

Анализ сетевых данных

Лекция 3. Random Graph models.
Centrality measures.

30 сентября 2020

План на сегодня

Модели случайных графов

- Preferential attachment
- Configuration model
- Stochastic block model
- Watts-Strogatz model
- Geometric model

Метрики центральности

- Degree centrality
- Harmonic centrality
- Betweenness centrality
- Closeness centrality
- Eigenvector centrality
- PageRank centrality

Зачем?

Когда решается какая то задача в которой фигурируют сети, мы предполагаем что вершины этой сети связаны не просто так “не случайно”, т.е. существует нечто ненаблюдаемое, что является причиной этой связи.

В таком случае бывает полезно зафиксировать какую то максимально общую характеристику эмпирической сети и проверить поведение другой, но уже “случайной” сети с такой же “общей характеристикой”.

Зачем?

- Средняя степень вершины
 - ◆ Erdos-Renyi model
- Scale-free property (power law degree distribution)
 - ◆ Preferential attachment (e.g. Barabasi-Albert model)
- Сеть с данным набором степеней вершин
 - ◆ Configuration model
- High clustering + low Average Path length
 - ◆ Watts-Strogatz model
- Clustering structure
 - ◆ Stochastic-block models (e.g. LFR benchmark, geometric graphs)

Erdos-Renyi random graph

- $G_{n,p}$ model:

Graph with n nodes and for each pair of nodes the probability of an edge between them is equal to p .

- $G_{n,m}$ model:

A randomly selected graph from the set of C_N^m graphs, with $N = n(n-1)/2$, where $n = \text{\#nodes}$ and $m = \text{\#edges}$

Random graph model (Erdos & Renyi, 1959)

Barabasi-Albert model

Цель

Моделировать сети с power law degree distribution.

Идея

Чем выше степень вершины тем выше вероятность того что новая вершина будет соединена с ней ребром.

Barabasi-Albert model

[Barabasi-Albert, 1999](#)

Как?

Детали реализации могут сильно отличаться

1. Стартуем с графа с m_0 вершинами.
2. Добавляем новую вершину и соединяем ее с $m < m_0$ существующими вершинами с вероятностью прямо пропорциональной степени вершины.

$$p_i = \frac{k_i}{\sum_j k_j}$$

Barabasi-Albert model

Power law distribution function:

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

Average path length (analytical result) :

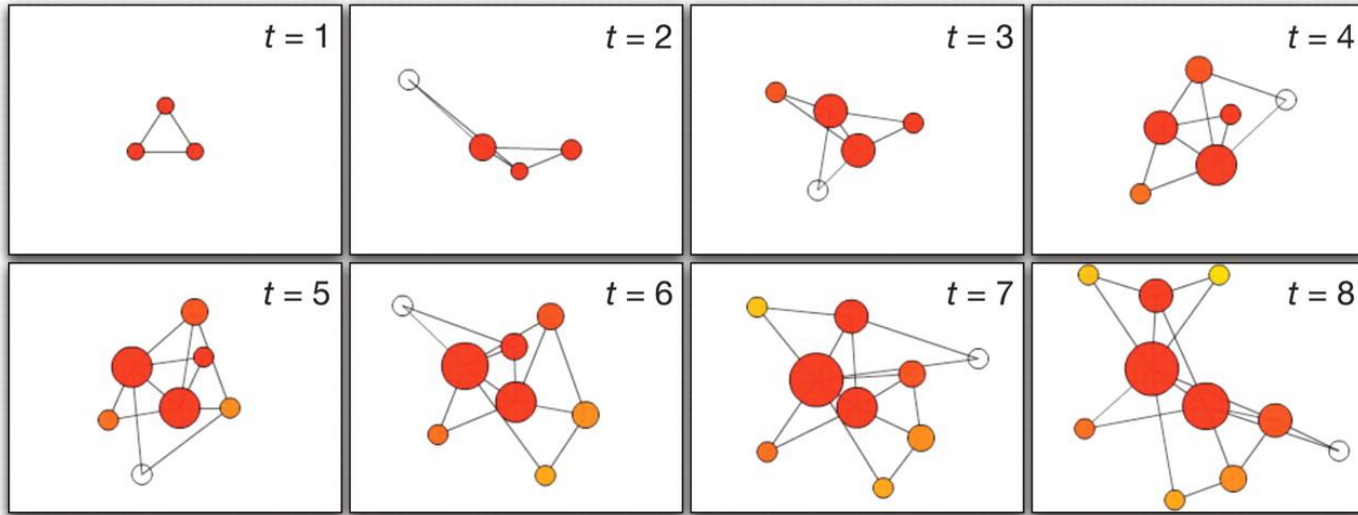
$$\langle L \rangle \sim \log(N) / \log(\log(N))$$

Clustering coefficient (numerical result):

