# COMPLEX NETWORKS

# DESCRIBING NODES

Node structural indices / centralities

# NODE

- We can measure nodes importance using so-called **centrality**.

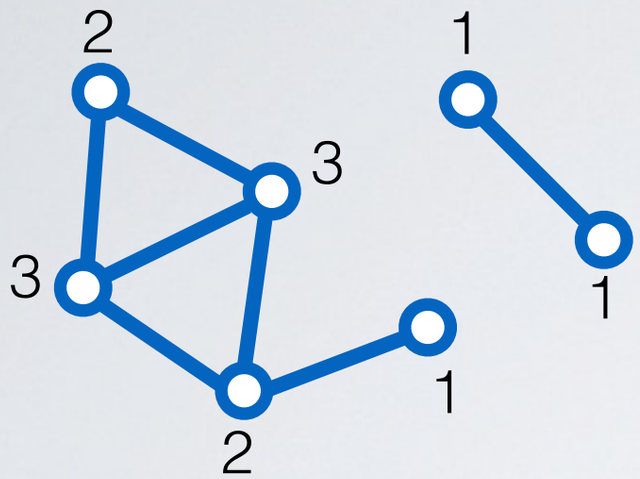
- Poor terminology: nothing to do with being central in general

- Usage:
  - Some centralities have straightforward interpretation
  - Centralities can be used as *node features* for machine learning on graph
    - (Classification, link prediction, …)

# Degree centrality - recap

## Number of connections of a node

- Undirected network



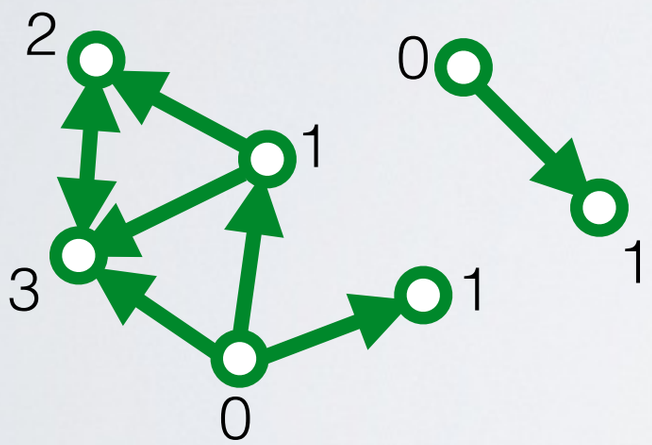

$$k_i = A_{i1} + A_{i2} + ... + A_{iN} = \sum_j^N A_{ij}$$

$$m = \frac{\sum_i k_i}{2} \qquad \text{where} \qquad m = |E|$$

mean degree $$\langle k \rangle = \frac{1}{N} \sum_i^N k_i$$

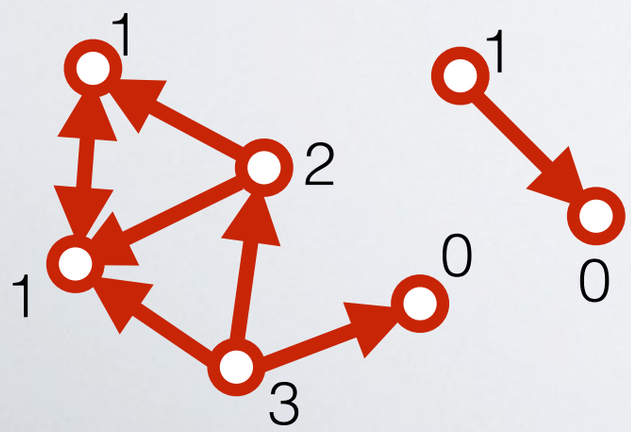
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

- Directed network




In degree $$k_i^{in} = \sum_j^N A_{ij}$$

Out degree $$k_j^{out} = \sum_i^N A_{ij}$$




$$m = \sum_i^N k_i^{in} = \sum_j^N k_j^{out} = \sum_{ij} A_{ij}$$

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^N k_i^{in} = \frac{1}{N} \sum_{j=1}^N k_j^{out} = \langle k^{out} \rangle$$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

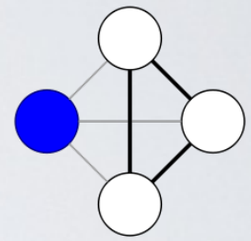
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

# NODE DEGREE

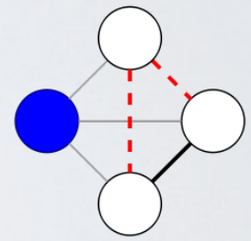
- **Degree**: how many neighbors
- Often enough to find important nodes
  - Main characters of a series talk with the more people
  - Largest airports have the most connections
  - …
- But not always
  - Facebook users with the most friends are spam
  - Webpages/wikipedia pages with most links are simple lists of references
  - …

# NODE CLUSTERING COEFFICIENT

- **Clustering coefficient**: density of neighborhood
- Tells you if the neighbors of the node are connected

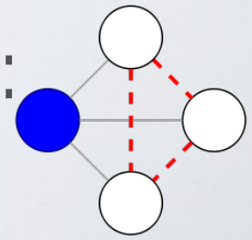


c = 1



c = 1/3

- Be careful!
  - Degree 2: value 0 or 1
  - Degree 1000: Not 0 or 1 (usually)
  - Ranking them is not meaningful
- Sometimes used as a proxy for "communities" belonging:
  - If node belong to single group: high CC
  - If node belong to several groups: lower CC



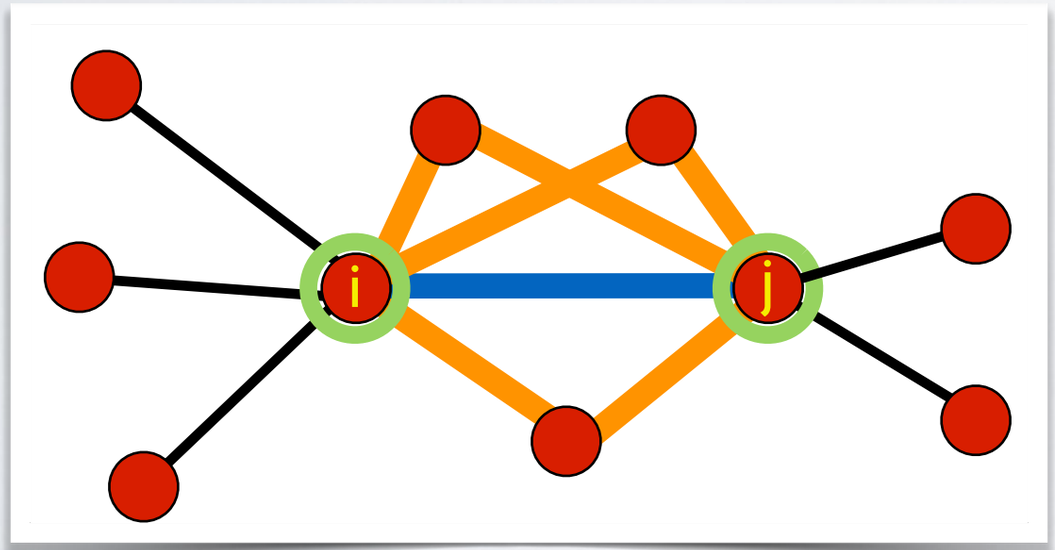
c = 0

# Link clustering coefficient: Overlap

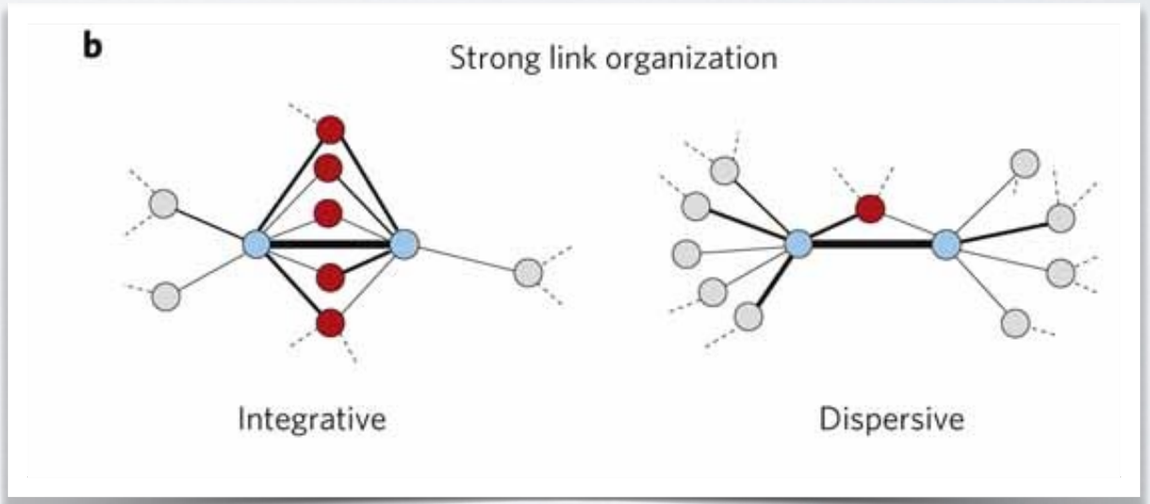
- Link property
- Fraction of common neighbors of a connected pair
- Jaccard Coefficient of neighborhoods

**Edge Clustering** $C^e$ of an edge $(u, v)$ is the fraction of the neighbors of at least one of the two nodes which are neighbors of both of them, i.e.,

$$C^e(u,v) = \frac{|N_u \cap N_v|}{|N_u \cup N_v| - 2}$$




$$C^e(i,j) = \frac{3}{8}$$



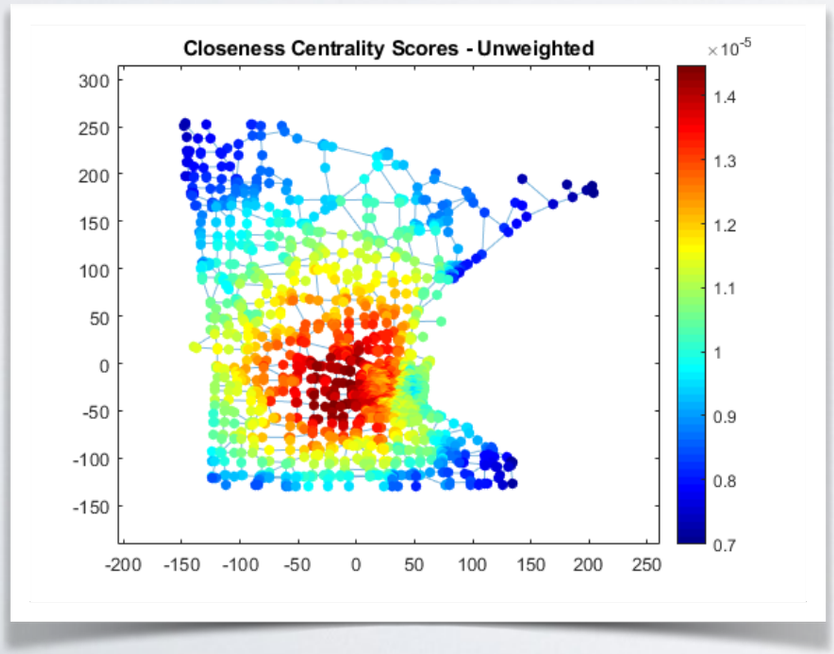

Pajevic, S., & Plenz, D. (2012). The organization of strong links in complex networks. *Nature Physics*, *8*(5), 429-436.

