COMPLEX NETWORKS

Centrality measures

NODE

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

- Bad term: nothing to do with being central in general

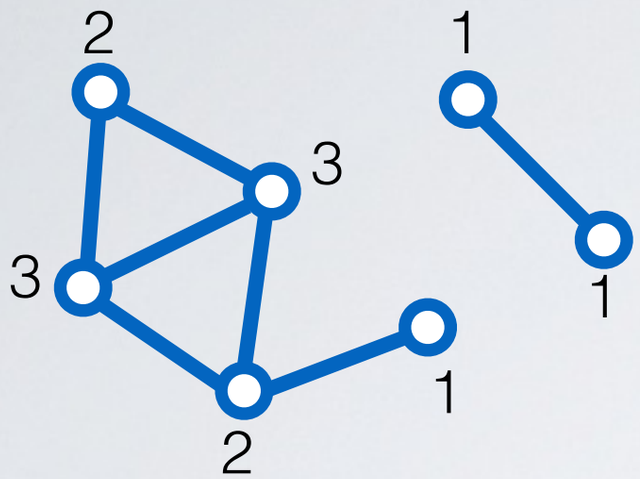
- Usage:
  - Some centralities have straightforward interpretation
  - Centralities can be used as *node features* for machine learning on graph
    - (Classification, link prediction, …)

Connectivity based centrality measures

Degree centrality - recap

Number of connections of a node

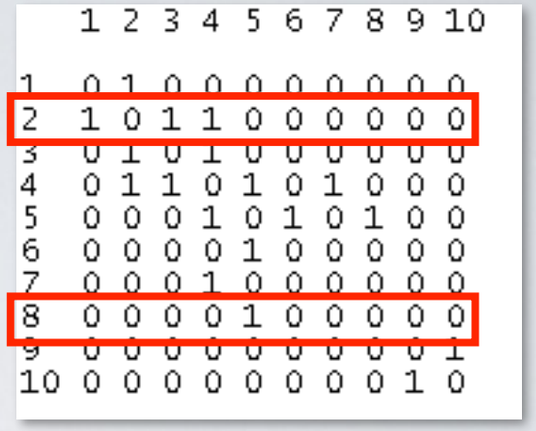
- Undirected network



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

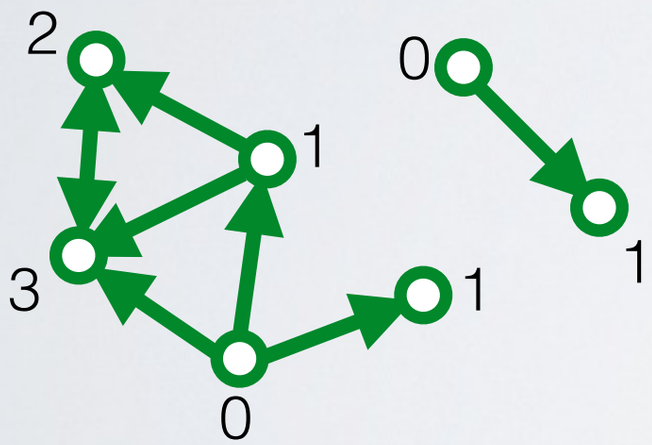
$$m = \frac{\sum_i k_i}{2} \qquad \text{where} \qquad m = |E|$$

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



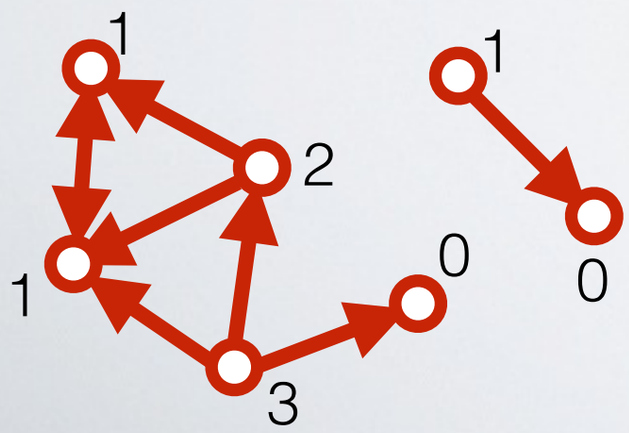
	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

- Directed network



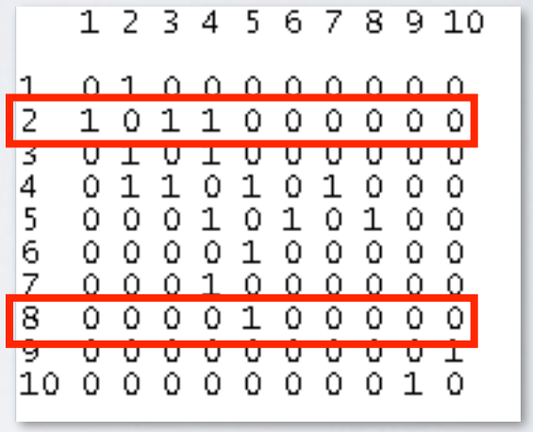
In degree $$k_i^{in} = \sum_j^N A_{ij}$$

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

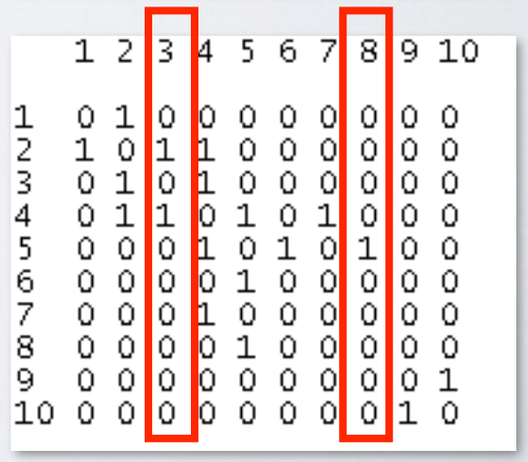


$$m = \sum_i^N k_i^{in} = \sum_j^N k_j^{out} = \sum_{ij} A_{ij}$$

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



	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



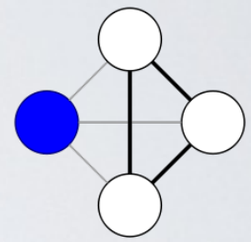
	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

NODE DEGREE

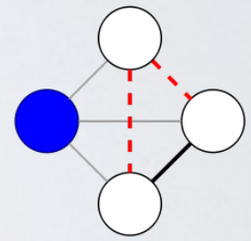
- **Degree**: how many neighbors
- Often enough to find important nodes
  - Main characters of a series talk with the more people
  - Largest airports have the most connections
  - …
- But not always
  - Facebook users with the most friends are spam
  - Webpages/wikipedia pages with most links are simple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

NODE CLUSTERING COEFFICIENT

- **Clustering coefficient**: closed triangles/triads
- Tells you if the neighbors of the node are connected

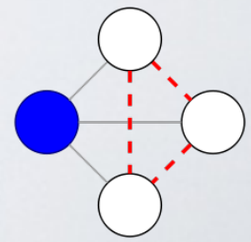


c = 1



c = 1/3

- Be careful!
  - Degree 2: value 0 or 1
  - Degree 1000: Not 0 or 1 (usually)
  - Ranking them is not meaningful
- Can be 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

RECURSIVE DEFINITIONS

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

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

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

x_i is the centrality of node i.

$A_{ij} = 1$ if there is an edge, 0 otherwise

RECURSIVE DEFINITION

- This problem can be solved by what is called the *power method:*

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

  - 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 for *real* and *irreducible square matrices* with *non-negative* entries
  - =>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 (λ) is defined by the following relation: $Ax = \lambda x$
  - x is a vector of size *n*, which can be interpreted as the scores of nodes
  - Ax yield a new vector of size *n*, which corresponds for each node to receive the sum of the scores of its neighbors (like in the power method)
  - The equality means that the new scores are proportional to the previous scores
- 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

