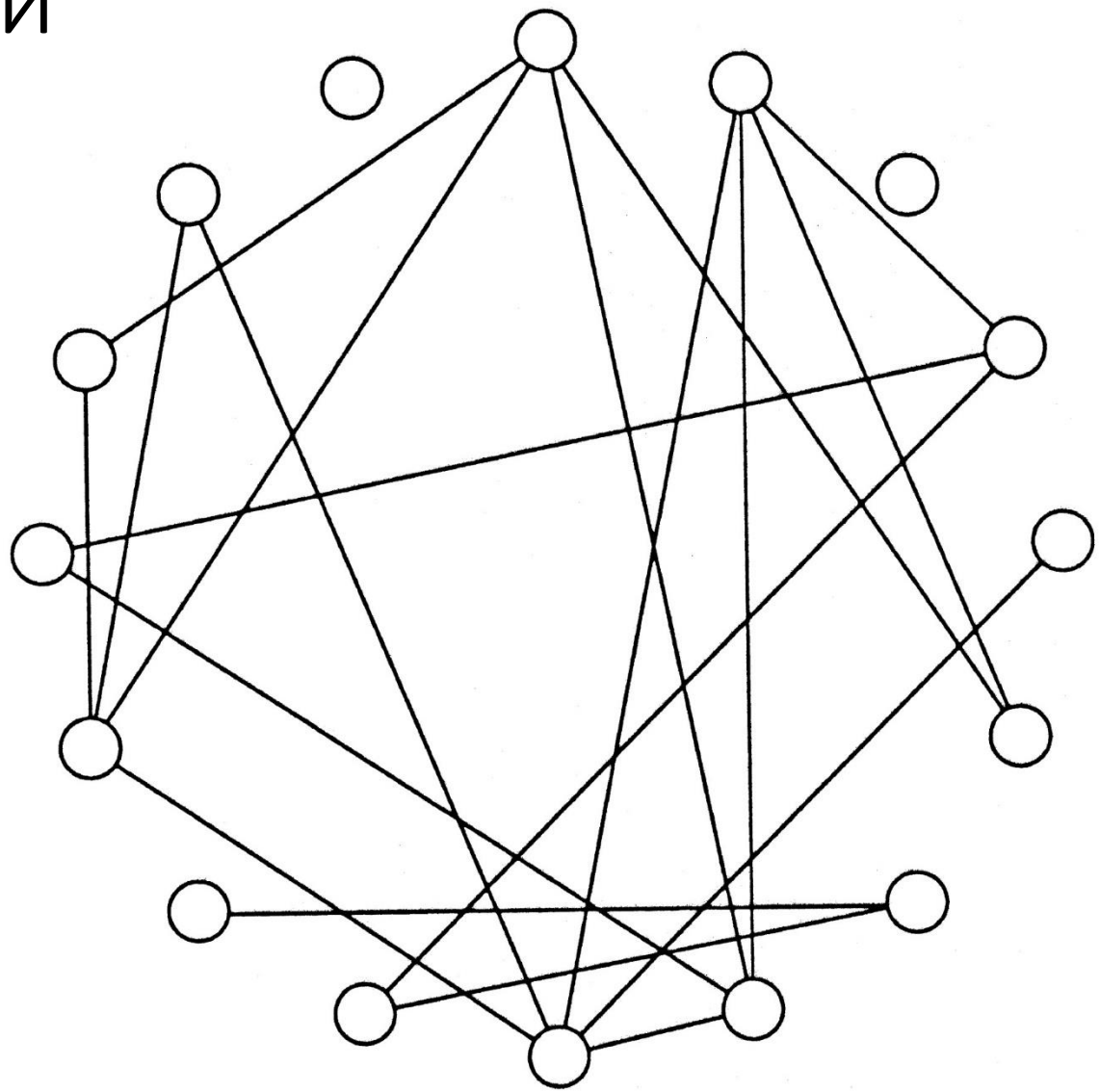
Среднее расстояние

Сообщества

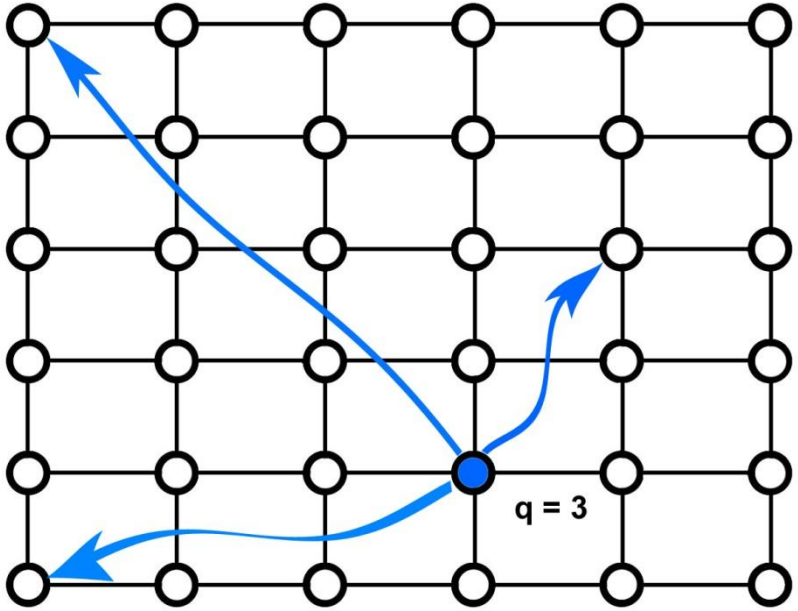
Модель Эрдеша-Реньи

- Самая простая
- N вершин
- Ребра создаются независимо с вероятностью p

- N вершин, M ребер, равномерное распределение
- Выбирается случайный граф



Модель Уоттса-Строгаца

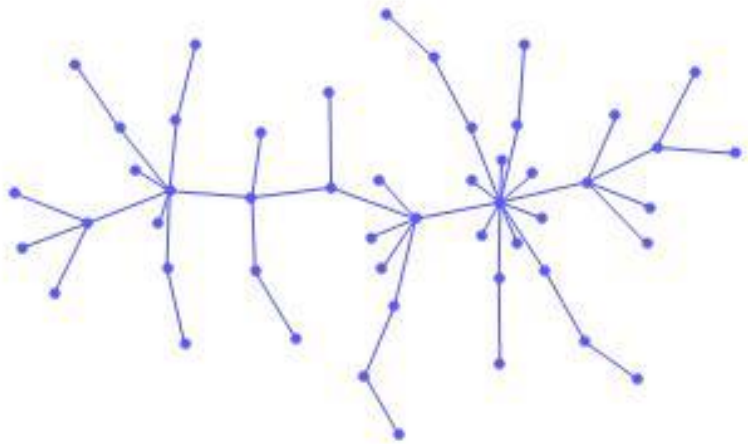


$$\frac{[d(u,v)]^{-r}}{\sum_{v:v\neq u}[d(u,v)]^{-r}}$$

- Все живут на сетке *n* x *n*
- “Решёточное расстояние” – число шагов по решетке между двумя точками
- *p* - диапазон локальных связей
- *q – количество дальних связей*

Модель Барабаши — Альберта

- Безмасштабная сеть
- Принципы
  - Рост сети
  - Принцип предпочтительного присоединения (ПП)



Модель Барабаши — Альберта

Алгоритм:

- Начальный граф с m_0узлами. $m_0 \geq 1$ и $\deg(v) \geq 1$

- Добавляется новый узел, соединяется с существующими узлами с вероятностью

$$p_i = \frac{k_i}{\sum_j k_j}, \text{где } k_i - \text{степень узла } i$$

Модель Барабаши — Альберта

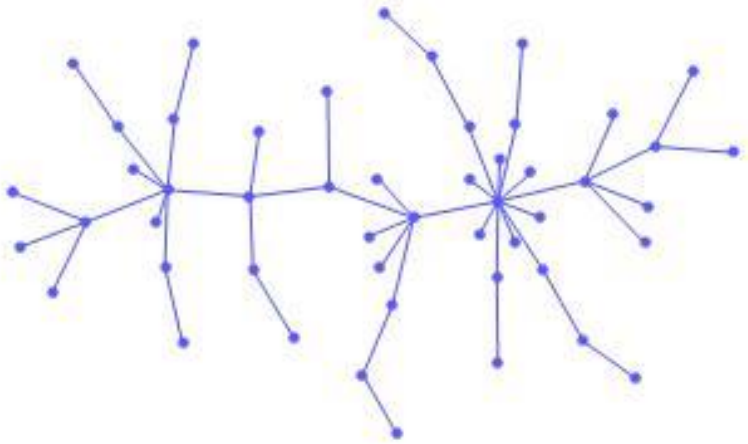
Алгоритм:

- Начальный граф с m_0 узлами. $m_0 \geq 1$ и $\deg(v) \geq 1$

- Добавляется новый узел, соединяется с существующими узлами с вероятностью

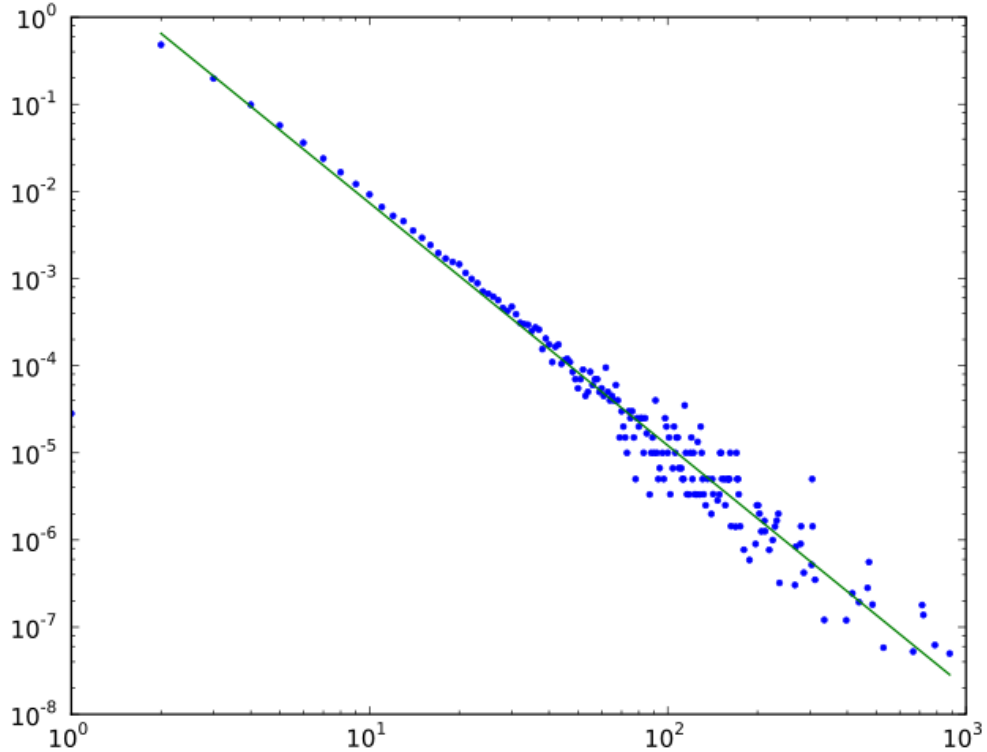
$$p_i = \frac{k_i}{\sum_j k_j}, \text{где } k_i - \text{степень узла } i$$