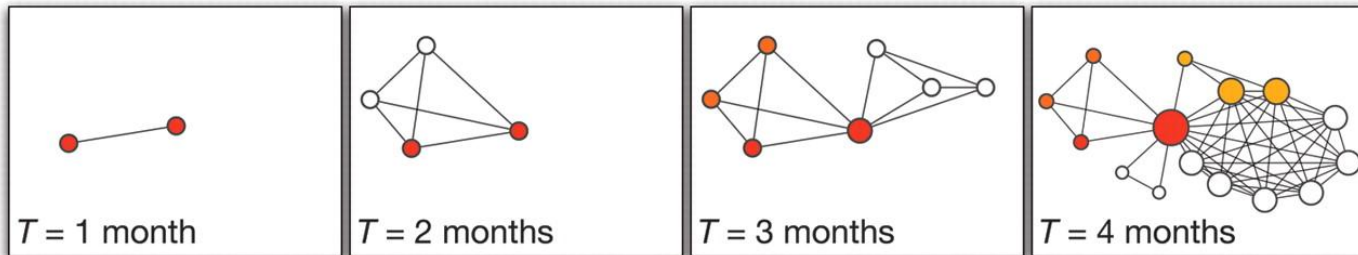
$$C \sim N^{-0.75}$$

Barabasi-Albert model

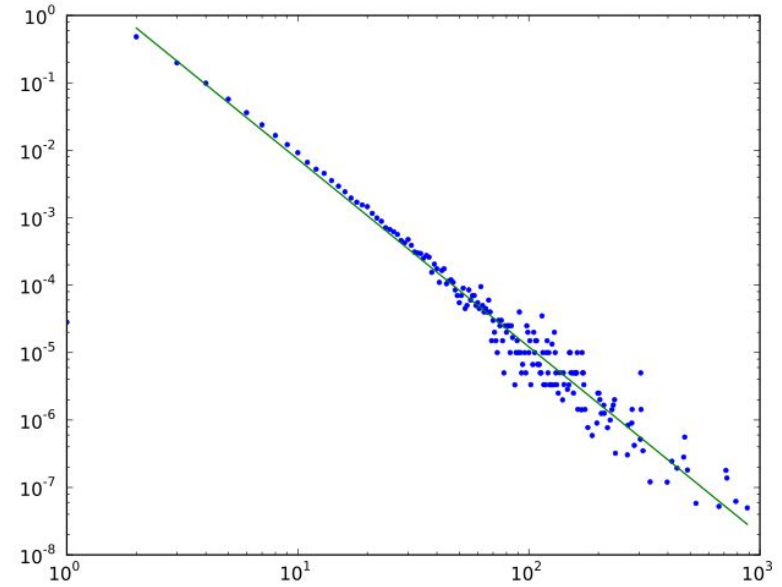
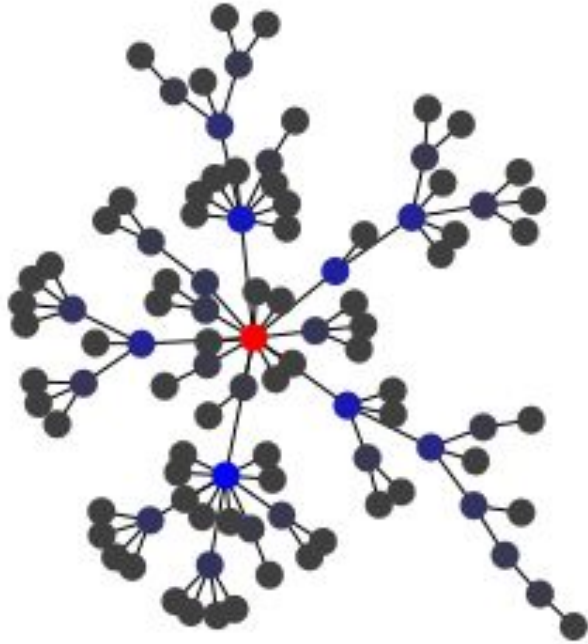
Scale-Free Model