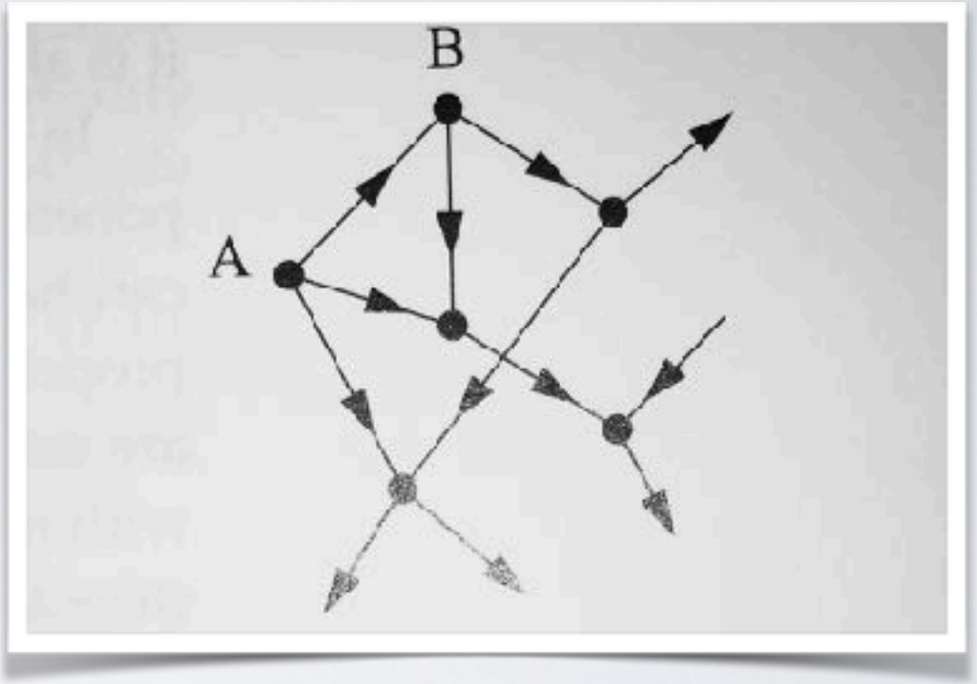
normalized=>

	leading eig	random	step 1	step 2	step 3	step 4	step 5	step 6	step 7	step 8	step 9	step 10
0	0.355491	0.935089	0.355491	0.355491	0.355491	0.355491	0.355491	0.355491	0.355491	0.355491	0.355491	0.355491
1	0.265960	0.085427	0.240626	0.226817	0.257197	0.261238	0.262913	0.265743	0.264947	0.266126	0.265601	0.266067
2	0.317193	0.946762	0.198805	0.318385	0.289034	0.319438	0.308991	0.318290	0.314404	0.317583	0.316102	0.317300
3	0.211180	0.406925	0.141819	0.207191	0.193548	0.212811	0.206461	0.212442	0.209759	0.211824	0.210693	0.211481
4	0.075969	0.478530	0.080332	0.105676	0.077633	0.082889	0.075815	0.077905	0.075729	0.076621	0.075844	0.076224
5	0.079483	0.807562	0.088545	0.116411	0.082726	0.087845	0.079672	0.081756	0.079303	0.080228	0.079368	0.079770
6	0.079483	0.461360	0.108581	0.108909	0.084797	0.087185	0.079863	0.081698	0.079320	0.080222	0.079369	0.079769
7	0.170960	0.414581	0.107283	0.175380	0.152884	0.174590	0.165953	0.172741	0.169381	0.171774	0.170394	0.171326
8	0.227404	0.172052	0.177318	0.223427	0.209979	0.230078	0.220224	0.228890	0.224296	0.227976	0.225992	0.227586
9	0.102674	0.100008	0.079303	0.095221	0.093842	0.102635	0.099201	0.103035	0.101167	0.102854	0.101977	0.102733
10	0.075969	0.381324	0.100368	0.098174	0.079704	0.082229	0.076006	0.077847	0.075746	0.076616	0.075845	0.076224

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)

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)} \quad \Rightarrow \quad x_i = \alpha \sum_{j=1}^{n} A_{ij} \frac{x_j}{k_j^{\text{out}}} + \beta$$

With by convention β=1 and α a parameter (usually 0.85)

PAGERANK

Intuition

$$x_i = \alpha \sum_{j=1}^{n} A_{ij} \frac{x_j}{k_j^{\text{out}}} + \beta$$

Left term dominates when nodes have many neighbors.
Right term dominate with few neighbors

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

PageRank - as Random Walk

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

- In the Google Matrix, Elements in each row are summing up to 1
- It is a stochastic matrix which can be interpreted as a stationary transition matrix of a random walk process

- Probability that the RW will be in node i next step depends only on the current node j and the transition probability $j \rightarrow i$ determined by the stochastic matrix
- Consequently this is a **first-order Markov process**
- **Stationary probabilities** (i.e., when walk length tends towards ∞) of the RW to be in node i gives the PageRank of the actual node

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

- If $\alpha<1$, it assures that the RW will never be stuck at nodes with $k^{out}=0$, but it can restart the RW from a randomly selected other node

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

It measures the relative degree of influence of a node within a network

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

connected pairs of nodes in distance k

attenuation factor to penalise influence by distance

- Attenuation factor α must be smaller than $1/|\lambda_0|$, i.e. the reciprocal of the absolute value of the largest eigenvalue of $\mathbf{A}$.

Matrix form:

$$\vec{C}_{Katz} = ((I - \alpha A^T)^{-1} - I)\vec{I}$$

- where I is the identity matrix, and $\vec{I}$ is the identity vector
- Katz centrality is useful for directed networks (citation nets, WWW) where Eigenvector centrality fails

KATZ CENTRALITY

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

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 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.
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

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

It measures the relative degree of influence of a node within a network

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

connected pairs of nodes in distance k

attenuation factor to penalise influence by distance

- Attenuation factor α must be smaller than $1/|\lambda_0|$, i.e. the reciprocal of the absolute value of the largest eigenvalue of $\mathbf{A}$.

Matrix form:

$$\vec{C}_{Katz} = ((I - \alpha A^T)^{-1} - I)\vec{I}$$

- where I is the identity matrix, and $\vec{I}$ is the identity vector
- Katz centrality is useful for directed networks (citation nets, WWW) where Eigenvector centrality fails

Geometric centrality measures

CLOSENESS CENTRALITY

$$C_{cl}(i) = \frac{n-1}{\sum_{d_{ij}<\infty} d_{ij}}$$

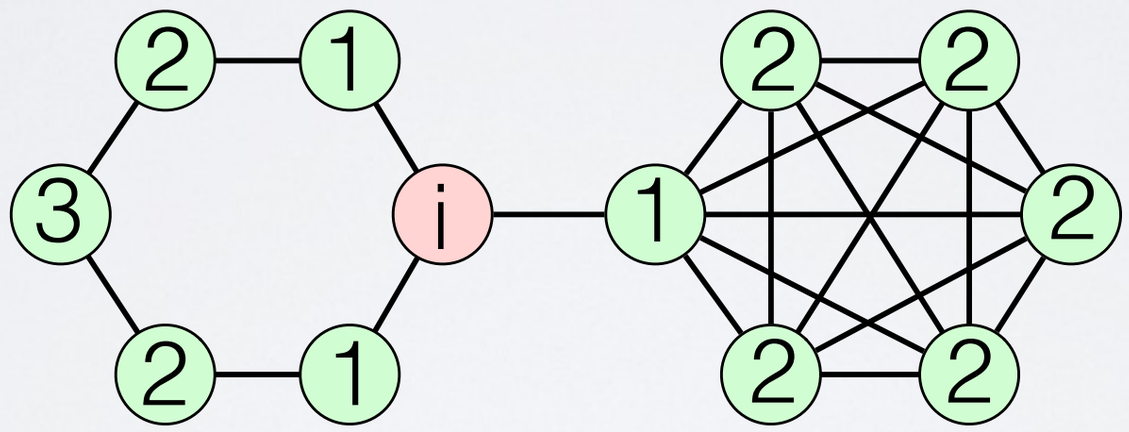
- **Farness**: average of length of shortest paths to all other nodes.

- **Closeness**: inverse of the Farness (normalized by number of nodes)
  - Highest closeness = More central
  - Closness= 1: directly connected to all other nodes

