Link Analysis and Structural Similarity

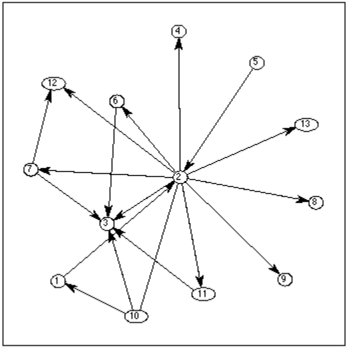
I. Makarov & L.E. Zhukov

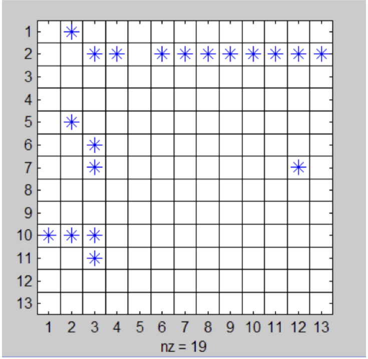
BigData Academy MADE from VK

Social Network Analysis and Machine Learning on Graphs

Lecture outline

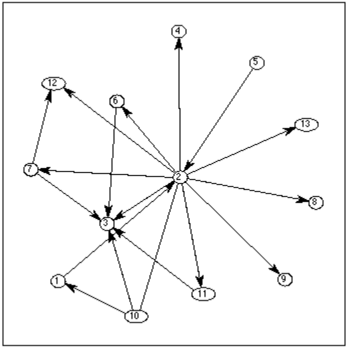
Graph $G(E, V)$, $|V| = n$, $|E| = m$
Adjacency matrix $\mathbf{A}^{n \times n}$, A_{ij}, edge $i \to j$





Graph is directed, matrix is non-symmetric: $\mathbf{A}^T \neq \mathbf{A}$, $A_{ij} \neq A_{ji}$

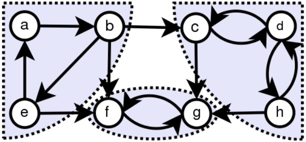
Graph theory



- sinks: zero out degree nodes, $k_{out}(i) = 0$, absorbing nodes
- sources: zero in degree nodes, $k_{in}(i) = 0$

Graph theory

- Graph is **strongly connected** if every vertex is reachable form every other vertex.
- **Strongly connected components** are partitions of the graph into subgraphs that are strongly connected



- In strongly connected graphs there is a path is each direction between any two pairs of vertices

image from Wikipedia

Graph theory

- A directed graph is **aperiodic** if the greatest common divisor of the lengths of its cycles is one (there is no integer $k > 1$ that divides the length of every cycle of the graph)

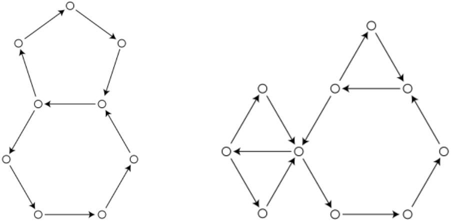
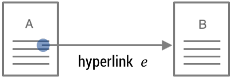


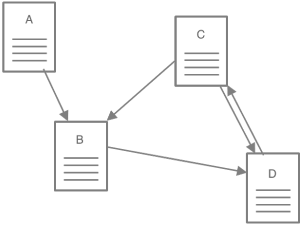
image from Wikipedia

Web as a graph

- Hyperlinks - implicit endorsements



- Web graph - graph of endorsements (sometimes reciprocal)



Random walk

- Random walk on a directed graph:

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

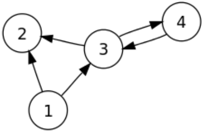
$$\mathbf{D}_{ii} = diag\{d_i^{out}\}$$

$$\mathbf{p}^{t+1} = (\mathbf{D}^{-1}\mathbf{A})^T \mathbf{p}^t$$

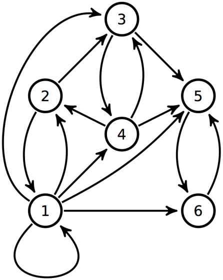
$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$$