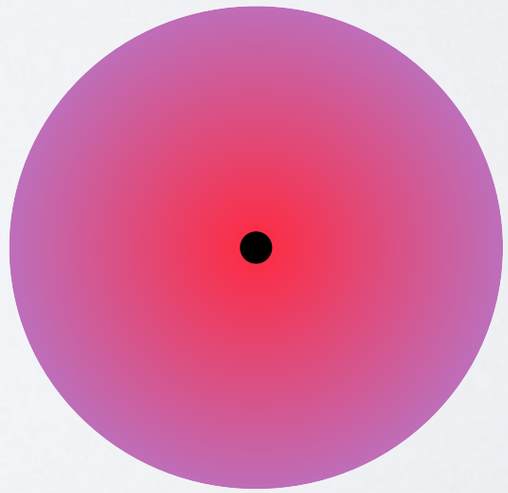
# FARNESS, CLOSENESS

# HARMONIC CENTRALITY

# FARNESS, CLOSENESS

- How close the node is to all other nodes
- Parallel with the center of a figure:
  - Center of a circle is the point of shorter average distance to any points in the circle



# FARNESS, CLOSENESS

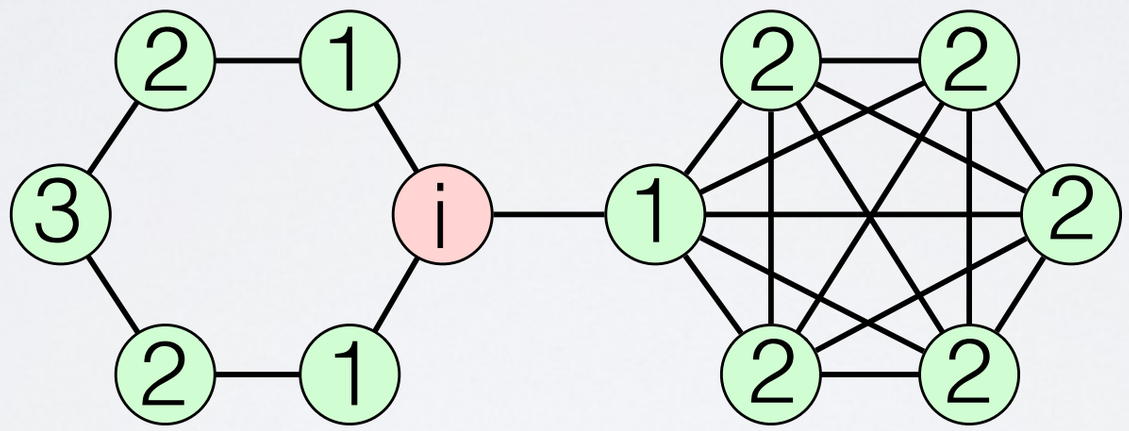
**Farness:** Average distance to all other nodes in the graph

$$\text{Farness}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \ell_{u,v}$$

# CLOSENESS CENTRALITY

**Closeness:** Inverse of the farness, i.e., how close the node is to all other nodes in term of shortest paths.

$$\text{Closeness}(u) = \frac{N-1}{\sum_{v \in V \setminus u} \ell_{u,v}}$$



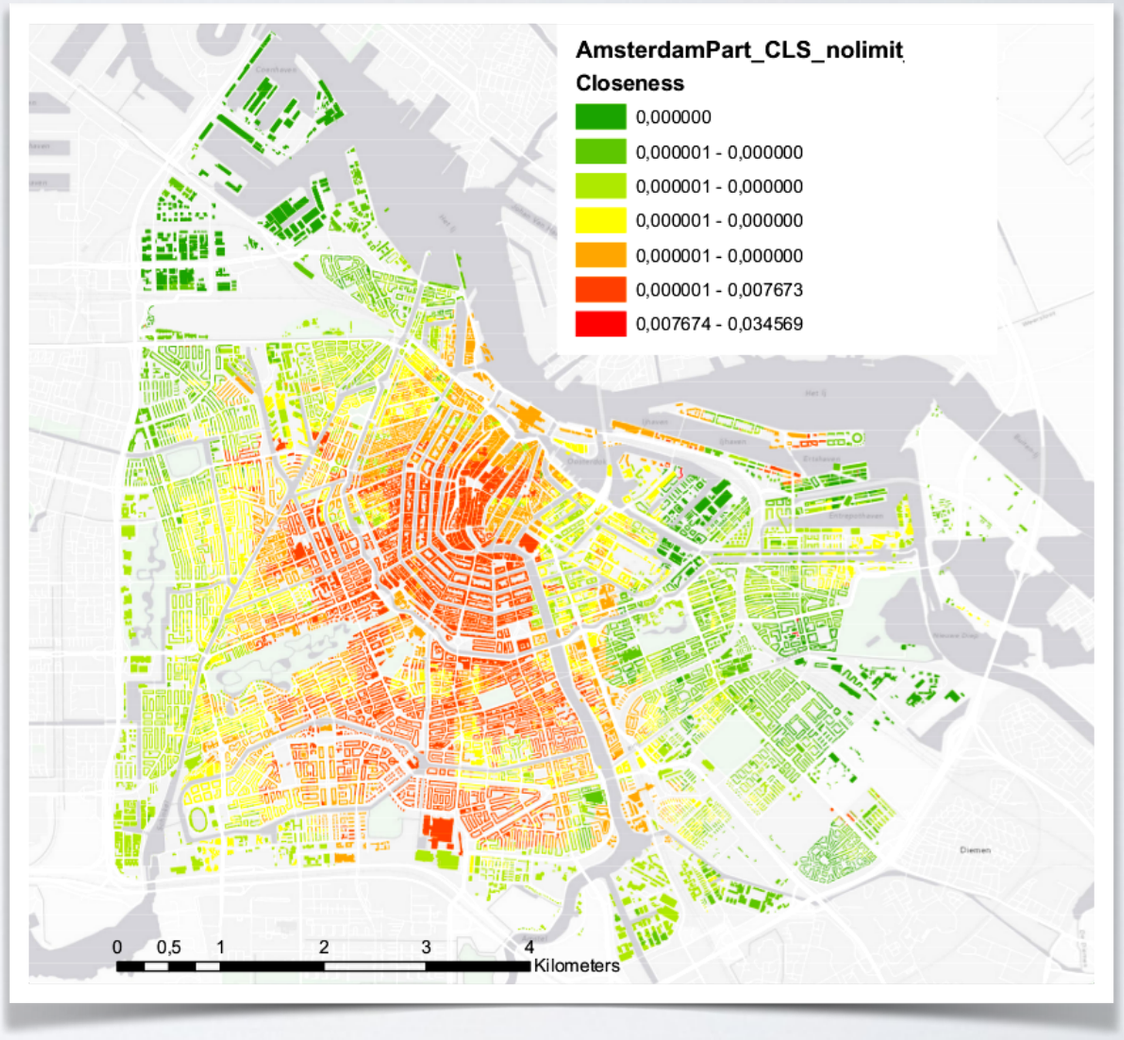

$$C_{cl}(i) = \frac{12-1}{(3\times 1 + 7\times 2 + 1\times 3)} = \frac{11}{20} = 0.55$$

# CLOSENESS CENTRALITY

**Closeness:** Inverse of the farness, i.e., how close the node is to all other nodes in term of shortest paths.

$$\text{Closeness}(u) = \frac{N-1}{\sum_{v \in V \setminus u} \ell_{u,v}}$$

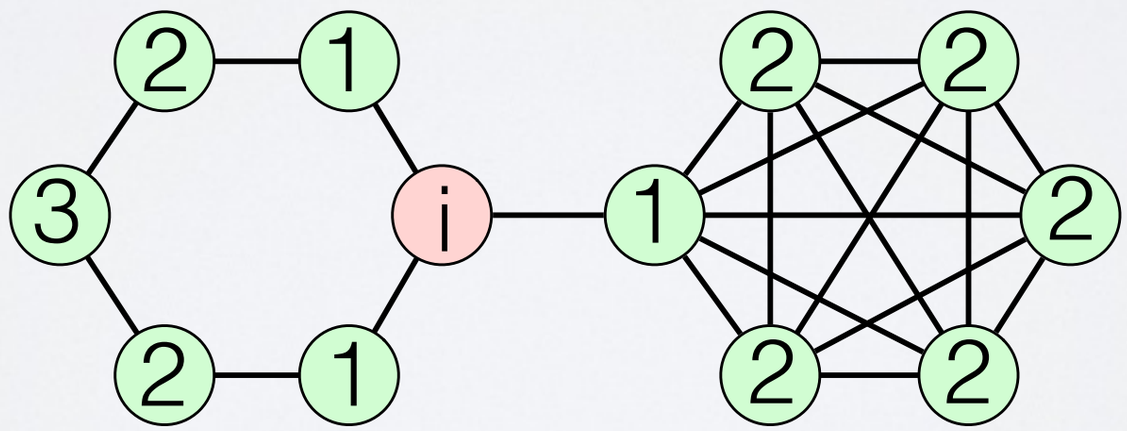
1=all nodes are at distance one



# Harmonic Centrality

**Harmonic centrality:** A variant of the closeness defined as the average of the inverse of distance to all other nodes (Harmonic mean). Well defined on disconnected network with $\frac{1}{\infty} = 0$. Its interpretation is the same as the closeness.

$$\text{Harmonic}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \frac{1}{\ell_{u,v}}$$




$$C_h(i) = \frac{1}{12-1}\left(3 \times \frac{1}{1} + 7 \times \frac{1}{2} + 1 \times \frac{1}{3}\right) = \frac{41}{66} = 0.6212$$

# BETWEENNESS CENTRALITY

- Measure how much the node plays the role of a bridge
- Betweenness of *u: f*raction of all the shortest paths between all the pairs of nodes going through u.