Свойства



Степенное распределение

$$P(k) \sim k^{-3}$$



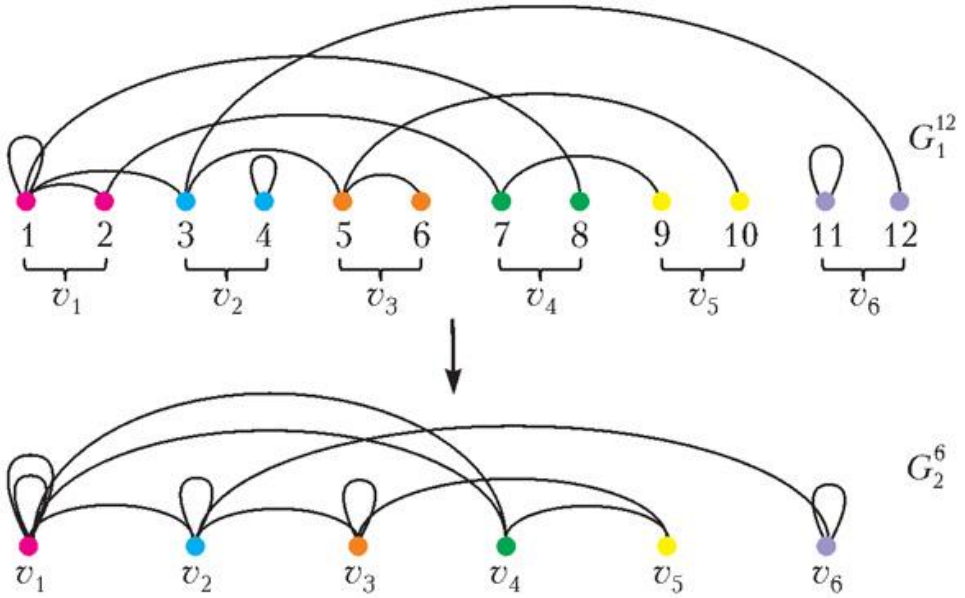
Распределение степеней модели БА, которое подчиняется степенному закону

Длинна пути $l \sim \frac{\ln N}{\ln \ln N}$

Коэффициент кластеризации $C \sim N^{-0.75}$

Модель Боллобаша – Риордана

- $G_1^1 = (\{1\}, \{1,1\})$
- $G_1^{n-1} \rightarrow G_1^n$: добавим вершину n к графу G_1^{n-1} и ребро v: $p(\{n,n\}) = \frac{1}{2n-1}$, $p(\{n,i\}) = \frac{\deg(i)}{2n-1}$ (Принцип предпочтительного присоединения)
- G_n^k: $v_i = \{ki+1, \dots, ki+k\}, \quad i = 0, \dots, k-1$