- Power iterations

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



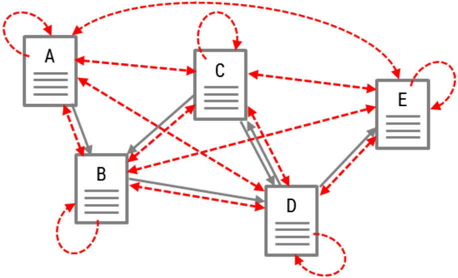
a)



b)

PageRank

"PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The **probability** that the random surfer visits a page is its **PageRank**."



Sergey Brin and Larry Page, 1998

PageRank formulation

- Power iterations:

$$\mathbf{p} \leftarrow \alpha \mathbf{P}^T \mathbf{p} + (1-\alpha)\frac{\mathbf{e}}{n}, \qquad \alpha \text{ - teleportation coefficient}$$

- Sparse linear system:

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

- Eigenvalue problem ($\lambda = 1$):

$$\left(\alpha \mathbf{P}^T + (1-\alpha)\mathbf{E}\right)\mathbf{p} = \lambda \mathbf{p}$$

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$$

Perron-Frobenius Theorem

Perron-Frobenius theorem (Fundamental Theorem of Markov Chains)
If matrix is

- stochastic (non-negative and rows sum up to one, describes Markov chain)
- irreducible (strongly connected graph)
- aperiodic

then

$$\exists \lim_{t \to \infty} \bar{\mathbf{p}}^t = \bar{\pi}$$

and can be found as a left eigenvector

$$\bar{\pi} P = \lambda \bar{\pi}, \quad \text{where} \quad ||\bar{\pi}||_1 = 1, \lambda = 1$$

$\bar{\pi}$ - stationary distribution of Markov chain, row vector

Oscar Perron, 1907, Georg Frobenius,1912.

PageRank variations

- Power iterations

$$\mathbf{p} \leftarrow \alpha \mathbf{P}^T \mathbf{p} + (1-\alpha)\mathbf{v}, \quad \mathbf{v} \text{ - teleportation vector}$$

$$\mathbf{P}' = \alpha \mathbf{P} + (1-\alpha)\mathbf{e}\mathbf{v}^T$$

$$\mathbf{p} \leftarrow \mathbf{P}'^T \mathbf{p}, \; ||\mathbf{p}|| = 1$$

- Topic specific PageRank

v - set of pages on specific topics

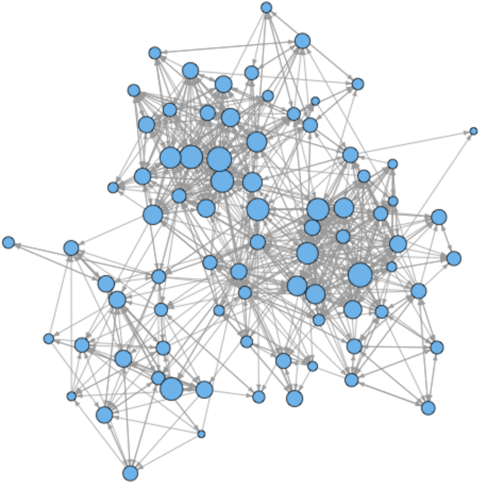
- TrustRank

v - set of trusted pages

- Personalized PageRank

v - set of personal preference pages

PageRank



PageRank beyond the Web

1. GeneRank
2. ProteinRank
3. FoodRank
4. SportsRank
5. HostRank
6. TrustRank
7. BadRank
8. ObjectRank
9. ItemRank
10. ArticleRank
11. BookRank
12. FutureRank
13. TimedPageRank
14. SocialPageRank
15. DiffusionRank
16. ImpressionRank
17. TweetRank
18. TwitterRank
19. ReversePageRank
20. PageTrust
21. PopRank
22. CiteRank
23. FactRank
24. InvestorRank
25. ImageRank
26. VisualRank
27. QueryRank
28. BookmarkRank
29. StoryRank
30. PerturbationRank
31. ChemicalRank
32. RoadRank
33. PaperRank
34. Etc...