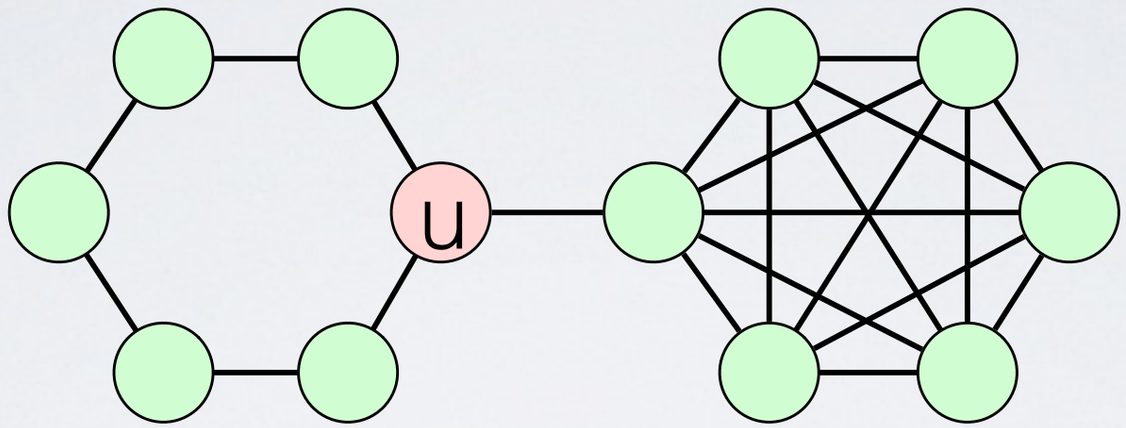
$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

with $\sigma_{st}$ the number of shortest paths between nodes $s$ and $t$ and $\sigma_{st}(v)$ the number of those paths passing through $v$.
The betweenness tends to grow with the network size. A normalized version can be obtained by dividing by the number of pairs of nodes, i.e., for a directed graph: $C_B^{\text{norm}}(v) = \frac{C_B(v)}{(N-1)(N-2)}$.

# Betweenness Centrality

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

directed graph: $C_B^{\text{norm}}(v) = \frac{C_B(v)}{(N-1)(N-2)}$.




$$C_B(u) = 2\frac{5 * 6 + 1 + \frac{1}{2} + \frac{1}{2}}{11 * 10} = \frac{64}{110}$$

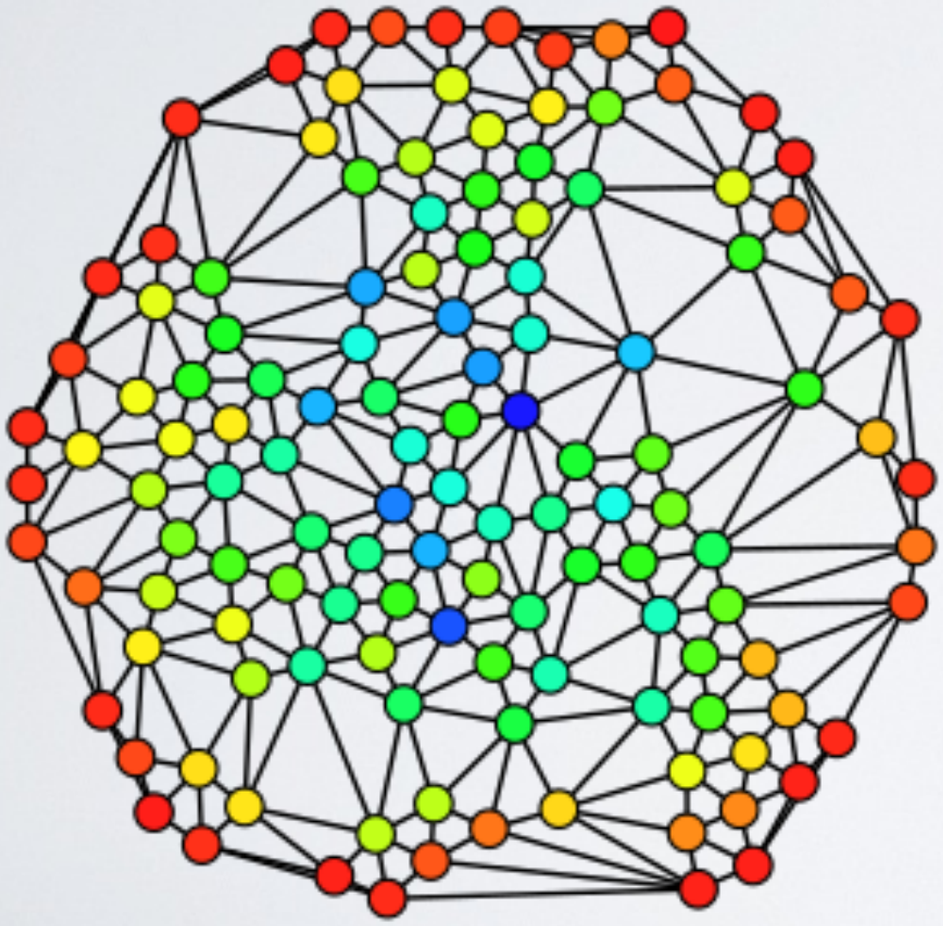
**Exact computation:**

**Floyd-Warshall:** *O(n³) time complexity*
*O(n²) space complexity*

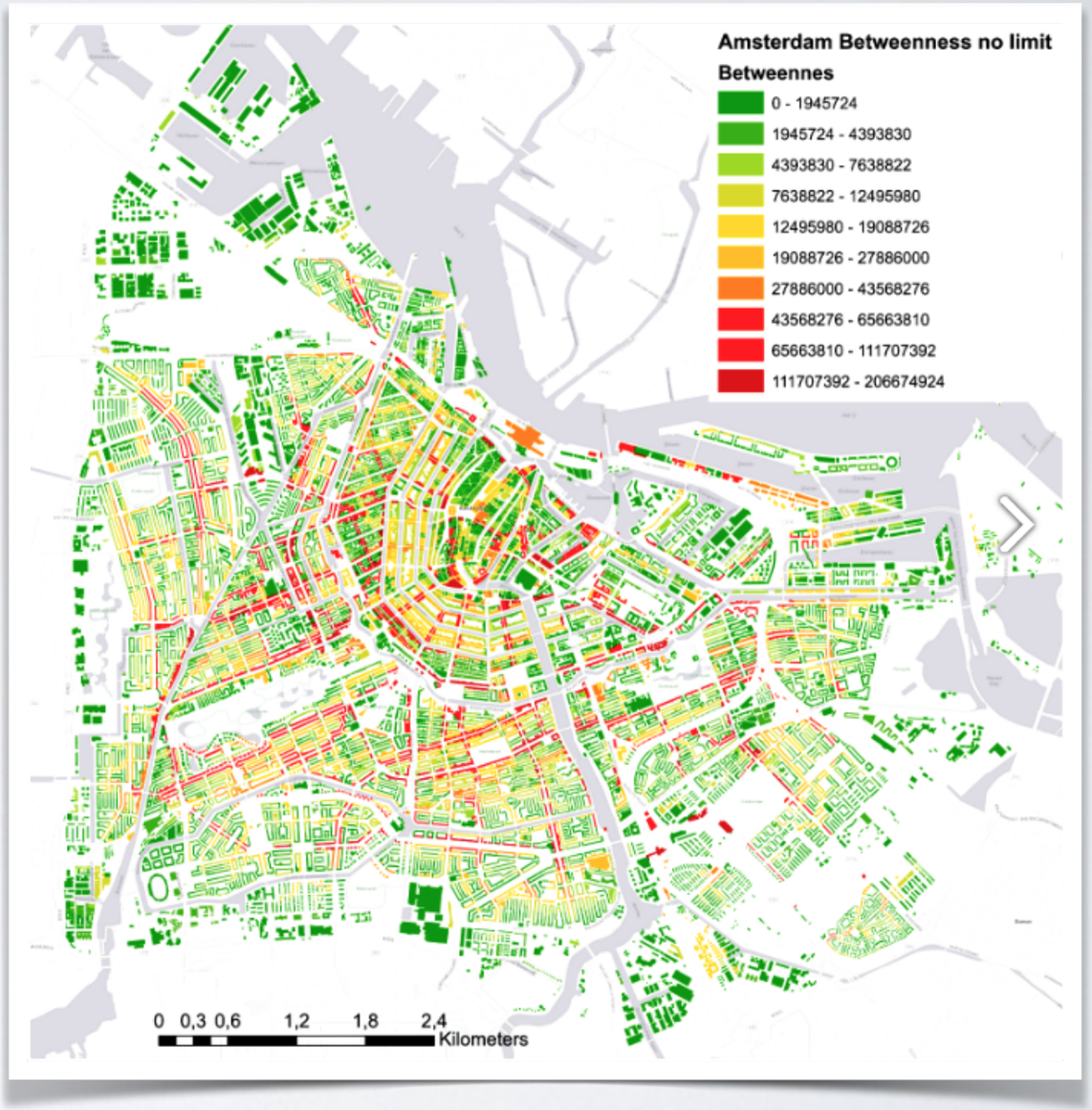
**Approximate computation**

**Dijskstra:** *O(n(m+n log n)) time complexity*

# BETWEENNESS CENTRALITY



(blue higher)




(red higher)

# EDGE - BETWEENNESS

Same definition as for nodes

Can you guess the edge of highest betweenness in the European rail network ?