Scientific Collaboration Network



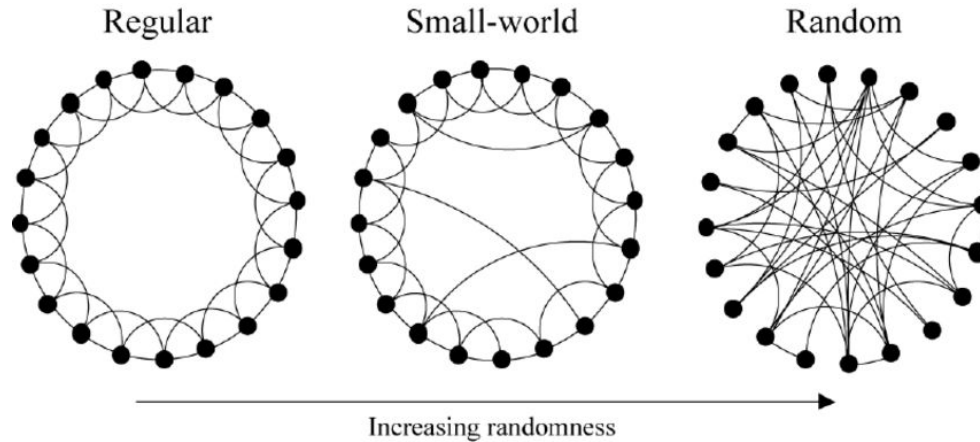
Barabasi-Albert model



Watts-Strogatz model

Начнем с регулярного графа и сделаем его случайным:

1. Начнем с регулярной решетки с n вершинами и k степенью каждой вершины ($k \ll n$)
2. С вероятностью p свапнем **существующее** ребро.

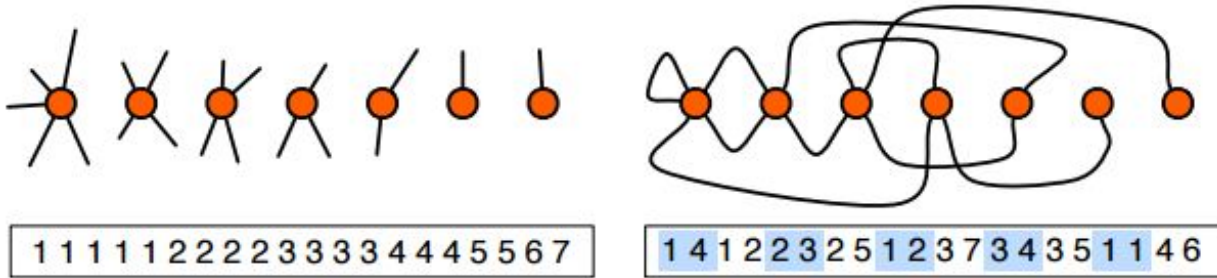


$$0 < p < 1$$

Configuration model

Позволяет моделировать сети с произвольным эмпирическим/теоретическим законом распределения степеней вершин.

1. Сгенерируем набор вершин с заданными степенями заданными “полу-ребрами” (stubs/half-edges).
2. Выберем случайные две вершины со свободными ребрами и объединим их.



Configuration model

Issues

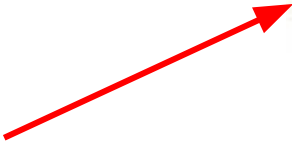
- multi edges

$$\frac{1}{2} \left[\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right]^2$$

- self-loops

$$\frac{\langle k^2 \rangle - \langle k \rangle}{2\langle k \rangle}$$

- low clustering

$$\frac{1}{n} \frac{[\langle k^2 \rangle - \langle k \rangle]^2}{\langle k \rangle^3}$$


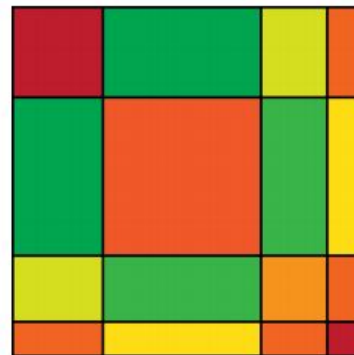
See [Aaron Clauset](#) for details

Stochastic-block models

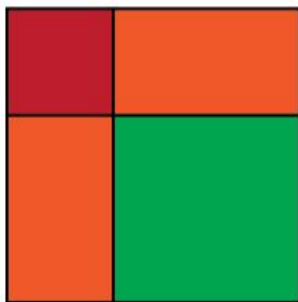
Block model assumes:

1. Each node is assigned to a single community.
2. For every pair of community types, there is a fixed probability of forming a connection.

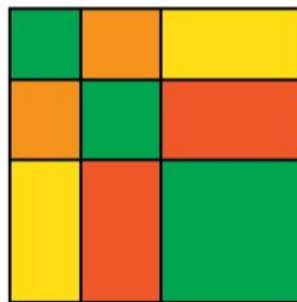
$$\Pr(A_{ij} = 1 \mid c_i = a, c_j = b) = \omega_{ab}$$