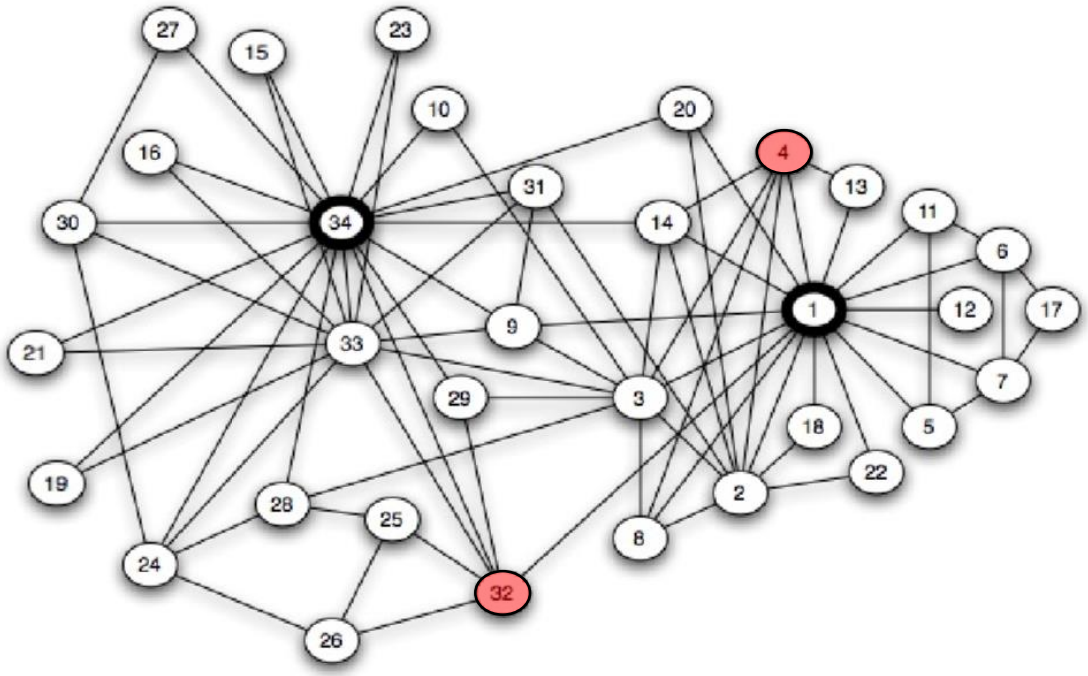


Двухуровневые модели

- Принципиальная схема генерации такая же
- Вероятность присоединения к старой вершине:

$$P(k_i) \sim k_i + C \sum_{(i,j)\in E} k_j \,, \qquad C \in [0,1]$$

Деанонимизация социальных сетей



Вершина 4 может идентифицировать себя: степень 6, соединен с обоими лидерами

Вершина 32 может идентифицировать себя: степень 5, соединен со вторым лидером

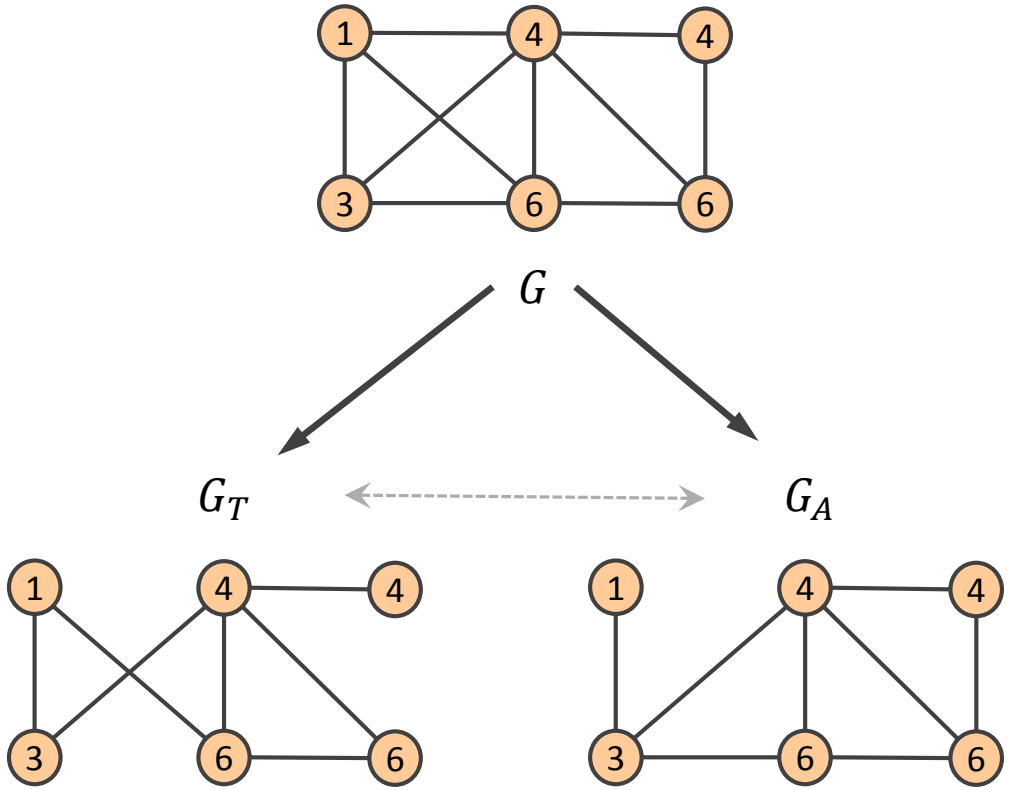
Wayne Zachary’s Ph.D. work (1970-72): observe social ties and rivalries in a university karate club.