# RECURSIVE DEFINITIONS

# RECURSIVE DEFINITIONS

- Recursive importance:
  - **Important nodes** are those connected ***to important nodes***
- Several centralities based on this idea:
  - Eigenvector centrality
  - PageRank
  - …

# RECURSIVE DEFINITION

- We would like scores such as :
  - Each node has a score (centrality),
  - If every node "sends" its score to its neighbors, the sum of all scores received by each node will be equal to its original score

$$C_u^{t+1} = \frac{1}{\lambda} \sum_{v \in N_u^{in}} C_v^t \quad (1)$$

- With $\lambda$ a normalisation constant

# RECURSIVE DEFINITION

- This problem can be solved by what is called the *power method:*
  - 1) We initialize all scores to random values
  - 2)Each score is updated according to the desired rule, until reaching a stable point (after normalization)

- Why does it converge?
  - Perron-Frobenius theorem (see next slide)
  - =>True for undirected graphs with a single connected component

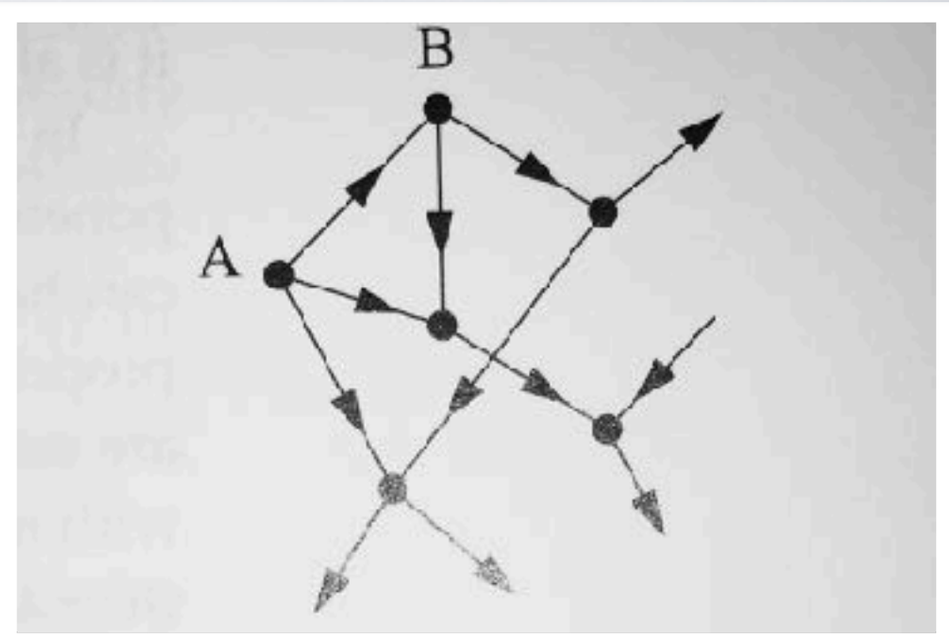
# EIGENVECTOR CENTRALITY

- What we just described is called the Eigenvector centrality
- A couple eigenvector ($x$) and eigenvalue ($\lambda$) is defined by the following relation: $Ax = \lambda x$
  - $x$ is a column vector of size *n*, which can be interpreted as the scores of nodes
- What Perron-Frobenius algorithm says is that the power method will always converge to the *leading eigenvector*, i.e., the eigenvector associated with the highest eigenvalue

# Eigenvector Centrality

**Some problems in case of directed network:**

- Adjacency matrix is asymmetric
- 2 sets of eigenvectors (Left & Right)
- 2 leading eigenvectors
  - Use right eigenvectors : consider nodes that are pointing towards you




**But problem with source nodes (0 in-degree)**

-Vertex A is connected but has only outgoing link = Its centrality will be 0

-Vertex B has outgoing and an incoming link, but incoming link comes from A
= Its centrality will be 0

-etc.

**Solution**: Only in strongly connected component

**Note**: Acyclic networks (citation network) do not have strongly connected component

# PageRank Centrality

- Eigenvector centrality generalised for directed networks

# PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

*Computer Science Department,*
*Stanford University, Stanford, CA 94305, USA*
sergey@cs.stanford.edu and page@cs.stanford.edu

# PageRank Centrality

- Eigenvector centrality generalised for directed networks

# PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

*Computer Science Department,*
*Stanford University, Stanford, CA 94305, USA*
sergey@cs.stanford.edu and page@cs.stanford.edu

**Abstract**

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

# PageRank Centrality

## (Side notes)

-"We chose our system name, Google, because it
is a common spelling of googol, or $10^{100}$ and fits well with our goal of building very large-scale search "

-"[…] at the same time, search engines have migrated from the academic domain to the commercial. **Up until now most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art and to be advertising oriented (see Appendix A). With Google, we have a strong goal to push more development and understanding into the academic realm."**

-"[...], we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers."

# PageRank Centrality

## (Side notes)

**Sergey Brin** received his B.S. degree in mathematics and computer science from the University of Maryland at College Park in 1993. Currently, he is a Ph.D. candidate in computer science at Stanford University where he received his M.S. in 1995. He is a recipient of a National Science Foundation Graduate Fellowship. His research interests include search engines, information extraction from unstructured sources, and data mining of large text collections and scientific data.

**Lawrence Page** was born in East Lansing, Michigan, and received a B.S.E. in Computer Engineering at the University of Michigan Ann Arbor in 1995. He is currently a Ph.D. candidate in Computer Science at Stanford University. Some of his research interests include the link structure of the web, human computer interaction, search engines, scalability of information access interfaces, and personal data mining.

# PAGERANK

- 2 main improvements over eigenvector centrality:
  - In directed networks, problem of source nodes
    - => Add a constant centrality gain for every node
  - Nodes with very high centralities give very high centralities to all their neighbors (even if that is their only in-coming link)
    - => What each node "is worth" is divided equally among its neighbors (normalization by the degree)

$$C_u^{t+1} = \frac{1}{\lambda} \sum_{v \in N_u^{in}} C_v^t \quad \Rightarrow \quad C_u^{t+1} = \alpha \sum_{v \in N_u^{in}} \frac{C_v^t}{k_v^{out}} + \beta$$

With by convention $\beta$=1 and $\alpha$ a parameter (usually 0.85) controlling the relative importance of $\beta$

# PAGERANK

## Matrix interpretation

Principal eigenvector of the "Google Matrix":
First, define matrix S as:

-Normalization by columns of A

-Columns with only 0 receives 1/n

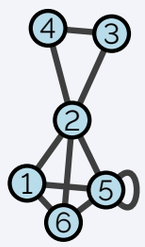
-Finally, $G_{ij} = \alpha S_{ij} + (1 - \alpha)/n$

(a)
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

(c)
$$S = \begin{pmatrix} 0 & 1/2 & 1/3 & 0 & 1/5 \\ 1 & 0 & 1/3 & 1/3 & 1/5 \\ 0 & 1/2 & 0 & 1/3 & 1/5 \\ 0 & 0 & 1/3 & 0 & 1/5 \\ 0 & 0 & 0 & 1/3 & 1/5 \end{pmatrix}$$

(e)
$$G = \begin{pmatrix} 0.03 & 0.455 & 0.313 & 0.03 & 0.2 \\ 0.88 & 0.03 & 0.313 & 0.313 & 0.2 \\ 0.03 & 0.455 & 0.03 & 0.313 & 0.2 \\ 0.03 & 0.03 & 0.313 & 0.03 & 0.2 \\ 0.03 & 0.03 & 0.03 & 0.313 & 0.2 \end{pmatrix}$$

**Graph**




**$A$ - Adjacency Mat.**

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

**Random W. mat.**

$$\begin{pmatrix} 0 & \frac{1}{5} & 0 & 0 & \frac{1}{4} & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{3} \\ 0 & \frac{1}{5} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{5} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{5} & 0 & 0 & \frac{1}{4} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{5} & 0 & 0 & \frac{1}{4} & 0 \end{pmatrix}$$

# PageRank - as Random Walk

**Main idea: The PageRank computation can be interpreted as a Random Walk process with restart**

**Teleportation probability:** the parameter $\alpha$ gives the probability that in the next step of the RW will follow a Markov process or with probability $1$-$\alpha$ it will jump to a random node

**Pagerank** score of a node thus corresponds to the probability of this random walker to be on this node after an infinite number of hops.

# PAGERANK

- Then how do Google rank when we do a research?
- Compute pagerank (using the power method for scalability)
- Create a subgraph of documents related to our topic
- Of course now it is certainly much more complex, but we don't really know:
  "Most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art" [Page, Brin, 1997]

# KATZ CENTRALITY

- Measure of the influence potential of a node

$$C_{\mathrm{Katz}}(u) = \sum_{\ell=1}^{\infty} \sum_{v=1}^{N} \alpha^{\ell} (A^{\ell})_{vu}$$

in which $(A^{\ell})_{vu}$ means the number of paths of length $\ell$ from $v$ to $u$, and $0 < \alpha < \frac{1}{\lambda_{\max}}$ an attenuation parameter smaller than the reciprocal of the largest eigenvalue of $A$, to allow computation in matrix form.

# KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

connected pairs of nodes in distance k

attenuation factor to penalise influence by distance

Katz centrality of node i=

# KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Repeat for all distances as long
As possible (convergence)

# KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Sum for each other node **j**

# KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

$\alpha$ is a parameter in $[0, \frac{1}{\lambda_{\max}}]$.

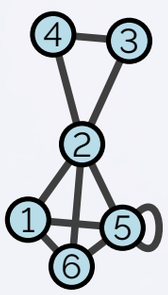
Its strength decreases at
each iteration (increased distance)

# KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Number of different paths from **i** *to* **j** of length k

**Graph**




**$A$ - Adjacency Mat.**

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$A^2$

$$\begin{pmatrix} 3 & 2 & 1 & 1 & 3 & 2 \\ 2 & 5 & 1 & 1 & 3 & 2 \\ 1 & 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 & 1 \\ 3 & 3 & 1 & 1 & 4 & 3 \\ 2 & 2 & 1 & 1 & 3 & 3 \end{pmatrix}$$

# KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Sum of paths to all other nodes at each distance multiplied by a factor decreasing with distance

# KATZ CENTRALITY

## Katz centrality and Eigenvector centrality

Katz centrality can also be understood[a] as a generalization of the eigenvector centrality, with a recursive definition, such as:

$$C_{\text{Katz}}(u) = \alpha \sum_{v \in N_u^{in}} C_{\text{Katz}}(u) + \beta$$

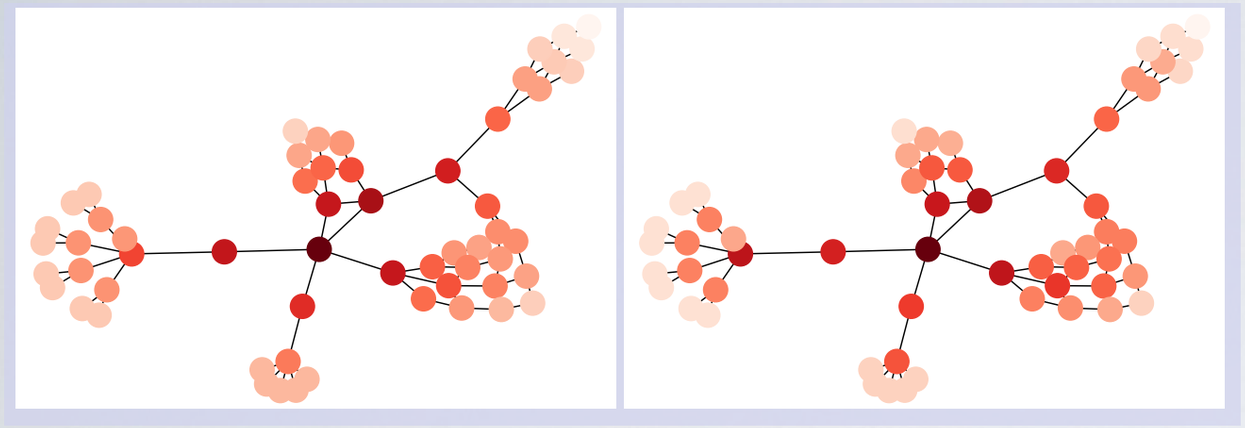
with $\beta = 1$. We can see that with $\beta = 0$ and $\alpha = 1/\lambda_{\max}$ Katz centrality is equivalent to the eigenvector centrality.