Core-periphery



Multipartite

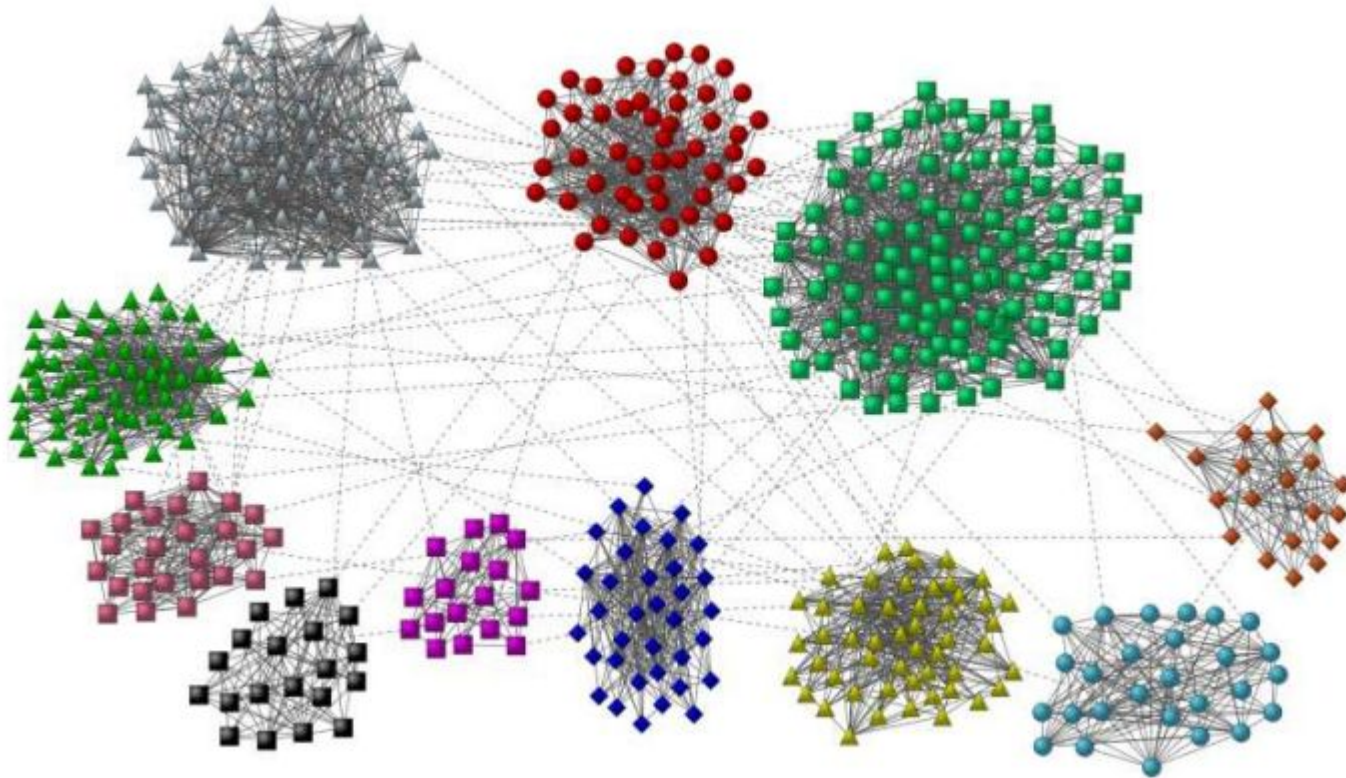


Block models can capture more exotic class-based organization than communities:

- Core-periphery/rich clubs
- Multi-partite

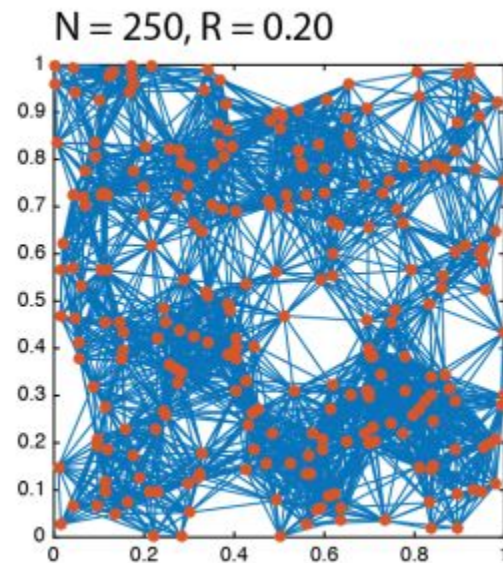
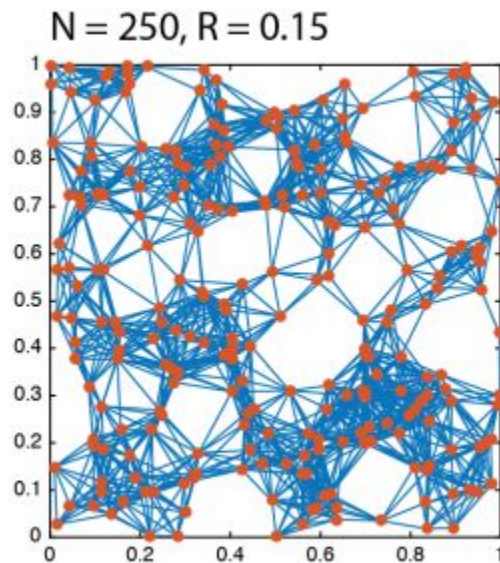
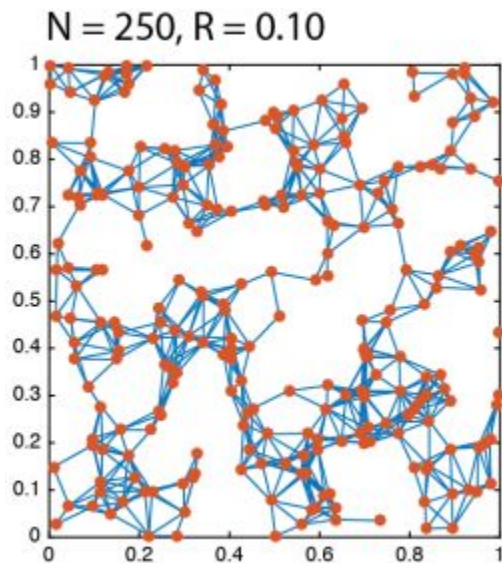
Stochastic-block models