Hubs and Authorities (HITS)

Citation networks. Reviews vs original research (authoritative) papers

- authorities, contain useful information, a_i
- hubs, contains links to authorities, h_i

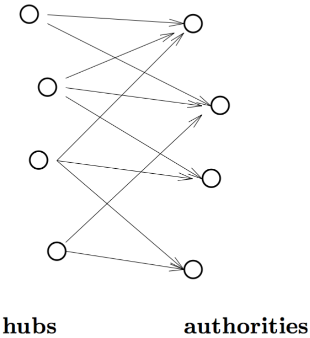
Mutual recursion

- Good authorities reffered by good hubs

$$a_i \leftarrow \sum_j A_{ji} h_j$$

- Good hubs point to good authorities

$$h_i \leftarrow \sum_j A_{ij} a_j$$



Jon Kleinberg, 1999

HITS

System of linear equations

$$\begin{aligned} \mathbf{a} &= \alpha \mathbf{A}^T \mathbf{h} \\ \mathbf{h} &= \beta \mathbf{A} \mathbf{a} \end{aligned}$$

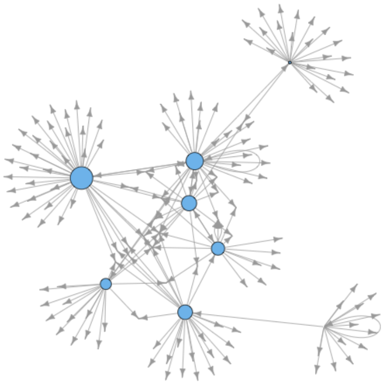
Symmetric eigenvalue problem

$$\begin{aligned} (\mathbf{A}^T \mathbf{A})\mathbf{a} &= \lambda \mathbf{a} \\ (\mathbf{A} \mathbf{A}^T)\mathbf{h} &= \lambda \mathbf{h} \end{aligned}$$

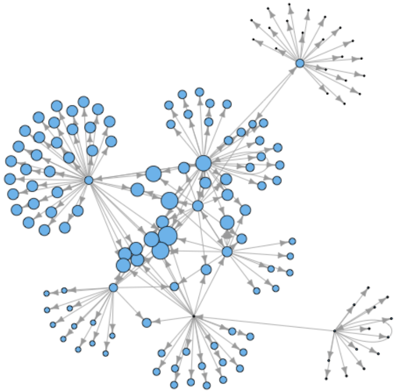
where eigenvalue $\lambda = (\alpha\beta)^{-1}$

Hubs and Authorities

Hubs



Authorities



References

- The PageRank Citation Ranknig: Bringing Order to the Web. S. Brin, L. Page, R. Motwany, T. Winograd, Stanford Digital Library Technologies Project, 1998
- Authoritative Sources in a Hyperlinked Environment. Jon M. Kleinberg, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms,
- Graph structure in the Web, Andrei Broder et all. Procs of the 9th international World Wide Web conference on Computer networks, 2000
- A Survey of Eigenvector Methods of Web Information Retrieval. Amy N. Langville and Carl D. Meyer, 2004
- PageRank beyond the Web. David F. Gleich, arXiv:1407.5107, 2014

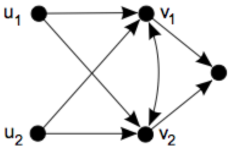
Patterns of relations

- Global, statistical properties of the networks:
  - average node degree (degree distribution)
  - average clustering
  - average path length
- Local, per vertex properties:
  - node centrality
  - page rank
- Pairwise properties:
  - node equivalence
  - node similarity
  - correlation between pairs of vertices (node values)