- Well defined only on connected networks

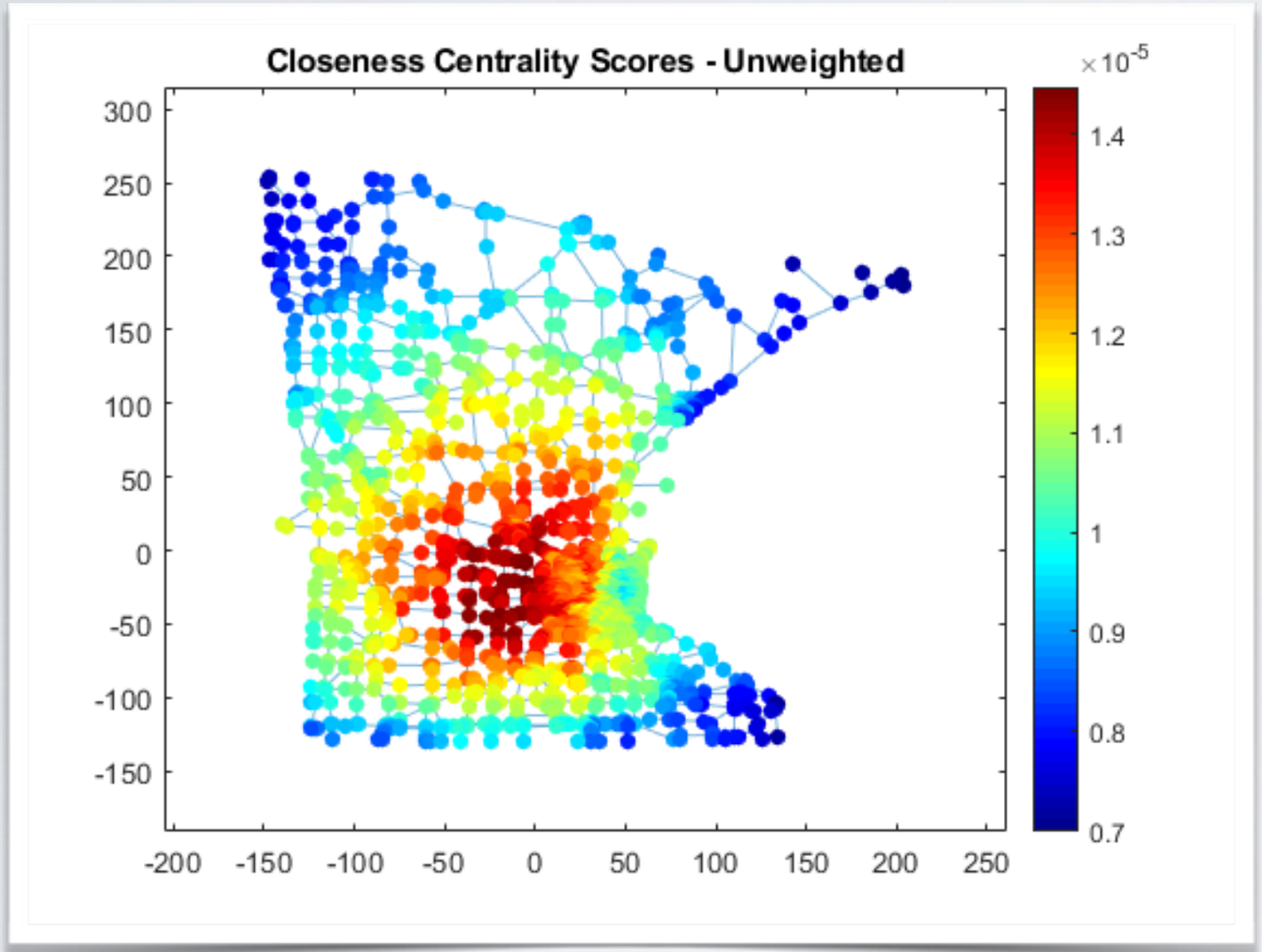
CLOSENESS CENTRALITY

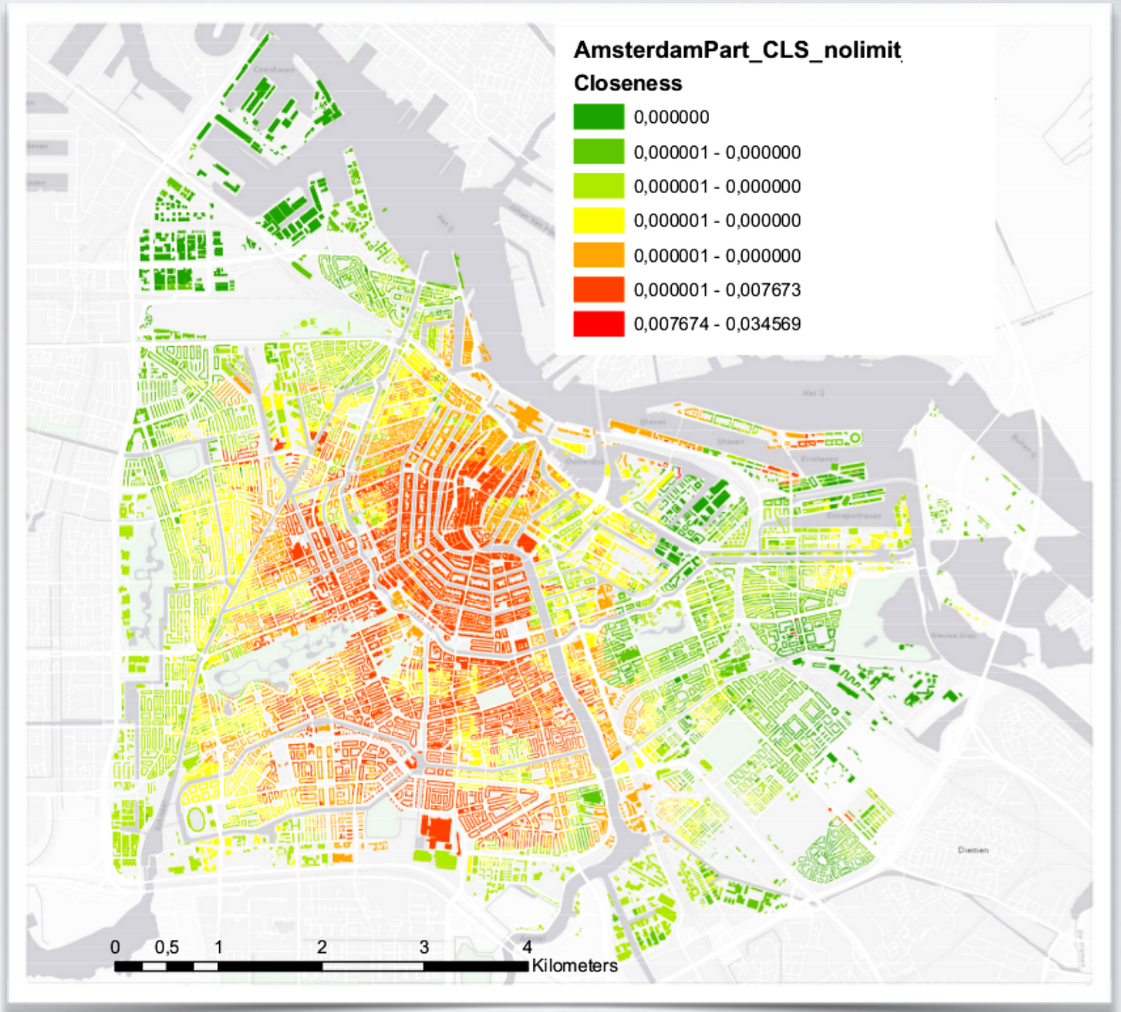
$$C_{cl}(i) = \frac{n-1}{\sum_{d_{ij}<\infty} d_{ij}}$$



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

CLOSENESS CENTRALITY



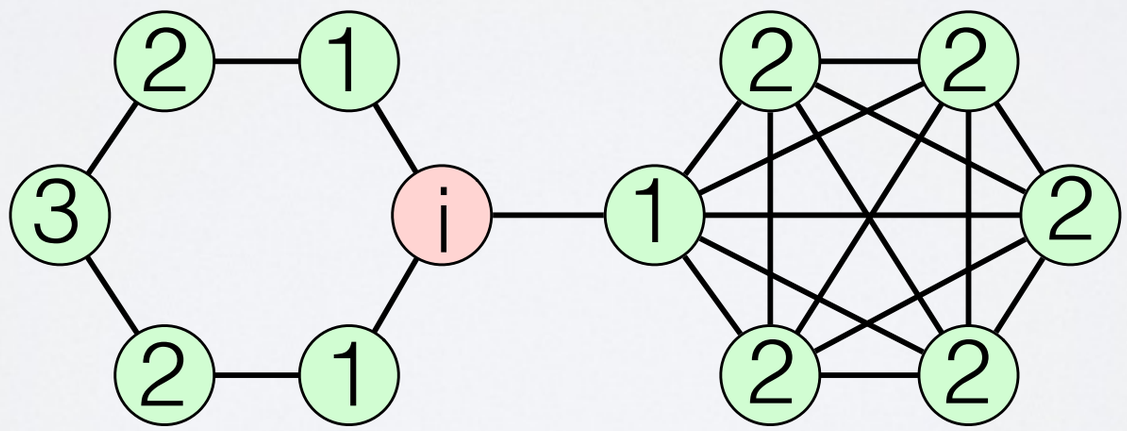


Harmonic Centrality

- Harmonic mean of the geodesic (shorted paths) distances from a given node to all others

$$C_h(i) = \frac{1}{n-1} \sum_{i \neq j} \frac{1}{d_{ij}}$$

- In case of no path between nodes i and j: $d_{ij} = \infty$
- Well **defined on disconnected networks**



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

Betweenness Centrality

Assumption: important vertices are bridges over which information flows

Practically: if information spreads via shortest paths, important nodes are found on many shortest paths

Notation: $\sigma_{jk}(i) =$ number of geodesic path from j to k via i: $j \to \ldots \to i \to \ldots \to k$

$\sigma_{jk} =$ number of geodesic path from j to k: $j \to \ldots \to k$

Definition:

$$C_b(i) = \sum_{j \neq k} \frac{\#\{\text{geodesic path: } j \to \ldots \to i \to \ldots \to k\}}{\#\{\text{geodesic path: } j \to \ldots \to k\}} = \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

Normalised definition:

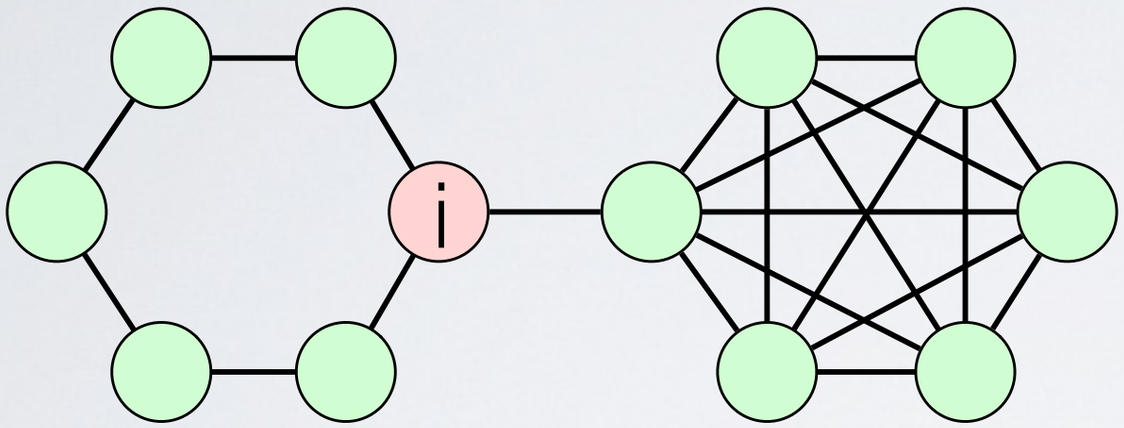
$$C_b(i) = \frac{1}{n^2} \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \qquad \text{where} \qquad C_b \in [0,1]$$

Total number of ordered vertex pairs

Betweenness Centrality

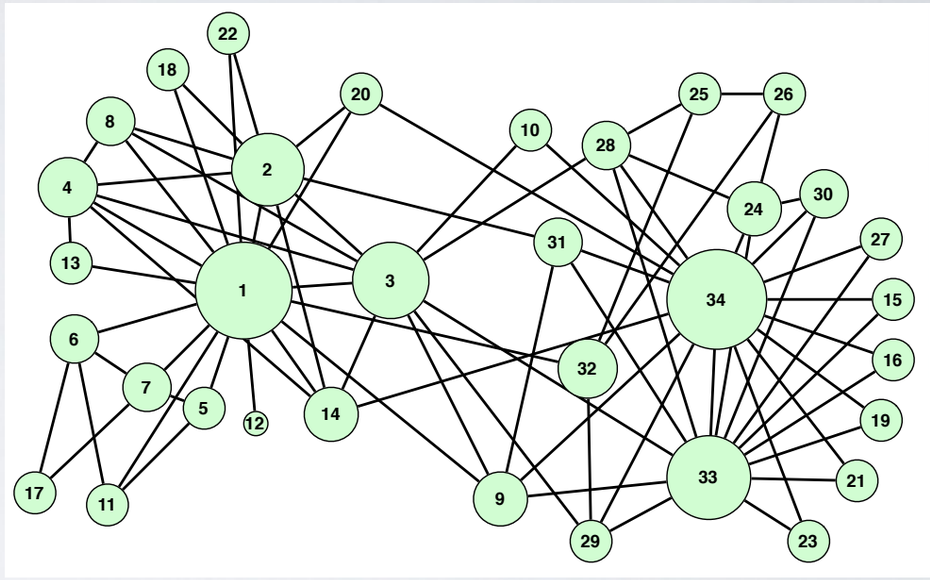
$$C_b(i) = \frac{1}{n^2} \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

where $C_b \in [0,1]$

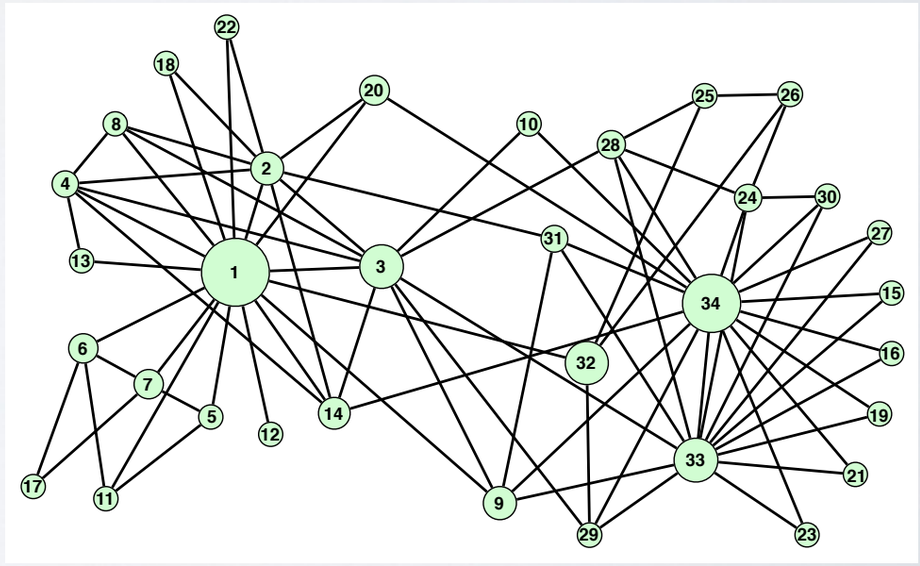


$$C_b(i) = \frac{78}{144}$$

Zachary's karate club network



degree



betweenness

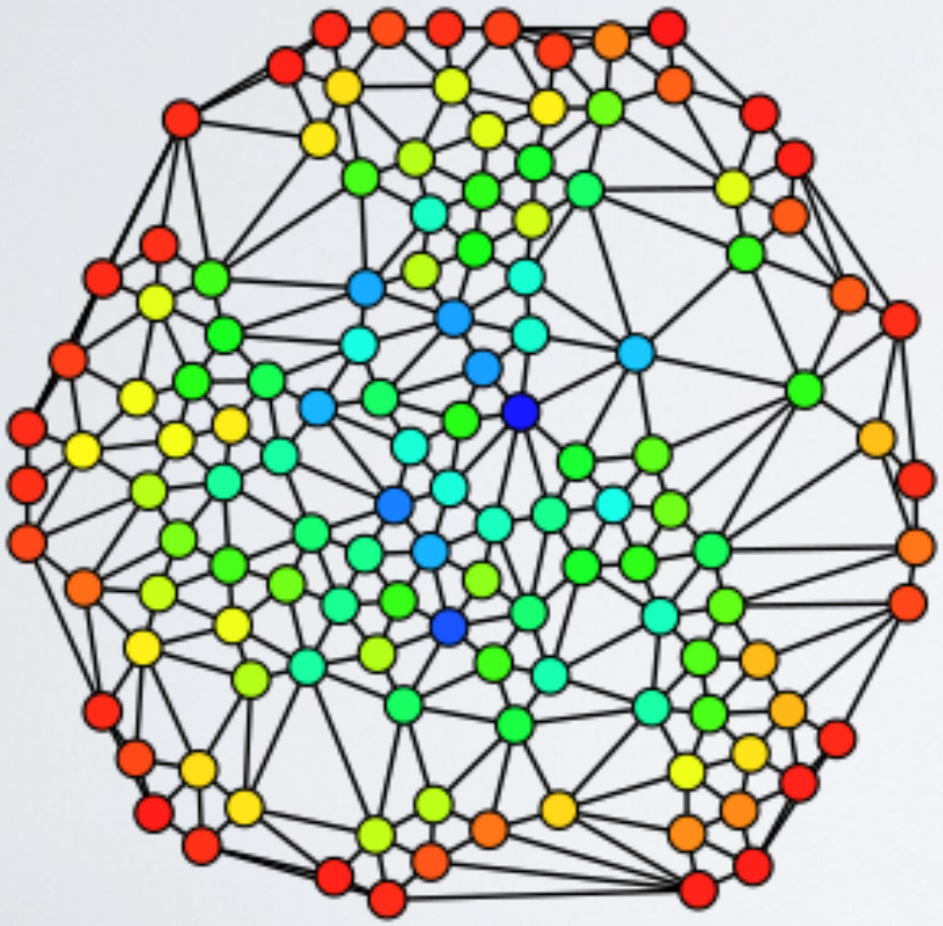
Exact computation:

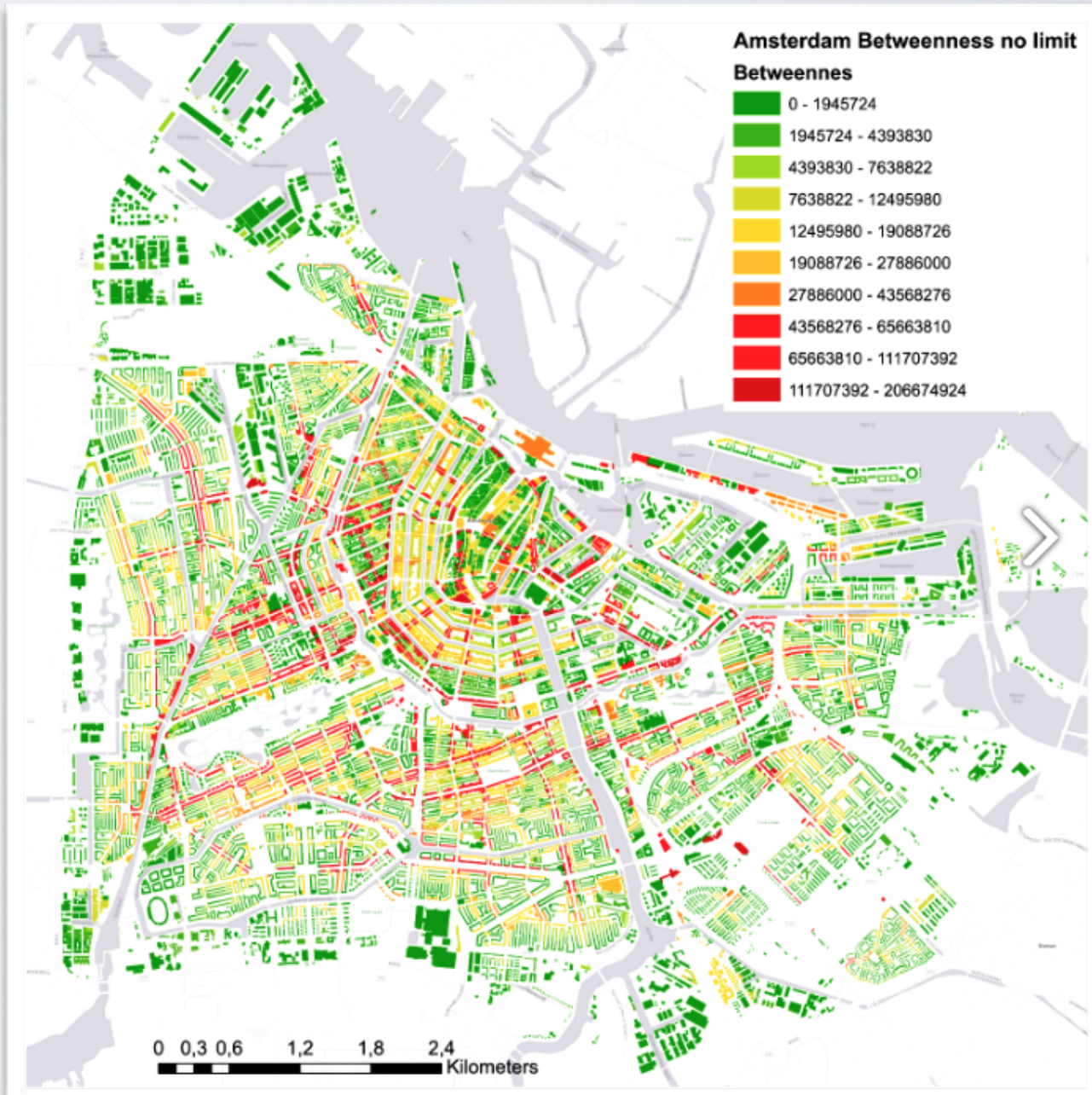
Floyd-Warshall: *$O(n^3)$ time complexity*
$O(n^2)$ space complexity