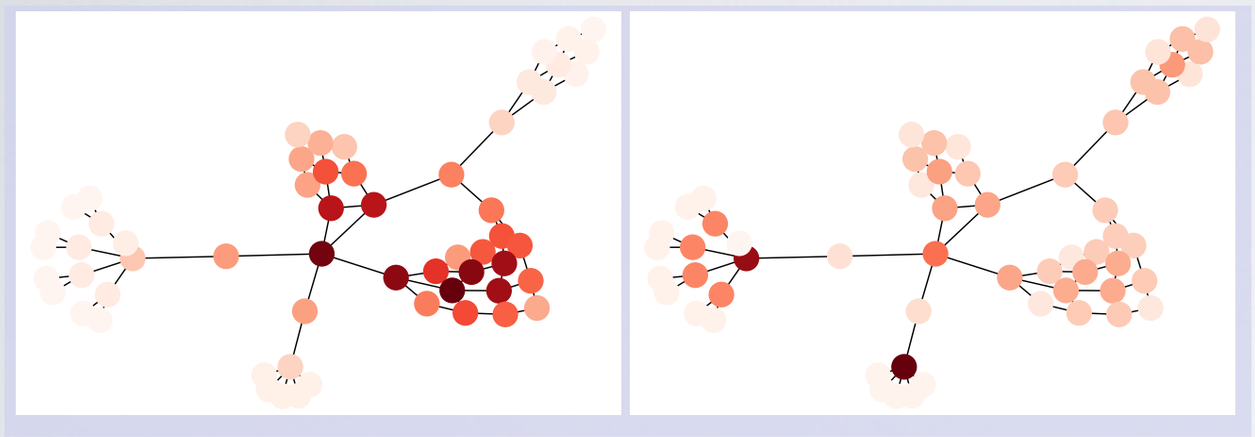
---

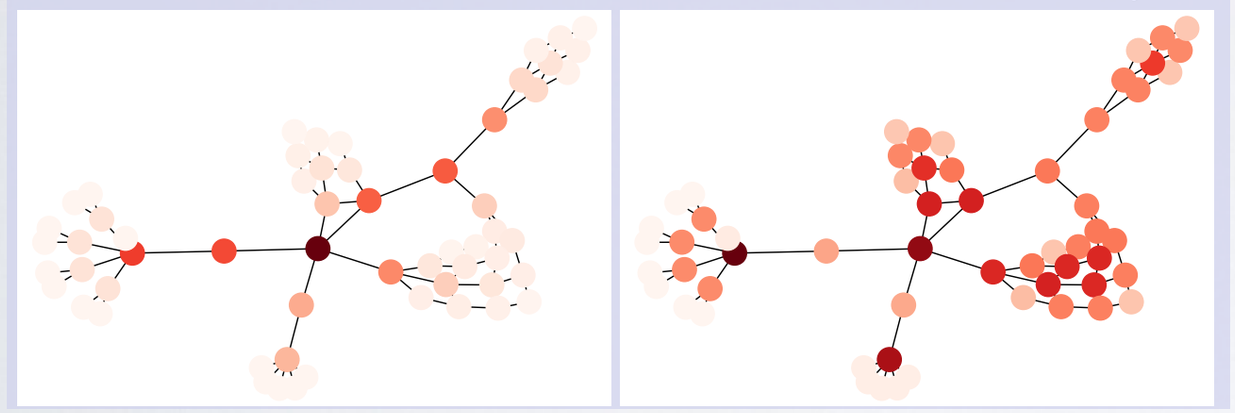
[a] Newman 2018.

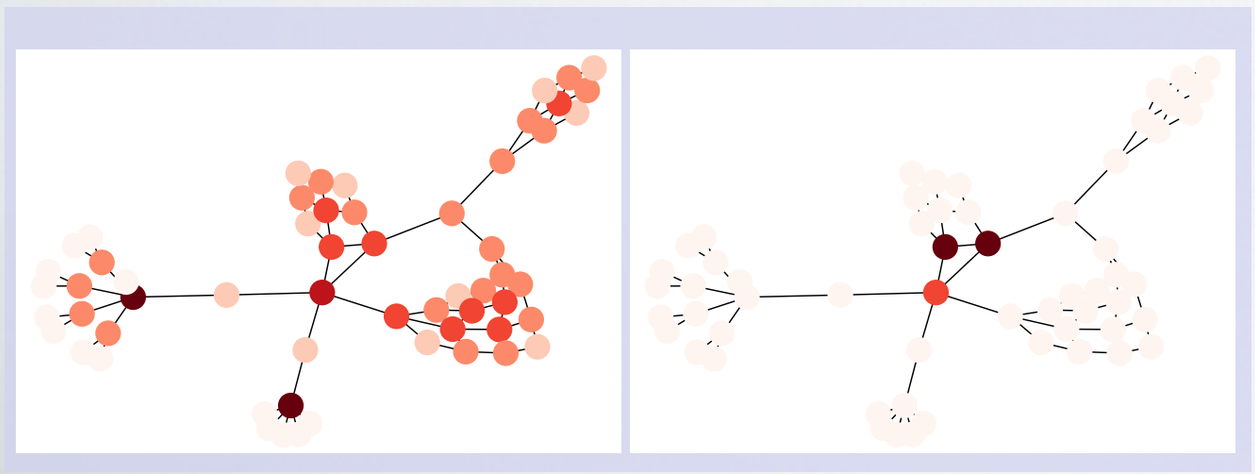
# OTHERS

- Many other centralities have been proposed
- The problem is how to interpret them ?
- Can be used as supervised tool:
  - Compute many centralities on all nodes
  - Learn how to combine them to find chosen nodes
  - Discover new similar nodes
  - (roles in social networks, key elements in an infrastructure, …)





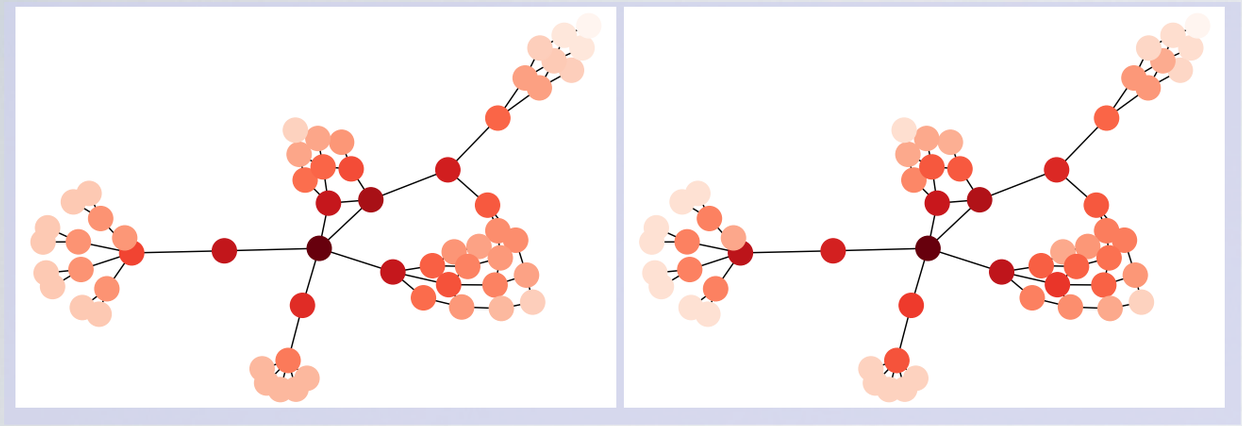


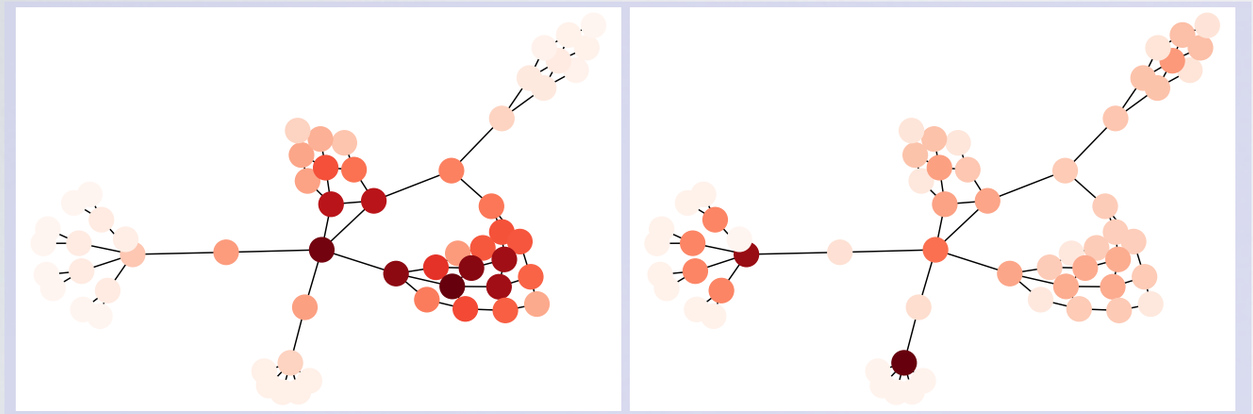


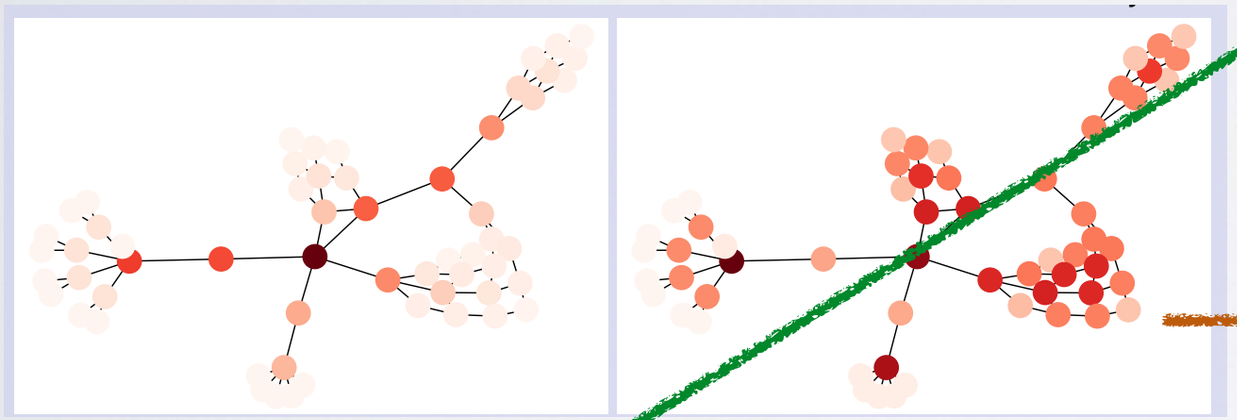
Which is which ?