Structural equivalence

Definition

Structural equivalence: two vertices are structurally equivalent if their respective sets of in-neighbors and out-neighbors are the same

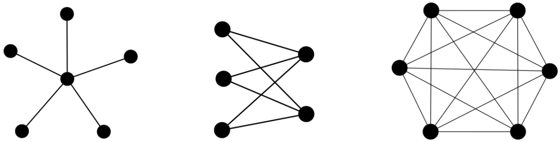


	u1	u2	v1	v2	w
u1	0	0	1	1	0
u2	0	0	1	1	0
v1	0	0	0	1	1
v2	0	0	1	0	1
w	0	0	0	0	0

rows and columns of adjacency matrix of structurally equivalent nodes are identical, "connect to the same neighbors"

Structural equivalence

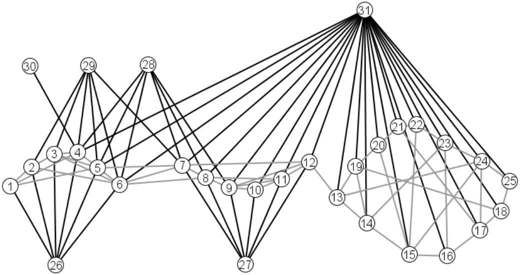
- In order for adjacent vertices to be structurally equivalent, they should have self loops.
- Sometimes called "strong structural equivalence"
- Sometimes relax requirements for self loops for adjacent nodes



Structural similarity

Definition

Two nodes are similar to each other if they share many neighbors.



Similarity measures

- Jaccard similarity

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

- Cosine similarity (vectors in n-dim space)

$$\sigma(v_i, v_j) = cos(\theta_{ij}) = \frac{\mathbf{v}_i^T \mathbf{v}_j}{|\mathbf{v}_i||\mathbf{v}_j|} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum A_{ik}^2} \sqrt{\sum A_{jk}^2}}$$

- Pearson correlation coefficient:

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

Similarity measures

- Unweighted undirected graph $A_{ik} = A_{ki}$, binary matrix, only 0 and 1
- $\sum_k A_{ik} = \sum_k A_{ik}^2 = k_i$ - node degree
- $\sum_k A_{ik} A_{kj} = (A^2)_{ij} = n_{ij}$ - number of shared neighbors

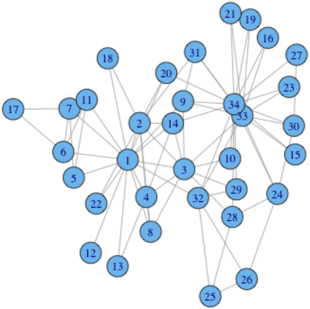
- Cosine similarity (vectors in n-dim space)

$$\sigma(v_i, v_j) = cos(\theta_{ij}) = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

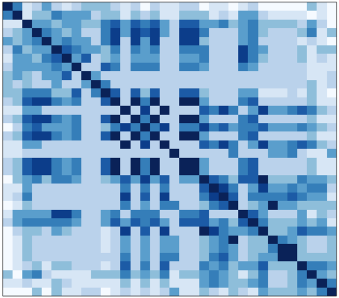
- Pearson correlation coefficient:

$$r_{ij} = \frac{n_{ij} - \frac{k_i k_j}{n}}{\sqrt{k_i - \frac{k_i^2}{n}} \sqrt{k_j - \frac{k_j^2}{n}}}$$

Similarity matrix



Graph



Node similarity matrix

Regular equivalence

Definition

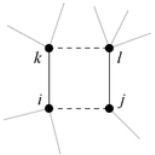
Two vertices are regularly equivalent if they are equally related to equivalent others.