Approximate computation

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

BETWEENNESS CENTRALITY



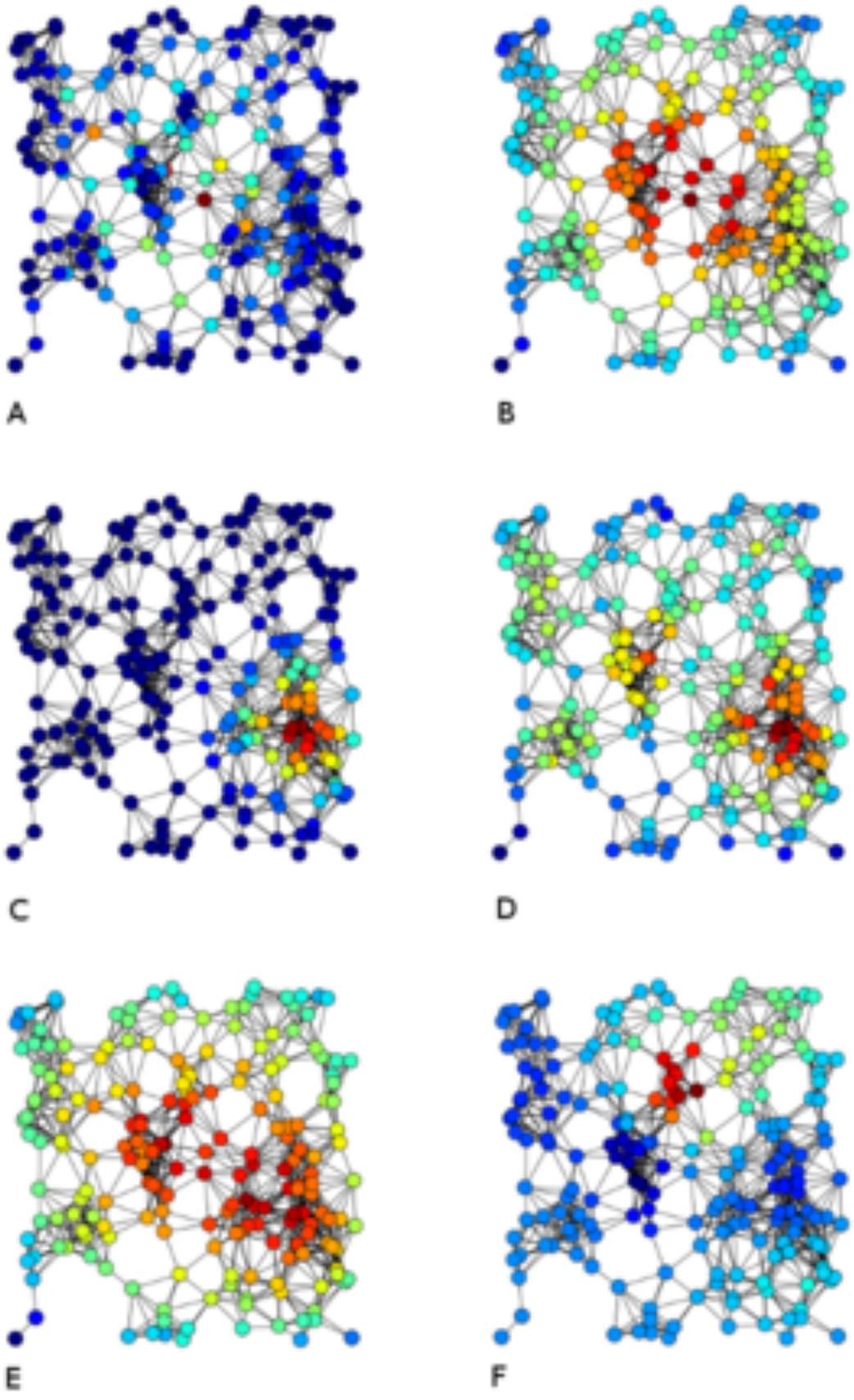


BETWEENNESS

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

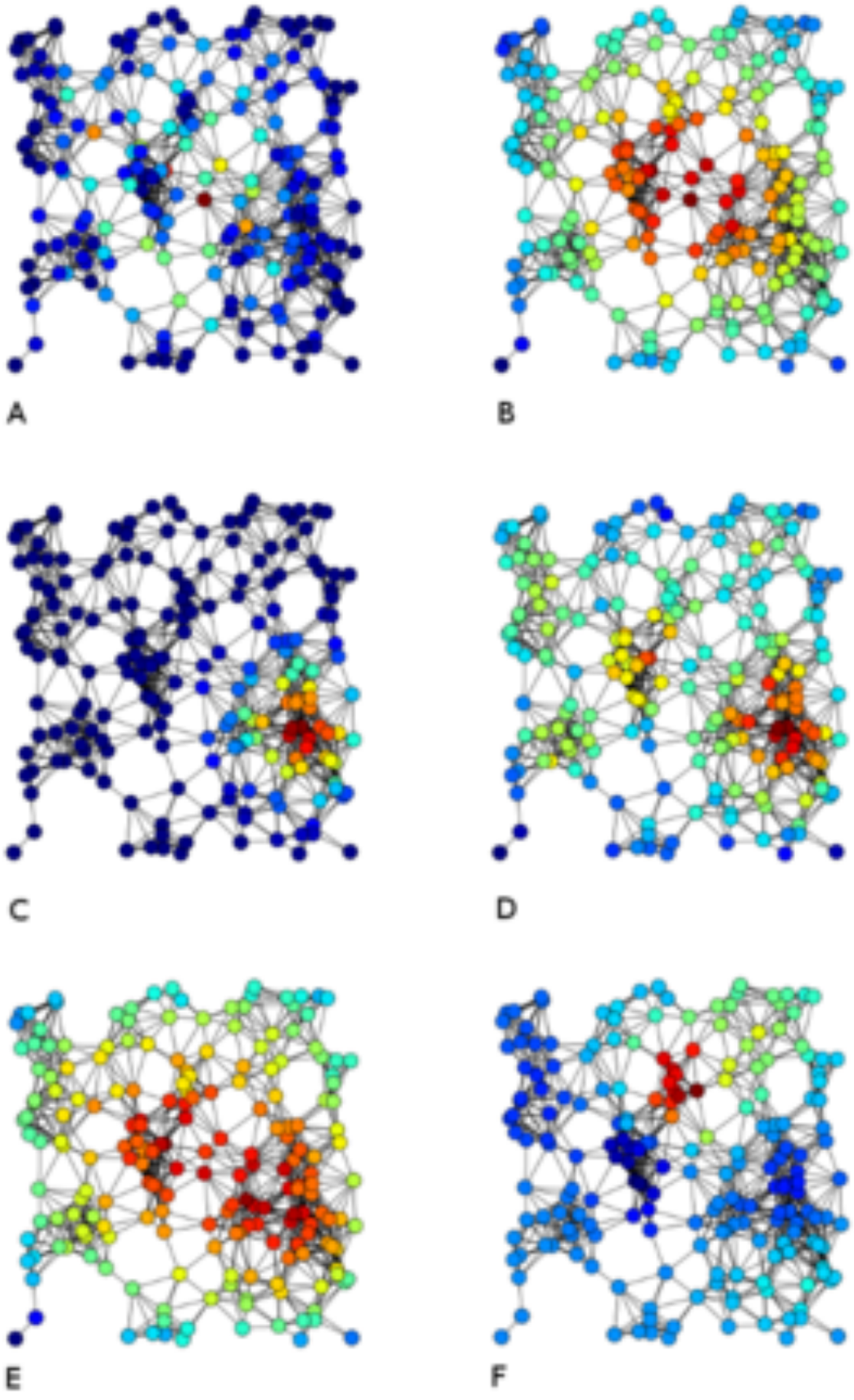
OTHERS

- 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, …)

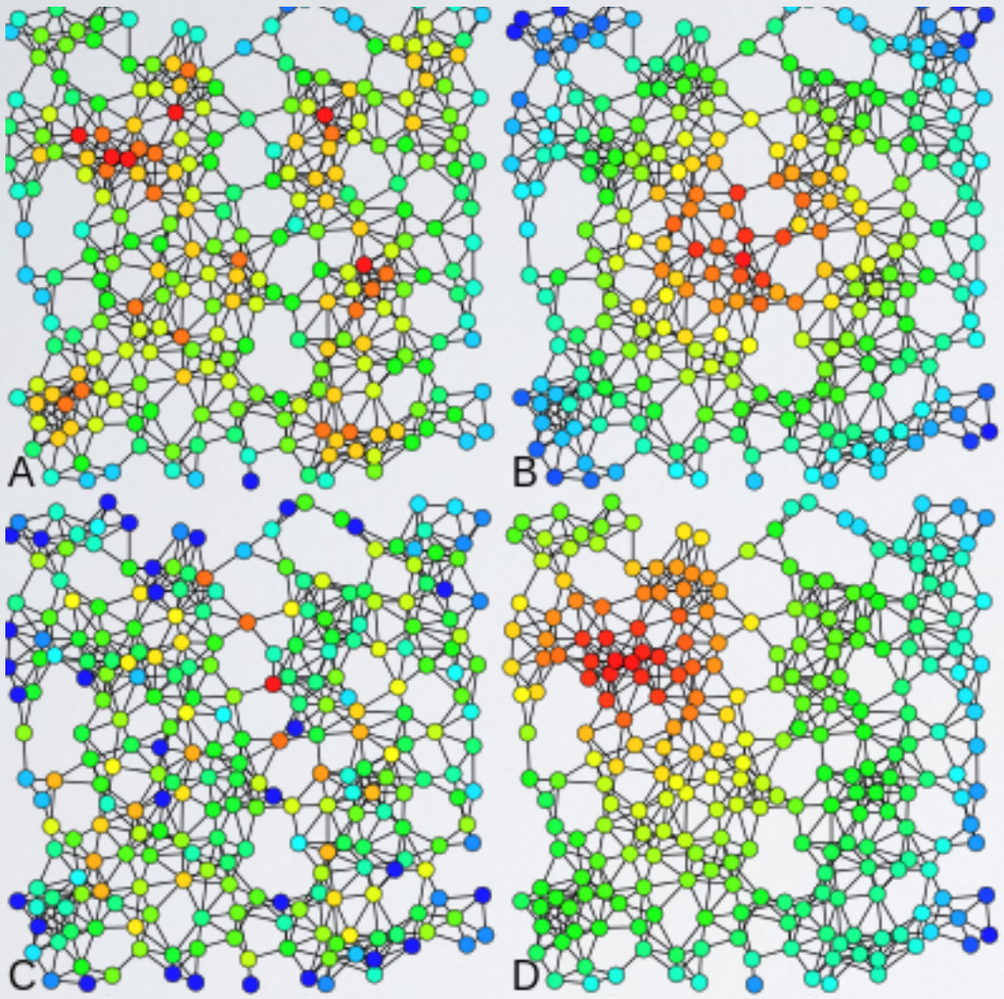


Which is which?

Harmonic
Closeness
Betweenness
Eigenvector
Katz
Degree



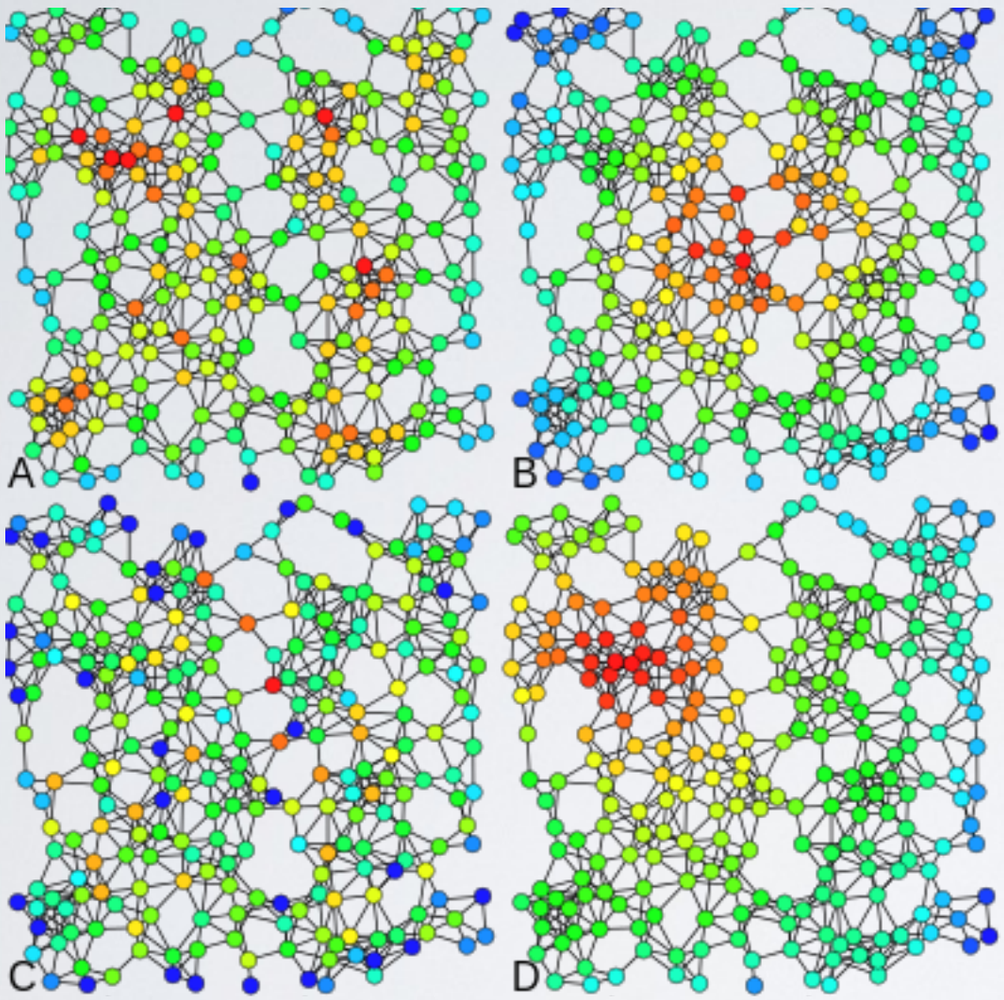
A: Betweenness
B:Closeness
C:Eigenvector
D:Degree
E:Harmonic
F: Katz



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*
- Centrality measures should be used with caution for (a) for exploratory purposes and (b) for characterisation

SOME EXAMPLES ON REAL NETWORKS

WIKIPEDIA

- What are the most important pages on Wikipedia ?
- Wikipedia network:
  - Nodes are pages
  - Links are hypertext links
- Wikipedia in english: Cultural bias !
- Results from http://wikirank-2019.di.unimi.it

WIKIPEDIA

Table 1

Page views	**harmonic centrality**	**indegree**	**PageRank**
0. Main Page	0. United States	0. United States	0. United States
1. Hyphen-minus	1. World War II	1. Association football	1. Association football
2. Louis Tomlinson	2. United Kingdom	2. World War II	2. France
3. Darth Vader	3. Association football	3. France	3. Iran
4. Lists of deaths by year	4. World War I	4. Germany	4. World War II
5. Exo (band)	5. France	5. India	5. Germany
6. List of stand-up comedians from the United Kingdom	6. Catholic Church	6. New York City	6. India
7. List of United States stand-up comedians	7. Germany	7. United Kingdom	7. Moth
8. List of stand-up comedians	8. China	8. Iran	8. United Kingdom
9. List of Australian stand-up comedians	9. India	9. London	9. Australia