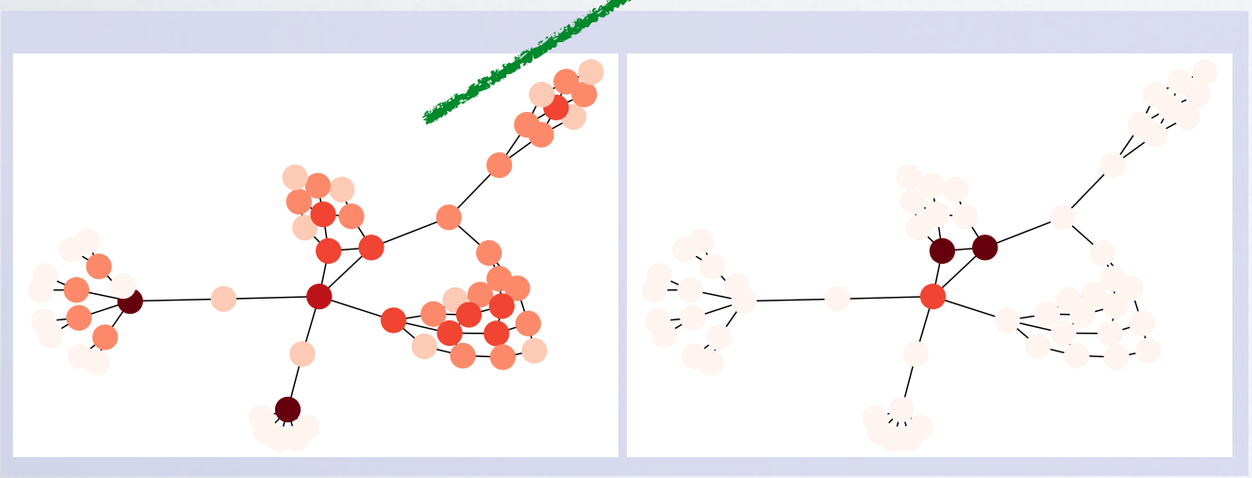
Degree
Clustering coefficient
Closeness

Harmonic Centrality
Betweenness
Katz
Eigenvector
PageRank



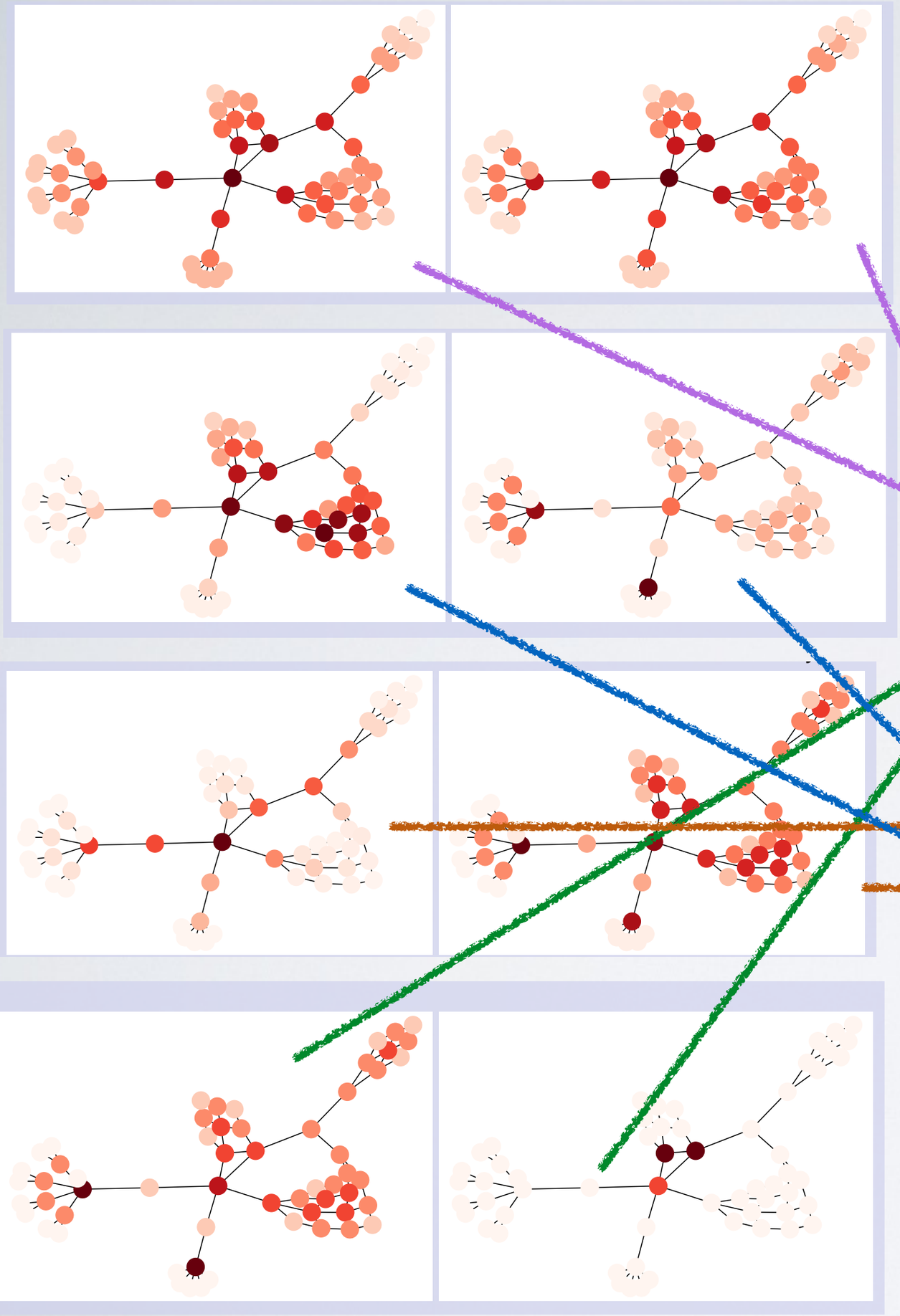






Which is which ?

Degree
Clustering coefficient
Closeness
Harmonic Centrality
Betweenness
Katz
Eigenvector
PageRank



Which is which ?