Общая постановка задачи

G – исходный граф

G_T, G_A - получены из G стохастическими процессами

Цель деанонимизации – найти соответствие между вершинами G_T, G_A



Примеры:

- Объединение информации из разных социальных сетей
- Деанонимизация данных

IJCNN Social Network Challenge

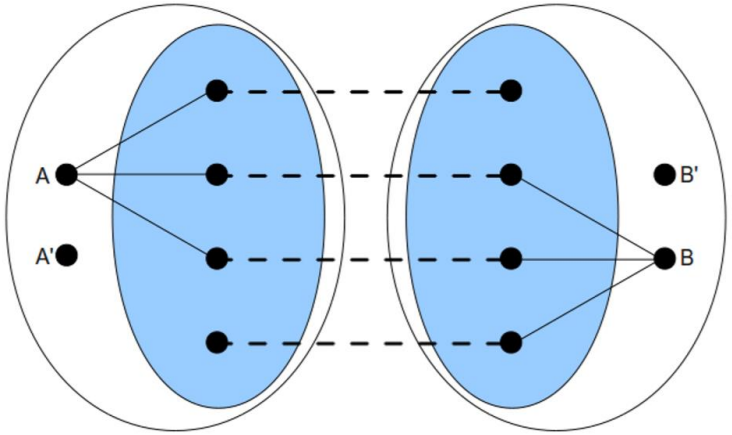
- LLP – задача прогнозирования ребер в графе
- Победители деанонимизировали данный граф и «подсмотрели» ответы
- Соответствие графов было не точным

#	Δ1w	Team Name * in the money	Score
1	—	IND CCA *	0.98115
2	↑98	wcuk	0.96954
3	↓1	vsh	0.95272
4	↓1	Jeremy Howard	0.94506
5	↑2	grec	0.92712

Принцип решения

- Поиск seed-вершин
  - Как-то идентифицировать небольшое количество вершин
  - Вручную, вершины высокой степени

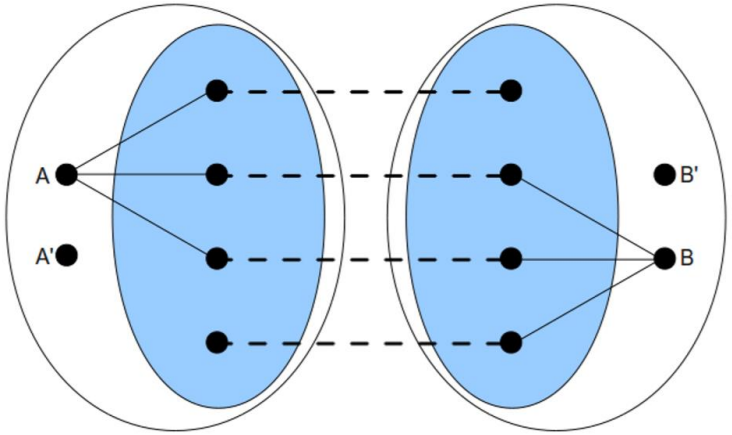
- Распространение
  - Итеративно используя уже идентифицированные вершины, распространить деанонимизацию на остальные
  - Косинусная мера схожести между соседними, уже размеченными вершинами.
  - «Достаточно» похожие связываются



Сходство между A и B $-$ $\frac{2}{\sqrt{3}\sqrt{3}}$

Принцип решения

- Распространение
  - Сначала на вершинах с высокой степенью, потом акцент на вершинах из тестовой выборки
  - Достаточная похожесть:
    - Хотя бы 4(3) пары соседних вершин связанных друг с другом
    - Сходство больше чем 0.5
    - Сходство между лучшим и следующим за ним отличается хотя бы на 0,2(нет)
  - Множественное соответствие до 3 кандидатов.