[LFR benchmark](#)

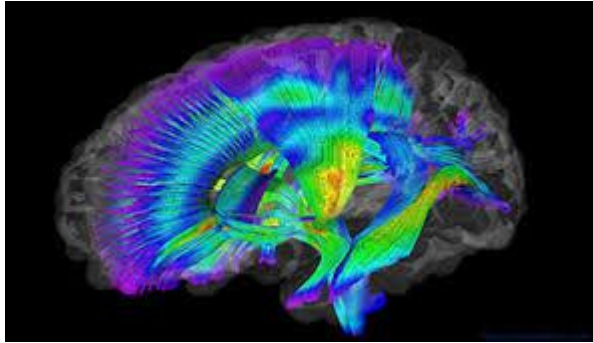


Geometric models

1. Sample N data points from some distribution
2. Connect points that are R distant or less



Geometric models



$$\mathcal{C}(E_1, E_2) = \iint_{E_1, E_2} \lambda(x, y) dx dy.$$

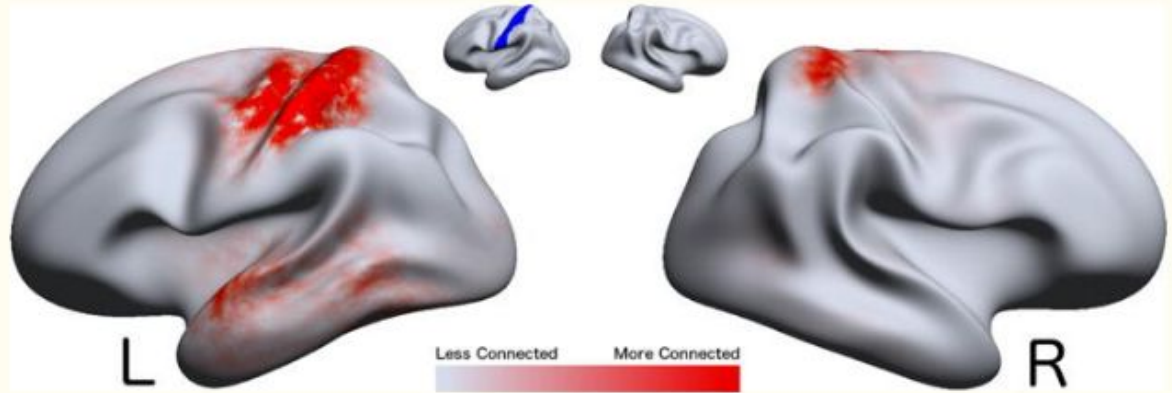


Fig. 2

A visualization of the marginal connectivity $M(x) = \int_{E_i} \hat{\lambda}(x, y) dy$ for the Left Post-central Gyrus region of the DK atlas (Region 57). The region is shown in blue on the inset. Red denotes higher connectivity regions with the blue region.

Метрики центральности



Метрики центральности

- Degree centrality
- Harmonic centrality
- Betweenness centrality
- Closeness centrality
- Eigenvector centrality
- PageRank centrality