Degree

Clustering coefficient

Closeness

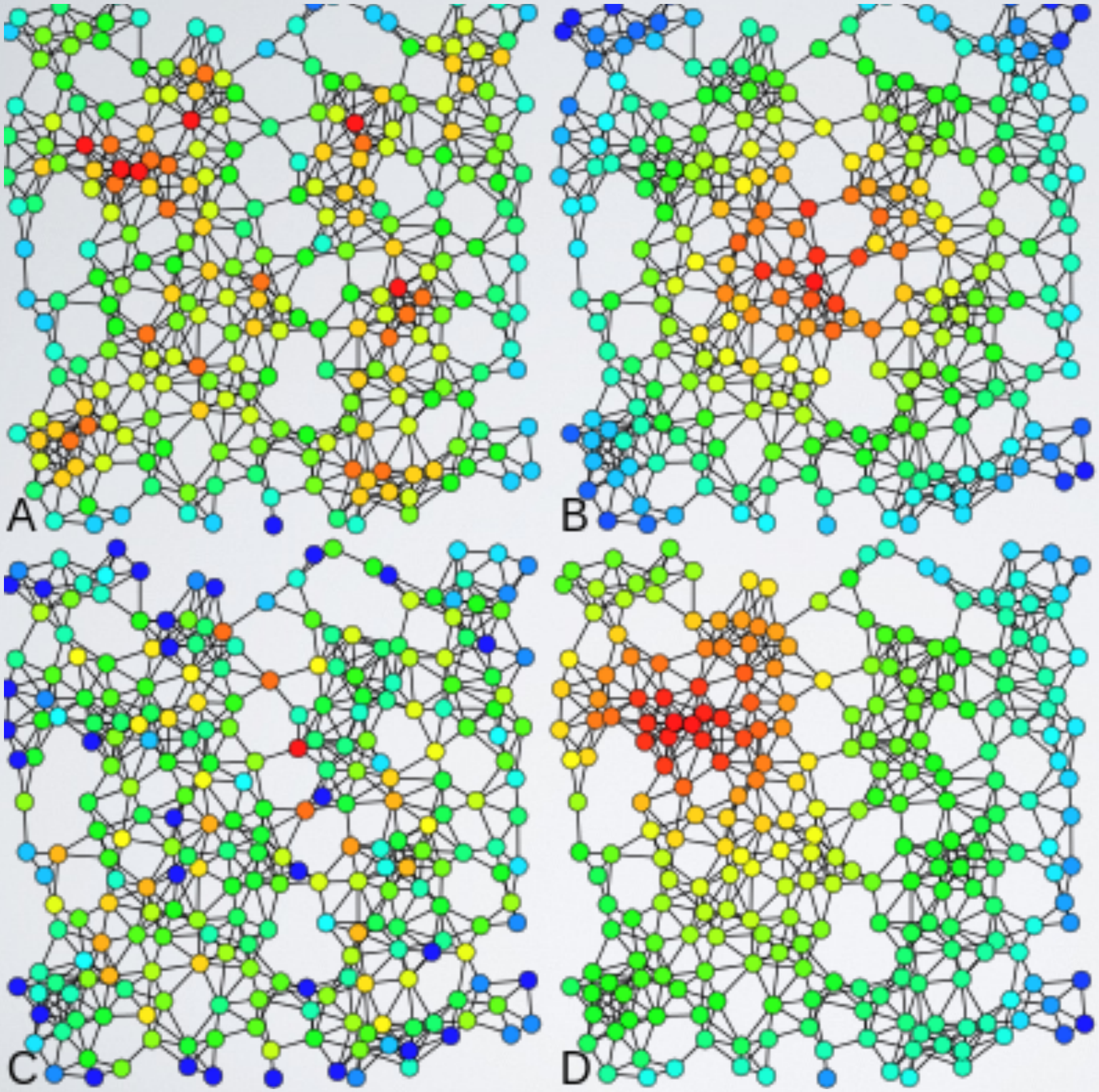
Harmonic Centrality

Betweenness

Katz

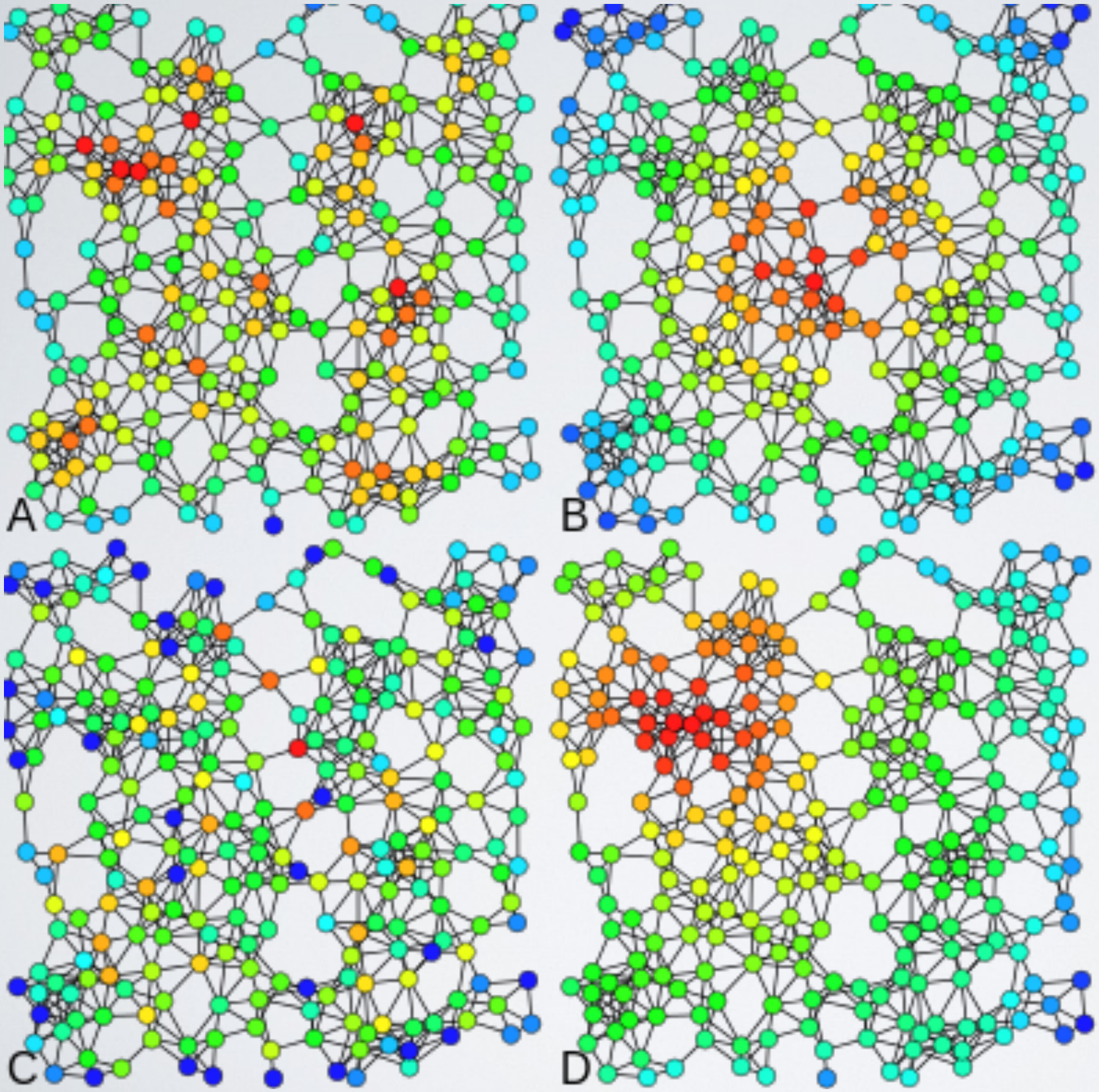
Eigenvector

PageRank




Try again :)

Degree
Betweenness
Closeness
Eigenvector




Try again :)

A: Degree
B:Closeness
C: Betweenness
D: Eigenvector

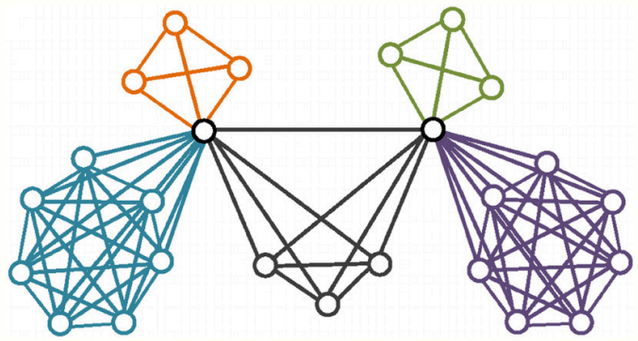
# Caveats of centrality measures

- Each centrality measures is a proxy of an underlying network process
- If this process is irrelevant for the actual network than the centrality measure makes no sense
  - *E.g. If information does not pass via the shortest paths in a network, betweenness centrality is irrelevant*

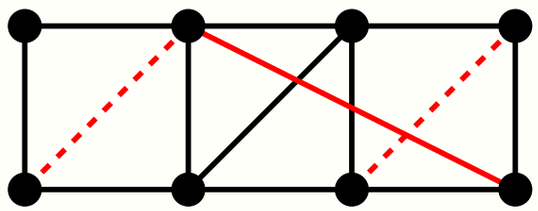
# M2 internships in ENS Lyon

With `christophe.crespelle@ens-lyon.fr` (your 2nd instructor)
Send me an email, we can meet

- On complex network analysis
  - subject 1 : Link communities in complex networks



- On graph algorithms
  - subject 2 : Polynomial kernel for chordal edge deletion



Full desription here (subjects 117, 120 and 128) :

`http://perso.ens-lyon.fr/laurent.lefevre/M2IF/StagesM2/sujets.html`

# Node similarity

**Similarity between nodes based on their neighborhood**

How much two nodes are similarly connected

- What does it mean that they have 3 neighbours in common?
- It is relative to their degree (different meaning for nodes with 3 or 100 neighbours)

➡**Normalisation to penalise nodes with small degrees**

We can define it using existing measures:

- Cosine Similarity
- Pearson Coefficient

# Cosine similarity

Cosine similarity between two non-zero vectors:

$$\cos\theta = \frac{x.y}{|x||y|}$$

Number of common neighbours:

$$n_{ij} = \sum_k A_{ik}A_{kj}$$

Vectors are the rows of adjacency matrix

$$\sigma_{ij} = \cos\theta = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}}$$

with properties for adjacency vectors as

$$A_{i,j} = 0/1 \qquad A_{ij}^2 = A_{ij} \qquad \sum_k A_{ik}^2 = \sum_k A_{ik} = k_i$$

**Cosine similarity:** $$\sigma_{ij} = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{k_i k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

Number of common neighbours normalised by the geometric mean of their degrees

# Pearson coefficient

Correlation between rows of the adjacency matrix

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

cov: covariance, expected product of deviations from individual expected values

$\sigma$: std deviation, square root of the expected squared deviation from the mean

**Intuition, numerator: Number of common neighbours compared to the expected number of common neighbours**

$$\sum_k (A_{ik} - \langle A_i \rangle)(A_{jk} - \langle A_j \rangle) = \sum_k A_{ik} A_{jk} - \frac{k_i k_j}{n}$$

## Properties

- $r(i,j)=0$ - if the number of common neighbours exactly as many as we would expect by chance
- $r(i,j)>0$ - if nodes have more neighbours in common than expected
- $r(i,j)<0$ - if nodes have fewer neighbours in common than expected

# ASSORTATIVITY - HOMOPHILY

# Homophily - Assortative mixing

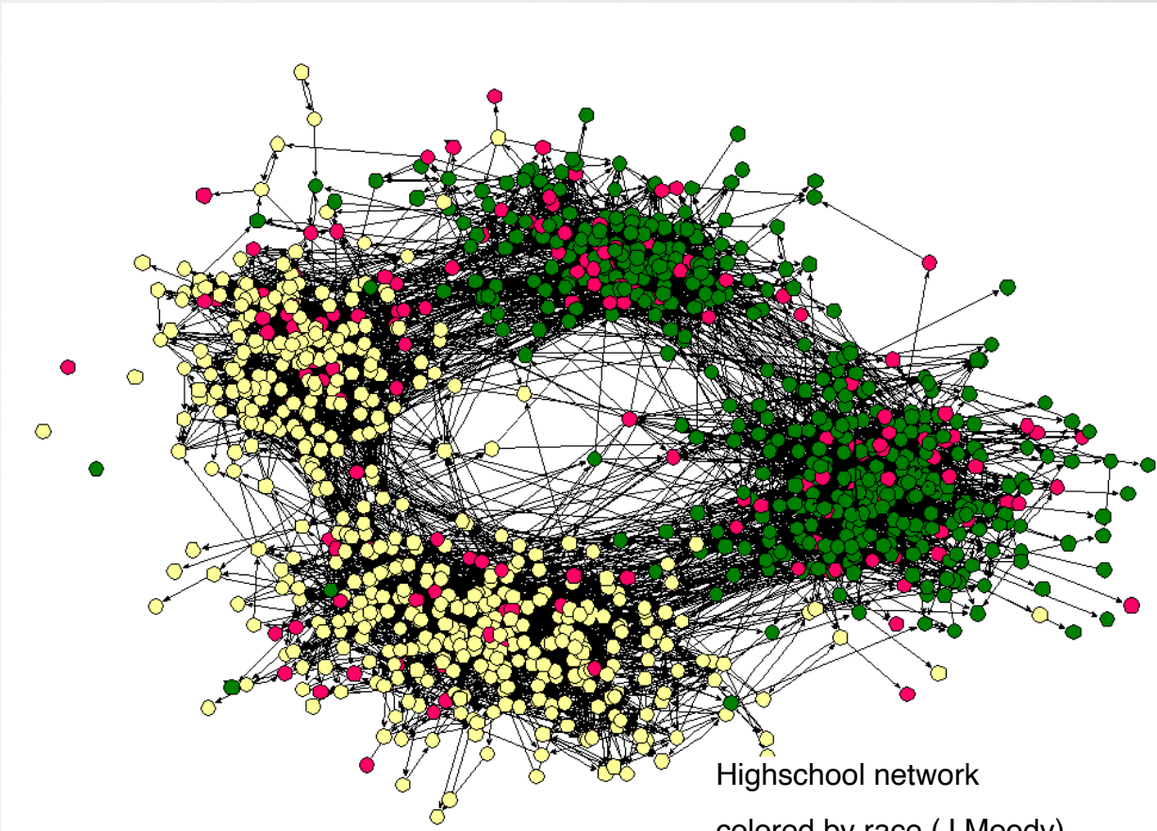
***"birds of a feather flock together"***

- Property of (social) networks that nodes of the same attitude tends to be connected with a higher probability than expected
- It appears as correlation between vertex properties of $x(i)$ and $x(j)$ if $(i,j) \in E$

**Vertex properties**

- age
- gender
- nationality
- political beliefs
- socioeconomic status
- habitual place
- obesity
- …

- Homophily can be a link creation mechanism or consequence of social influence (and it is difficult to distinguish)



Highschool network

colored by race (J Moody)

? Connected people of the same political opinion are connected because they were a priori similar (homophily) or they become similar after they become connected (social influence)?