- Quantitative measure - similarity score σ_{ij} (recursive definition):

$$\sigma_{ij} = \alpha \sum_{k,l} A_{ik} A_{jl} \sigma_{kl}$$

- should have high σ_{ii} - self similarity

$$\sigma_{ij} = \alpha \sum_{k,l} A_{ik} A_{jl} \sigma_{kl} + \delta_{ij}$$



Regular similarity

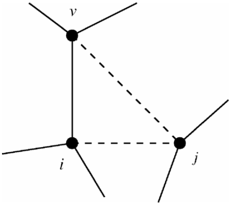
- A vertex j is similar to vertex i (dashed line) if i has a network neighbor v (solid line) that is itself similar to j

$$\sigma_{ij} = \alpha \sum_{v} A_{iv} \sigma_{vj} + \delta_{ij}$$

$$\boldsymbol{\sigma} = \alpha \mathbf{A} \boldsymbol{\sigma} + \mathbf{I}$$

- Closed form solution:

$$\boldsymbol{\sigma} = (\mathbf{I} - \alpha \mathbf{A})^{-1}$$



Leicht, Holme, and Newman, 2006

SimRank

- G - directed graph
- Two vertices are similar if they are referenced by similar vertices
- $s(a, b)$ - similarity between a and b, $I()$ - set of in-neighbours

$$s(a, b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{I(a)} \sum_{j=1}^{I(b)} s(I_i(a), I_j(b)), \quad a \neq b$$

$$s(a, a) = 1$$

- Matrix notation:

$$S_{ij} = \frac{C}{k_i k_j} \sum_{k,m} A_{ki} A_{mj} S_{km}$$

- Iterative solution starting from $s_0(i, j) = \delta_{ij}$

Jeh and Widom, 2002

Mixing patterns

Network mixing patterns

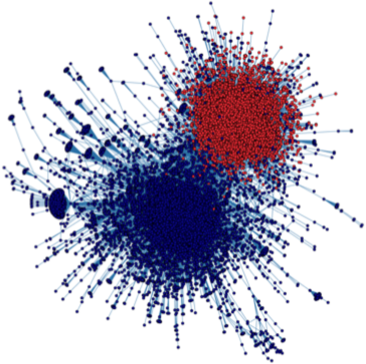
- **Assortative mixing**, "like links with like", attributed of connected nodes tend to be more similar than if there were no such edge
- **Disassortative mixing**, "like links with dislike", attributed of connected nodes tend to be less similar than if there were no such edge

Vertices can mix on any vertex attributes (age, sex, geography in social networks), unobserved attributes, vertex degrees

Examples:
assortative mixing - in social networks political beliefs, obesity, race
disassortative mixing - dating network, food web (predator/prey), economic networks (producers/consumers)

Assortative mixing

- Political polarization on Twitter: political retweet network ,red color - "right-learning" users, blue color - "left learning" users



- Assortative mixing = homophily

Conover et al., 2011

Mixing by categorical attributes

- Vertex categorical attribute (c_i -label): color, gender, ethnicity
- How much more often do attributes match across edges than expected at random?
- Modularity :

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

- m_c - number of edges between vertices with same attributes
  $\langle m_c \rangle$ - expected number of edges within the same class in random network
- Assortativity coefficient:

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

Mixing by scalar values

- Vertex scalar value (attribute) - x_i
- How much more similar are attributes across edges than expected at random?
- Average and covariance over edges

$$
\begin{aligned}
\langle x \rangle &= \frac{\sum_i k_i x_i}{\sum_i k_i} = \frac{1}{2m} \sum_i k_i x_i = \frac{1}{2m} \sum_{ij} A_{ij} x_i \\
var &= \frac{1}{2m} \sum_{ij} A_{ij} (x_i - \langle x \rangle)^2 = \frac{1}{2m} \sum_i k_i (x_i - \langle x \rangle)^2 \\
cov &= \frac{1}{2m} \sum_{ij} A_{ij} (x_i - \langle x \rangle)(x_j - \langle x \rangle)
\end{aligned}
$$

- Assortativity coefficient

$$
r = \frac{cov}{var} = \frac{\sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j}{\sum_{ij} \left(k_i \delta_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j}
$$