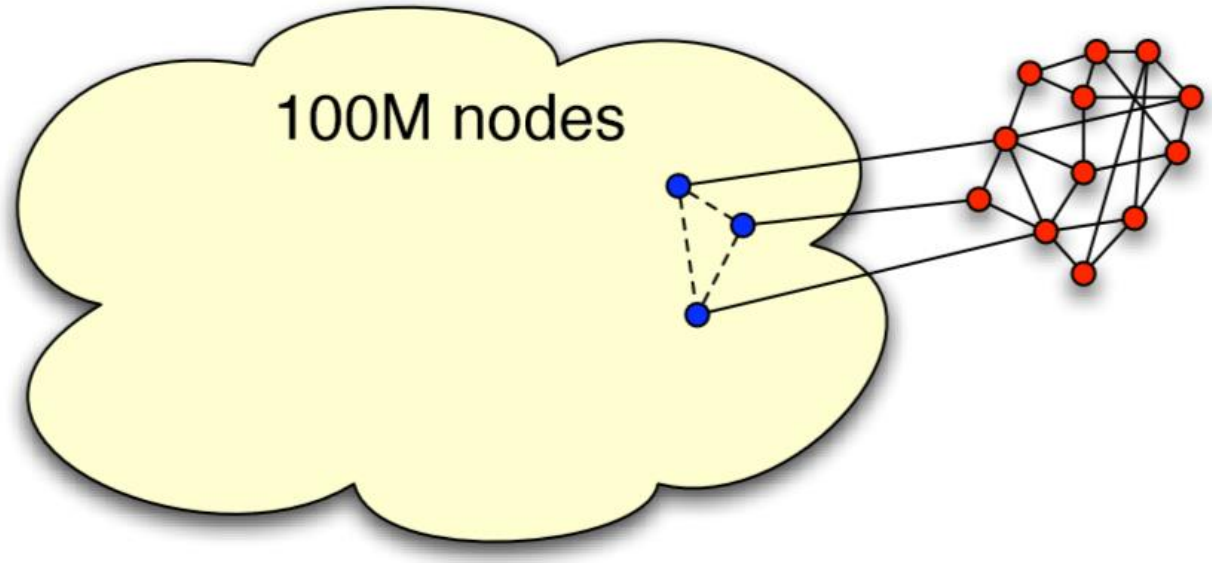


Сходство между A и B $-$ $\frac{2}{\sqrt{3}\sqrt{3}}$

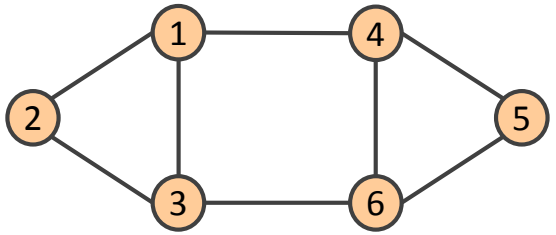
Деанонимизация. Активная атака

Принцип:

До извлечения данных создать случайный граф размера $\sim 2 \log n$, и встроить в сеть.



Матрица Лапласа



Собственное значение	0	1	3	3	4	5
Собственный вектор	1	1	−5	−1	−1	−1
	1	2	4	−2	1	0
	1	1	1	3	−1	1
	1	−1	−5	−1	1	1
	1	−2	4	−2	−1	0
	1	−1	1	3	1	−1

Матрица смежности

0	**1**	**1**	**1**	0	0
1	0	**1**	0	0	0
1	**1**	0	0	0	**1**
1	0	0	0	**1**	**1**
0	0	0	**1**	0	**1**
0	0	**1**	**1**	**1**	0

Матрица Лапласса

3	**-1**	**-1**	**-1**	0	0
-1	***2***	**-1**	0	0	0
-1	**-1**	***3***	0	0	**-1**
-1	0	0	***3***	**-1**	**-1**
0	0	0	**-1**	***2***	**-1**
0	0	**-1**	**-1**	**-1**	***3***

Python и не только

- http://current.cs.ucsb.edu/socialmodels/ - библиотека для питона по созданию графов
- https://networkx.github.io/ - библиотека NetworkX. Много чего умеет
- http://gephi.org Gephi: Java платформа для визуализации.