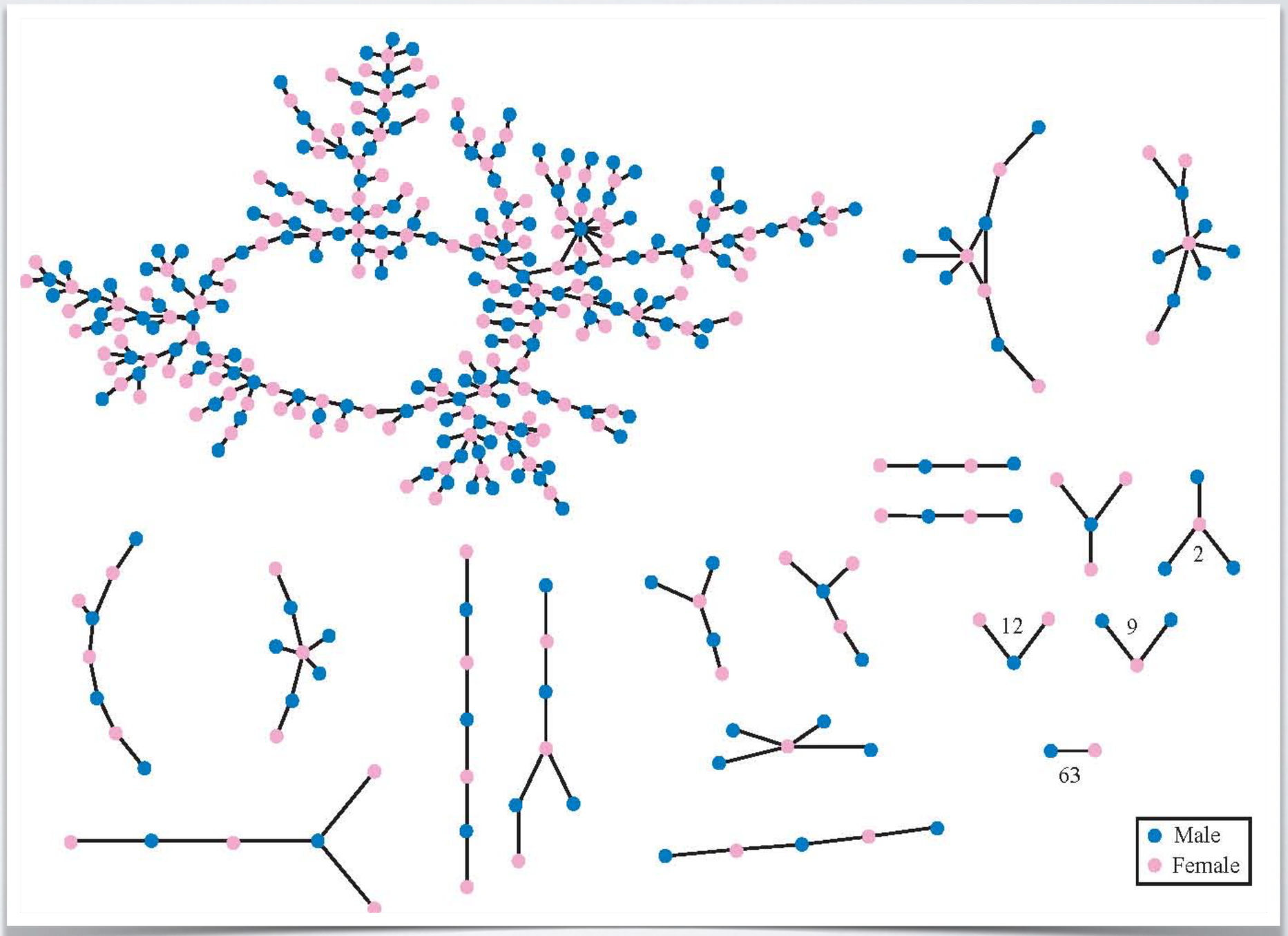
# Homophily - Assortative mixing

**Dissasortative mixing**

- Contrary of homophily, where dissimilar nodes are tend to be connected

**Examples**

- Sexual networks
- Predator - prey ecological networks



# Homophily - Assortative mixing

## To quantify homophily

| | | women | | | | |
|---|---|---|---|---|---|---|
| | | black | hispanic | white | other | $a_i$ |
| men | black | 0.258 | 0.016 | 0.035 | 0.013 | 0.323 |
| | hispanic | 0.012 | 0.157 | 0.058 | 0.019 | 0.247 |
| | white | 0.013 | 0.023 | 0.306 | 0.035 | 0.377 |
| | other | 0.005 | 0.007 | 0.024 | 0.016 | 0.053 |
| | $b_i$ | 0.289 | 0.204 | 0.423 | 0.084 | |

TABLE I: The mixing matrix $e_{ij}$ and the values of $a_i$ and $b_i$ for sexual partnerships in the study of Catania *et al.* [23]. After Morris [24].

## Discrete properties

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

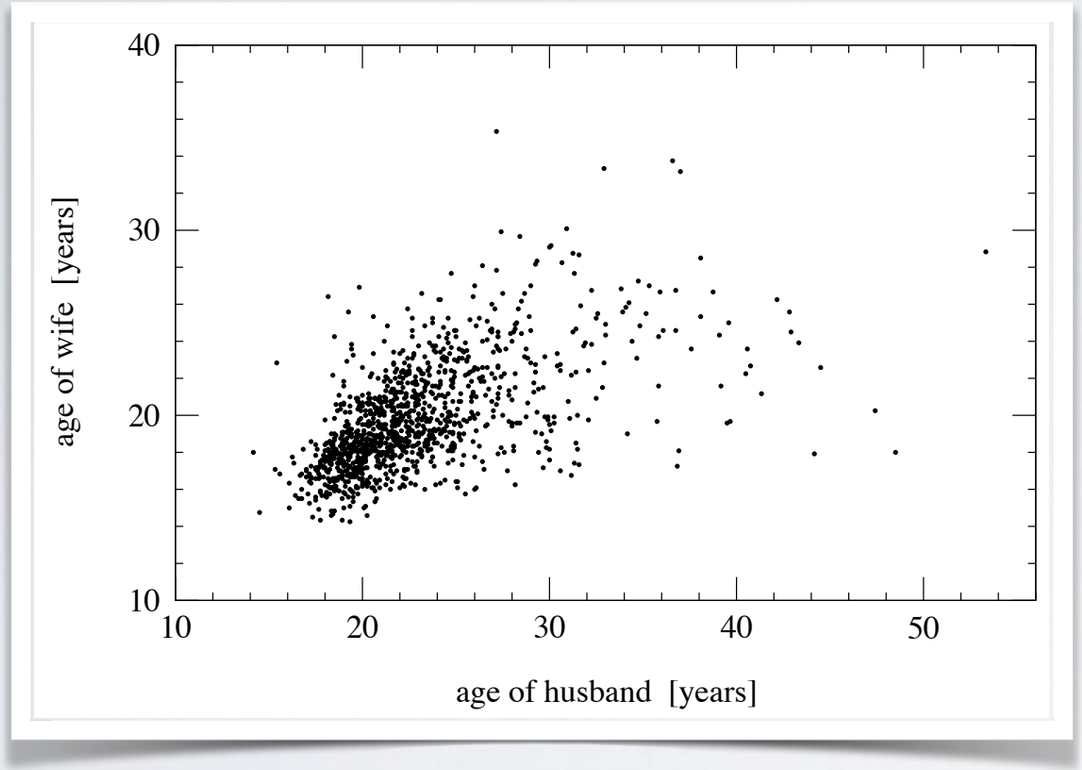
No assortative mixing : r=0 ($e_{ij} = a_i b_j$)

Perfectly assortative: r=1

Perfectly disassortative: -1<r<0

# Homophily - Assortative mixing

## To quantify homophily




## Scalar properties

Pearson correlation coefficient of properties at both extremities of edges

$e_{xy}$: fraction of edges joining nodes with values x and y

$$\sum_{xy} e_{xy} = 1, \qquad \sum_{y} e_{xy} = a_x, \qquad \sum_{x} e_{xy} = b_y$$

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b},$$

with $\sigma_a$ standard deviation of $a_x$

r=0, no assortative mixing,
r>0 assortative mixing,
r<0 disassortative mixing

# Degree-degree correlation

- A particular type of application is the degree correlation:
  - Are *important nodes* connected to other important nodes with a higher probability than expected?
  - The degree can be used as any other scalar property

| | network | type | size $n$ | assortativity $r$ | error $\sigma_r$ |
|---|---|---|---|---|---|
| social | physics coauthorship | undirected | 52 909 | 0.363 | 0.002 |
| | biology coauthorship | undirected | 1 520 251 | 0.127 | 0.0004 |
| | mathematics coauthorship | undirected | 253 339 | 0.120 | 0.002 |
| | film actor collaborations | undirected | 449 913 | 0.208 | 0.0002 |
| | company directors | undirected | 7 673 | 0.276 | 0.004 |
| | student relationships | undirected | 573 | −0.029 | 0.037 |
| | email address books | directed | 16 881 | 0.092 | 0.004 |
| technological | power grid | undirected | 4 941 | −0.003 | 0.013 |
| | Internet | undirected | 10 697 | −0.189 | 0.002 |
| | World-Wide Web | directed | 269 504 | −0.067 | 0.0002 |
| | software dependencies | directed | 3 162 | −0.016 | 0.020 |
| biological | protein interactions | undirected | 2 115 | −0.156 | 0.010 |
| | metabolic network | undirected | 765 | −0.240 | 0.007 |
| | neural network | directed | 307 | −0.226 | 0.016 |
| | marine food web | directed | 134 | −0.263 | 0.037 |
| | freshwater food web | directed | 92 | −0.326 | 0.031 |

M. E. J. Newman, Phys. Rev. Lett. **89,** 208701 (2002)

# Average nearest-neighbour degree

R. Pastor-Satorras, A. Vázquez, A. Vespignani, Phys. Rev. E 65, 066130 (2001)

- More detailed characterisation of degree-degree correlations
- $k_{annd}$: **a**verage **n**earest **n**eighbours **d**egree
- $k_{annd}$ can be written as:

$$k_{annd}(k) = \sum_{k'} k' P(k' \mid k) = \frac{\sum_{k'} k' e_{kk'}}{\sum_{k'} e_{kk'}}$$

  - where $P(k'|k)$ is the conditional probability that an edge of a node with degree $k$ points to a node with degree $k'$

- If there are no degree correlations:

$$k_{annd}(k) = \ldots = \frac{\langle k^2 \rangle}{\langle k \rangle}$$

  - $k_{annd}$ is independent of $k$ (nodes of any degrees should have the same nearest neighbors degree)

- If the network is assortative $k_{nn}(k)$ is a positive function
- If the network is disassortative $k_{nn}(k)$ is a negative function