WIKIPEDIA

Table 1

PageRank	**Harmonic Centrality**	**Indegree**	**Page Views**
0. Gone with the Wind (film)	0. Avatar (2009 film)	0. The Wizard of Oz (1939 film)	0. Black Panther (film)
1. The Wizard of Oz (1939 film)	1. Gone with the Wind (film)	1. Star Wars (film)	1. Deadpool 2
2. Cinema of Japan	2. The Wizard of Oz (1939 film)	2. Titanic (1997 film)	2. Venom (2018 film)
3. Star Wars (film)	3. The Godfather	3. Gone with the Wind (film)	3. A Quiet Place (film)
4. Titanic (1997 film)	4. Citizen Kane	4. Avatar (2009 film)	4. The Shape of Water
5. The Godfather	5. Casablanca (film)	5. The Godfather	5. Avengers: Endgame
6. Citizen Kane	6. Lawrence of Arabia (film)	6. The Lion King	6. Ant-Man and the Wasp
7. Avatar (2009 film)	7. On the Waterfront	7. The Matrix	7. The Greatest Showman
8. Casablanca (film)	8. Titanic (1997 film)	8. The Dark Knight (film)	8. A Star Is Born (2018 film)
9. Blade Runner	9. Schindler's List	9. Blade Runner	9. Ready Player One (film)

Similarity measures

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

σ: 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

Homophily - Assortative mixing

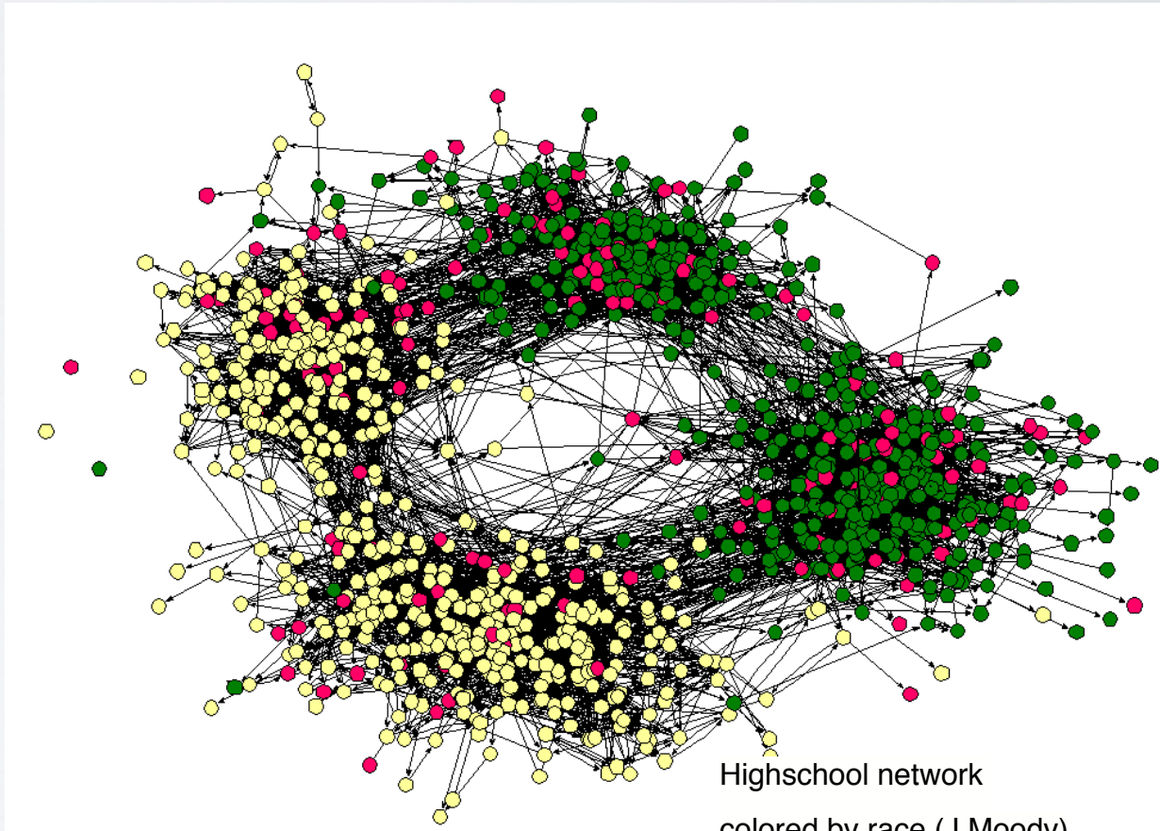
"birds of a feather flock together"

- 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)?