Mixing by node degree

- Assortative mixing by node degree, $x_i \leftarrow k_i$

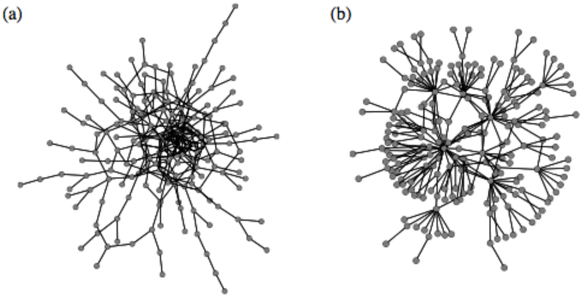
$$r = \frac{\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j}{\sum_{ij}\left(k_i \delta_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j}$$

- Computations:
  $S_1 = \sum_i k_i = 2m$
  $S_2 = \sum_i k_i^2$
  $S_3 = \sum_i k_i^3$
  $S_e = \sum_{ij} A_{ij} k_i k_j$
- Assortatitivity coefficient

$$r = \frac{S_e S_1 - S_2^2}{S_3 S_1 - S_2^2}$$

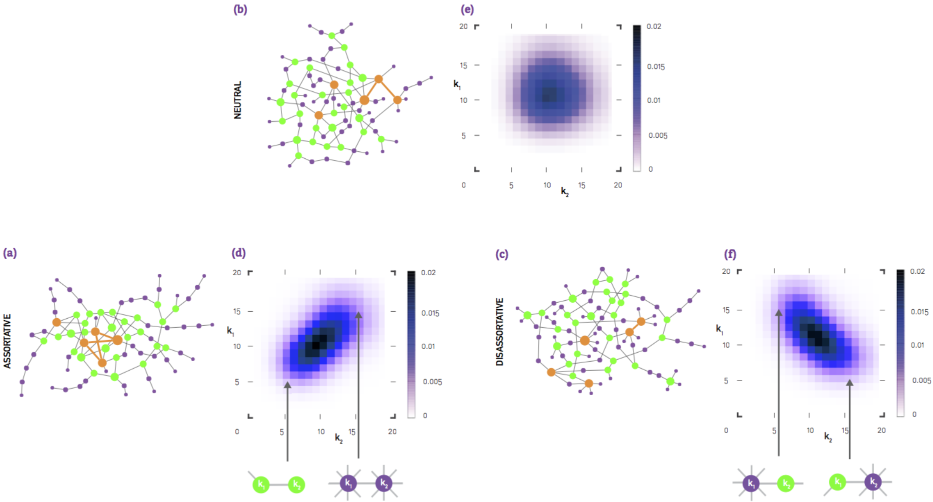
Mixing by node degree

- Assortative network: interconnected high degree nodes - core, low degree nodes - periphery
- Disassortative network: high degree nodes connected to low degree nodes, star-like structure



Assortative network Disassortative network

Degree correlation



from A.L. Barabasi, 2016

References

- White, D., Reitz, K.P. Measuring role distance: structural, regular and relational equivalence. Technical report, University of California, Irvine, 1985
- S. Borgatti, M. Everett. The class of all regular equivalences: algebraic structure and computations. Social Networks, v 11, p65-68, 1989
- E. A. Leicht, P.Holme, and M. E. J. Newman. Vertex similarity in networks. Phys. Rev. E 73, 026120, 2006
- G. Jeh and J. Widom. SimRank: A Measure of Structural-Context Similarity. Proceedings of the eighth ACM SIGKDD , p 538-543. ACM Press, 2002
- M. E. J. Newman. Assortative mixing in networks. Phys. Rev. Lett. 89, 208701, 2002.
- M. Newman. Mixing patterns in networks. Phys. Rev. E, Vol. 67, p 026126, 2003
- M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a Feather: Homophily in Social Networks, Annu. Rev. Sociol, 27:415-44, 2001.