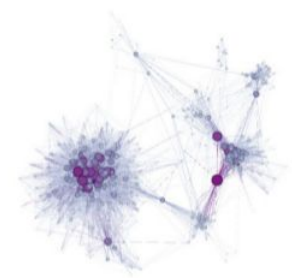
(a) Betweenness centrality



(b) Closeness centrality



(a) Degree centrality



(b) Page Rank

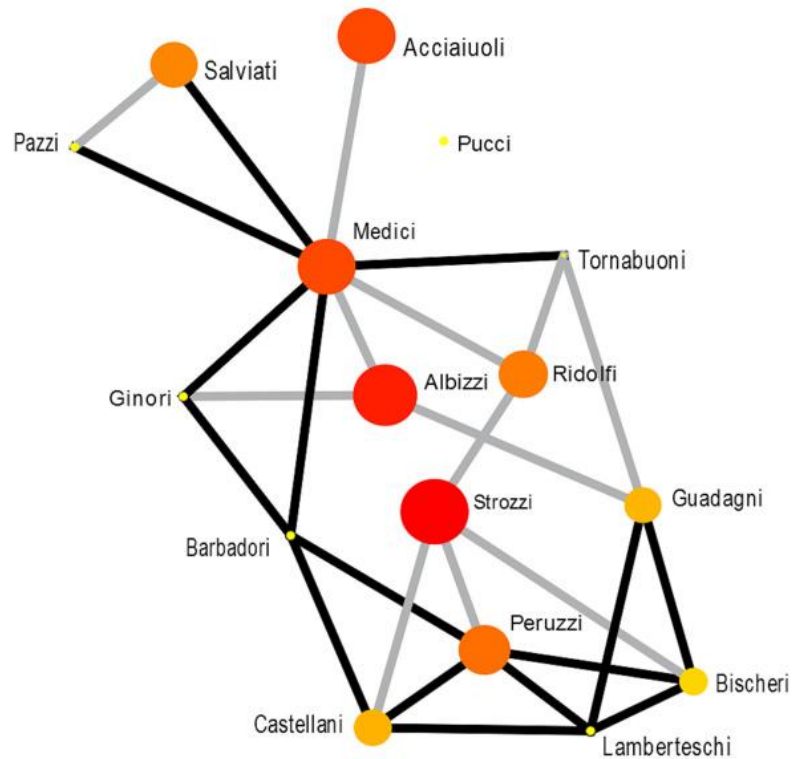
Метрики центральности

theorem

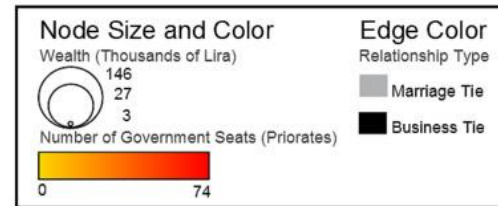
Q

Harmonic centrality	Indegree	PageRank	Page views
0. Poincaré conjecture	0. Pythagorean theorem	0. Strahler number	0. Bayes' theorem
1. Pythagorean theorem	1. Gödel's incompleteness theorems	1. Pythagorean theorem	1. Pythagorean theorem
2. Fermat's Last Theorem	2. Fermat's Last Theorem	2. Fermat's Last Theorem	2. Euler's formula
3. Bell's theorem	3. Bayes' theorem	3. Gödel's incompleteness theorems	3. Central limit theorem
4. Reductio ad absurdum	4. Central limit theorem	4. Central limit theorem	4. Binomial theorem
5. Gödel's incompleteness theorems	5. Reductio ad absurdum	5. Bayes' theorem	5. Fermat's Last Theorem
6. Vafa–Witten theorem	6. Chain rule	6. Law of large numbers	6. De Morgan's laws
7. Binomial theorem	7. Law of large numbers	7. Euler's formula	7. Gödel's incompleteness theorems
8. Bayes' theorem	8. Triangle inequality	8. Reductio ad absurdum	8. Chain rule
9. Arrow's impossibility theorem	9. Binomial theorem	9. Chain rule	9. Euler's identity

Флорентийские семьи



Padgett's Florentine Families



Degree centrality

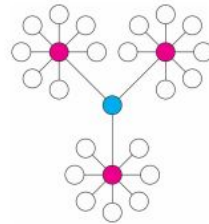
Degree centrality: number of nearest neighbors

$$C_D(i) = k(i) = \sum_j A_{ij} = \sum_j A_{ji}$$

Normalized degree centrality

$$C_D^*(i) = \frac{1}{n-1} C_D(i) = \frac{k(i)}{n-1}$$

High centrality degree -direct contact with many other actors



Next few slides from [Zhukov](#)

Closeness centrality

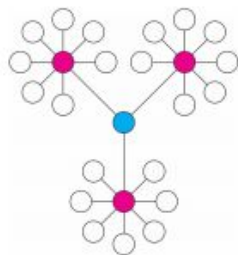
Closeness centrality: how close an actor to all the other actors in network

$$C_C(i) = \frac{1}{\sum_j d(i,j)}$$

Normalized closeness centrality

$$C_C^*(i) = (n - 1)C_C(i) = \frac{n - 1}{\sum_j d(i,j)}$$

High closeness centrality - short communication path to others, minimal number of steps to reach others



[*** Harmonic centrality $C_H(i) = \sum_j \frac{1}{d(i,j)}$ ***]