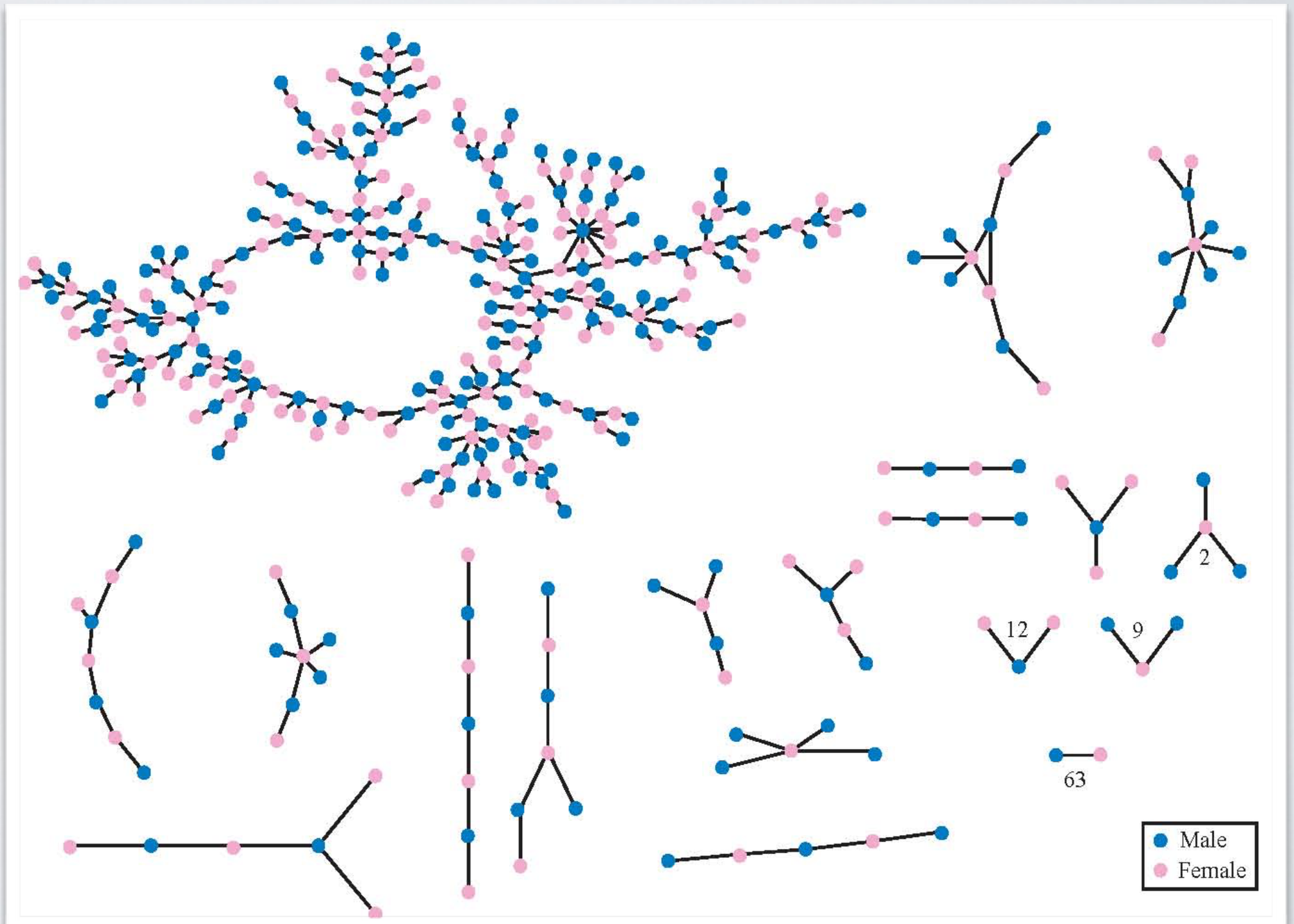
Homophily - Assortative mixing

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

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

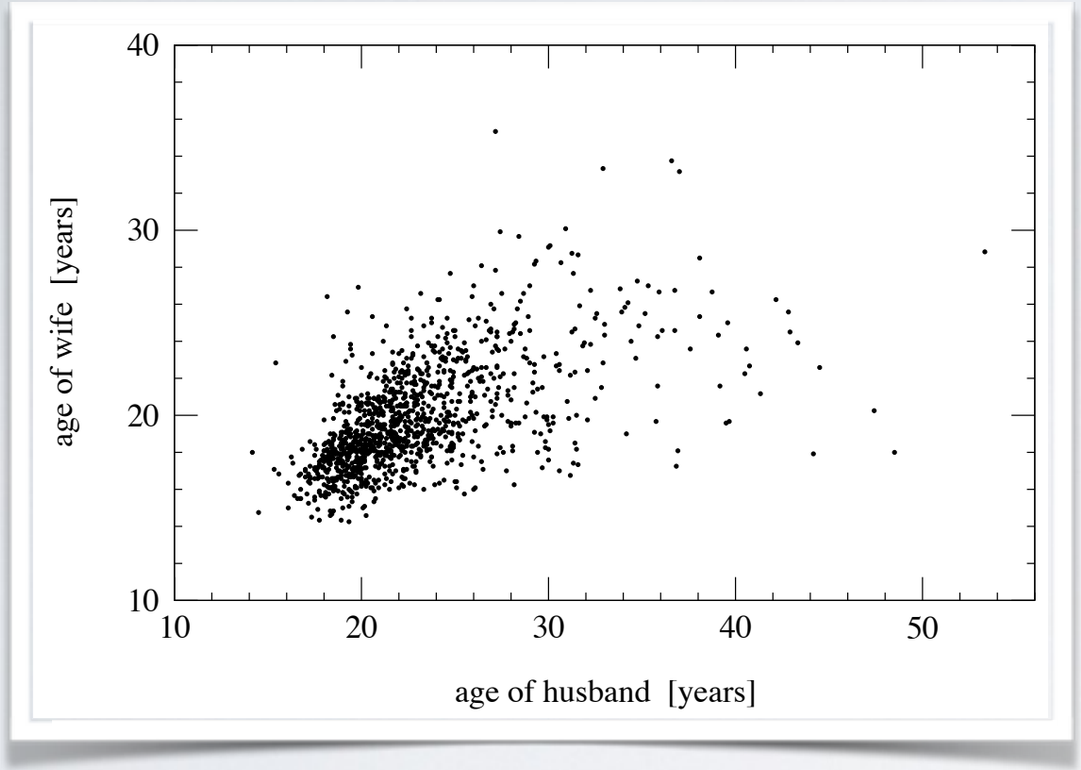
No assortative mixing : r=0 ($e_{ij} = a_i b_j$)

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 σ_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 σ_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'

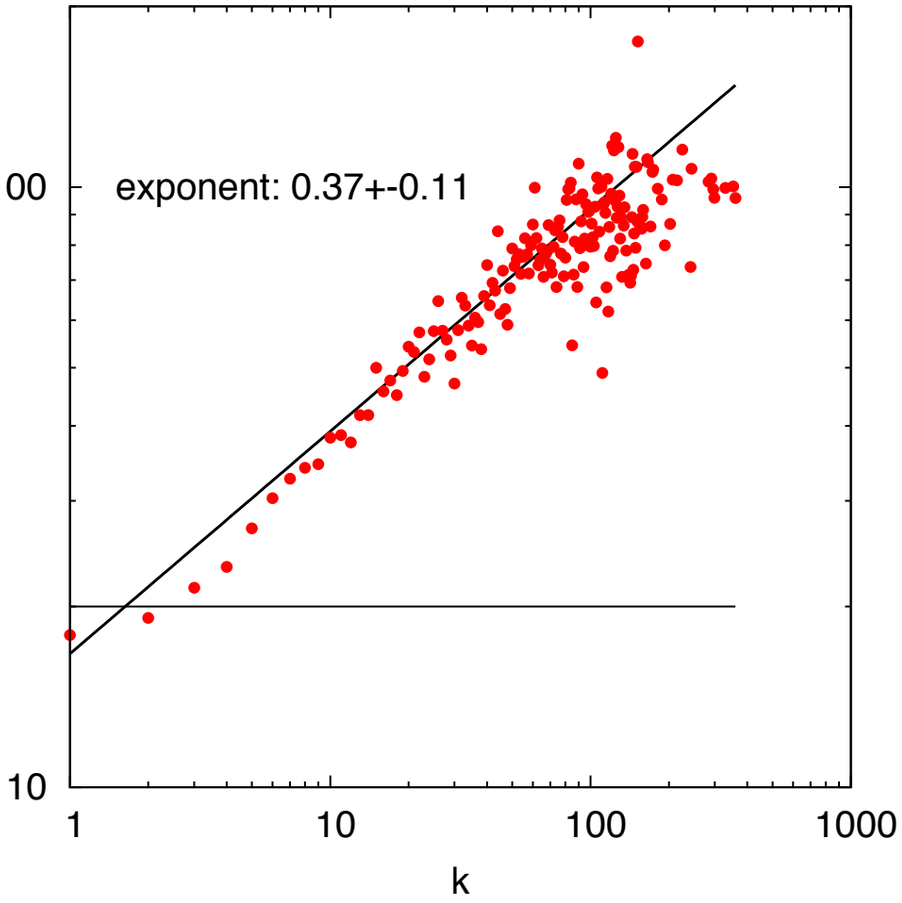
- If there are no degree correlations:

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

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

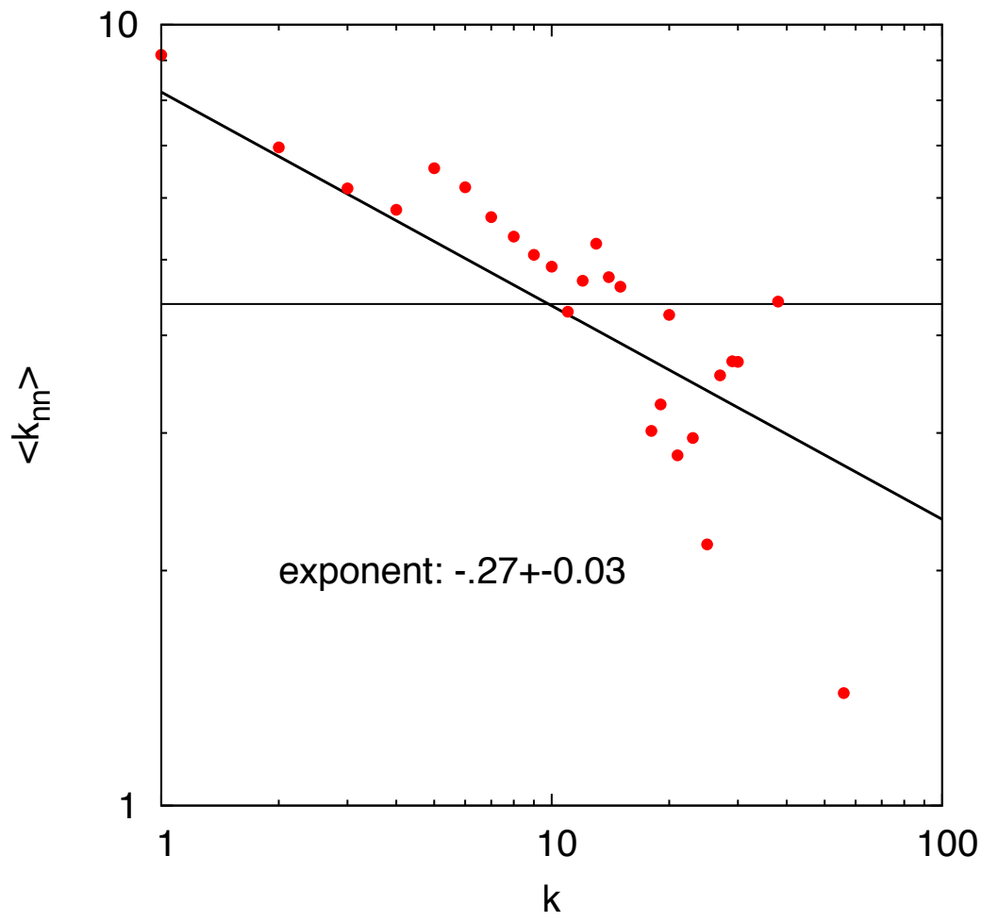
- If the network is assortative $k_{nn}(k)$ is a positive function
- If the network is disassortative $k_{nn}(k)$ is a negative function

Nearest neighbour degree



Astrophysics co-authorship network

Assortative



Yeast PPI

Disassortative

Nearest neighbour degree

/!\ These definitions suppose a finite variance.

One of the properties of power-law degree distributions is that they have **infinite variance**

Imagine a network with a node of degree 10 and 10 nodes of degree 1: by construction, they cannot have the same average degree of neighbors

Other measures need to be applied
(see for instance https://arxiv.org/pdf/1704.05707.pdf)

Rich-club coefficient

- How well connected are the well connected among themselves
- It is calculated on a list of node degree sorted in ascendant order as

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)}$$

- $N_{>k}$ denotes the number of nodes with degree k or larger than k
- $E_{>k}$ measures the number of links between them

Algorithm

- rank nodes by degree
- remove nodes in an ascendant degree order
- measure the density of the remaining network

- Results are usually compared to random references
  - configuration model of equivalent synthetic network
  - configuration model of the empirical network