Betweenness centrality

Betweenness centrality: number of shortest paths going through the actor

$\sigma_{st}(i)$

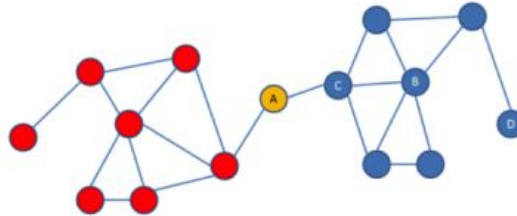
$$C_B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Normalized betweenness centrality

$$C_B^*(i) = \frac{2}{(n-1)(n-2)} C_B(i) = \frac{2}{(n-1)(n-2)} \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

High betweenness centrality - vertex lies on many shortest paths

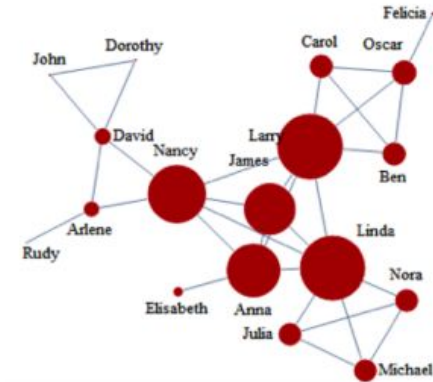
Probability that a communication from s to t will go through i



Eigenvector centrality

Importance of a node depends on the importance of its neighbors
(recursive definition)

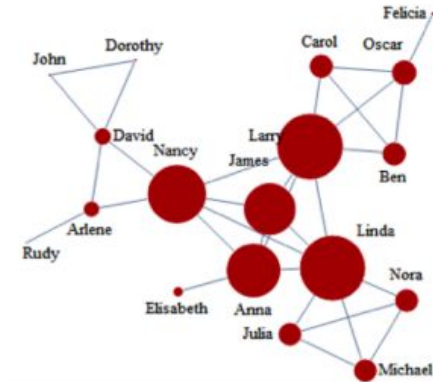
$$v_i \leftarrow \sum_j A_{ij} v_j$$
$$v_i = \frac{1}{r} \sum_j A_{ij} v_j$$



Eigenvector centrality

Importance of a node depends on the importance of its neighbors
(recursive definition)

$$v_i \leftarrow \sum_j A_{ij} v_j$$
$$v_i = \frac{1}{\lambda} \sum_j A_{ij} v_j$$



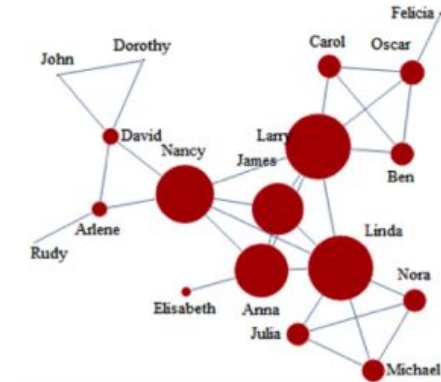
Eigenvector centrality

Importance of a node depends on the importance of its neighbors
(recursive definition)

$$v_i \leftarrow \sum_j A_{ij} v_j$$

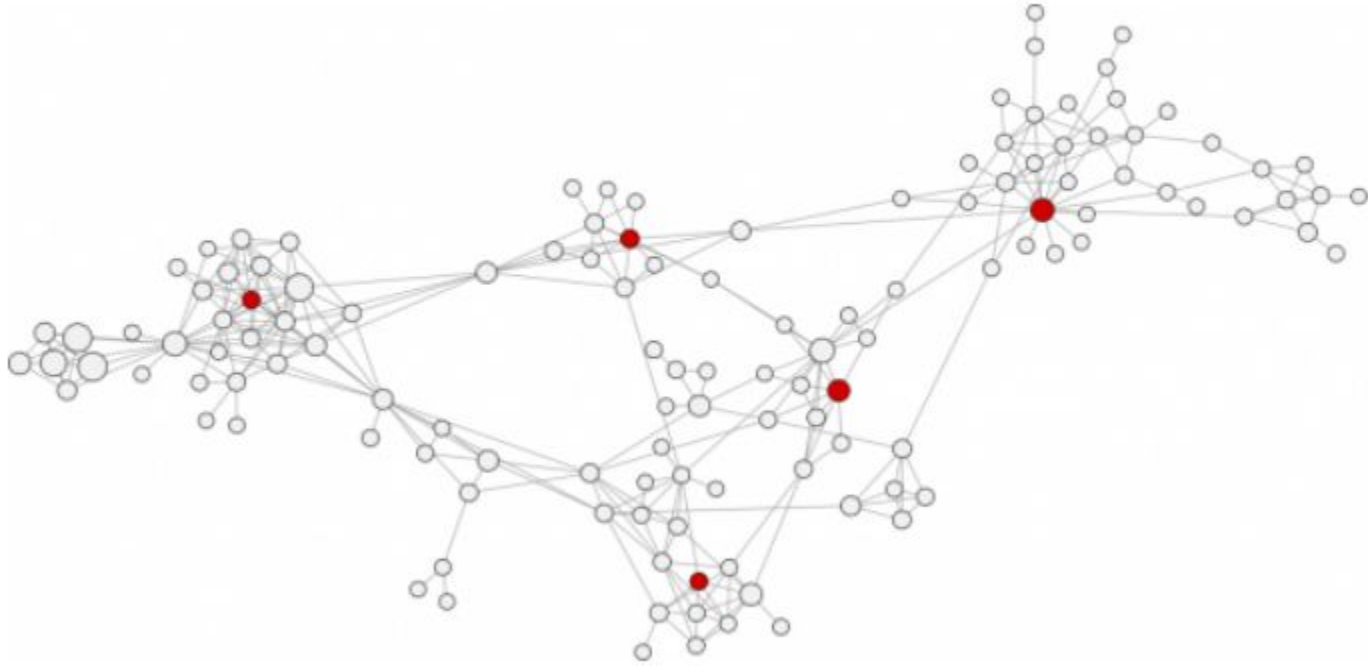
$$v_i = \frac{1}{\lambda} \sum_j A_{ij} v_j$$

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$



Select an eigenvector associated with largest eigenvalue $\lambda = \lambda_1$, $\mathbf{v} = \mathbf{v}_1$

Closeness centrality



Betweenness centrality



Eigenvector centrality



Directed case

- Вышеперечисленные метрики центральности (degree, closeness, betweenness) могут быть расширены на случай направленных графов с некоторыми ограничениями.
- Кроме того существуют метрики предложенные непосредственно для направленных графов (и в дальнейшем определенные для ненаправленных)

Katz centrality

Node prestige depends on prestige of directly connected actors

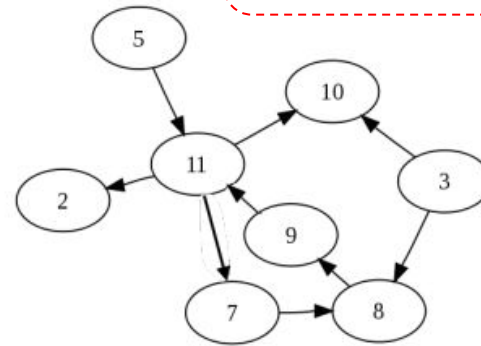
- iterate

$$p_i \leftarrow \sum_{j \in N(i)} p_j = \sum_j A_{ji} p_j$$

$$\mathbf{p}^{t+1} = \mathbf{A}^T \mathbf{p}^t, \quad \mathbf{p}^{t=0} = \mathbf{p}_0$$

- Difficulties:
 - Absorbing nodes
 - Source nodes
 - Cycles

** Solution to $\mathbf{p} = \mathbf{A}^T \mathbf{p}$ might not exist. Nontrivial solution only if $\det(\mathbf{I} - \mathbf{A}^T) = 0$. Need to constraint matrix



Важность вершины зависит от важности вершин на нее ссылающихся

Page rank

[Leskovec explanation](#) ~ 12 mins

[Google search engine](#) ~ 8 mins

The latter is pretty much outdated but the ideas stay the same (probably)

- Random walk on graph

$$p_i^{t+1} = \sum_{j \in N(i)} \frac{p_j^t}{d_j^{out}} = \sum_j \frac{A_{ji}}{d_j^{out}} p_j$$

$$\mathbf{P} = \mathbf{D}^{-1} \mathbf{A}, \quad \mathbf{D}_{ii} = \text{diag}\{d_i^{out}\}$$

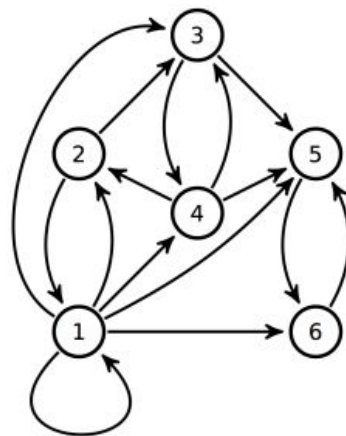
$$\mathbf{p}^{t+1} = \mathbf{P}^T \mathbf{p}^t$$

- with teleportation

$$\mathbf{p}^{t+1} = \alpha \mathbf{P}^T \mathbf{p}^t + (1 - \alpha) \frac{\mathbf{e}}{n}$$

Perron-Frobenius Theorem guarantees existence and uniqueness of the solution to

$$\mathbf{p} = \alpha \mathbf{P}^T \mathbf{p} + (1 - \alpha) \frac{\mathbf{e}}{n}$$



Флорентийские семьи (свадьбы)

