CS-E5740

Complex Networks

Network analysis: key measures and characteristics

Course outline

1. Introduction (motivation, definitions, etc.)
2. Static network models: random and small-world networks
3. Growing network models: scale-free networks
4. Percolation, error & attack tolerance of networks, epidemic models
5. Network analysis: key measures and characteristics
6. Social networks & (socio)dynamic models
7. Weighted networks
8. Clustering, sampling, inference
9. Temporal networks & multilayer networks

Universal and not-so-universal characteristics of complex networks

- For real-world networks,
  - path lengths are usually short,
  - the networks are almost always clustered,
  - degree distributions are usually broad, sometimes scale-free

- However, there are many other network characteristics that depend on the type of network at hand!
- This lecture:

  different network measures and their applications

Part 1:
Centrality measures
for *individual nodes*

Centrality = *How important is a node?*

...with respect to some assumptions.

Some example uses for centrality measures

Which are the most important…

1. … **airports** in the network of flights …

- … if you want to access most other places easily?
- … to screen for diseased people in order to stop epidemics?

2. … **routers** in the internet …

- … that should be close your data centres?
- …that are likely to have the most traffic?

3. … **people** in a social network …

- … to be vaccinated first to stop a disease spreading?
- … as initial targets of a viral marketing campaign?

Centrality measures

- Idea: **How important is a node in the network?**

1. Degree centrality

Important nodes have many connections

2. Betweenness centrality

Important nodes work as bridges

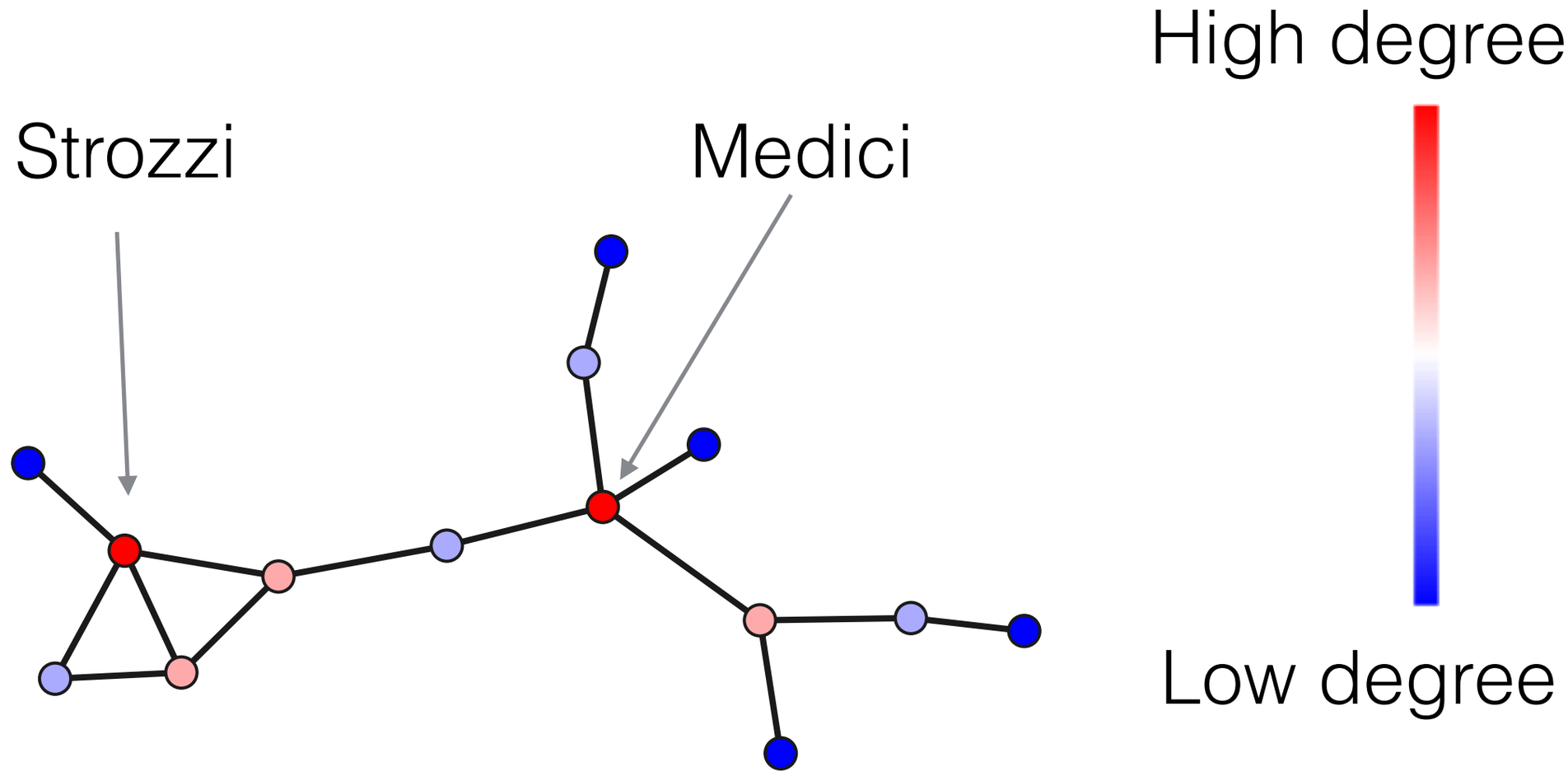
3. Closeness centralities

Important nodes are close to other nodes

4. Eigenvector centralities

Important nodes are connected to other important nodes

Network of alliances between Florentine families



$k_{Strozzi}=4$

$k_{Medici}=4$

D. Kent "The rise of the Medici: Faction in Florence, 1426 - 1434", OUP (1978)

Betweenness centrality

=

On how many shortest paths is a node on?

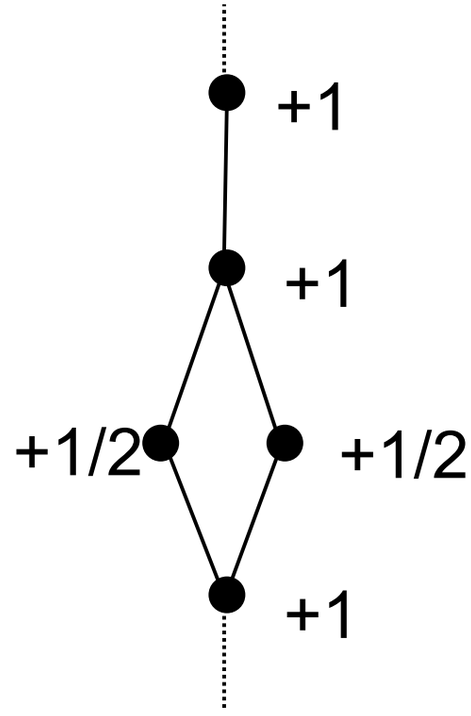
Centrality measures: betweenness

- *Measures "traffic" or flow through a node or a link*, if all nodes communicate to all others via the shortest paths

- Formally, *betweenness centrality = fraction of shortest paths going through node/link*

- If there are multiple shortest paths (there usually are), divide by multiplicity

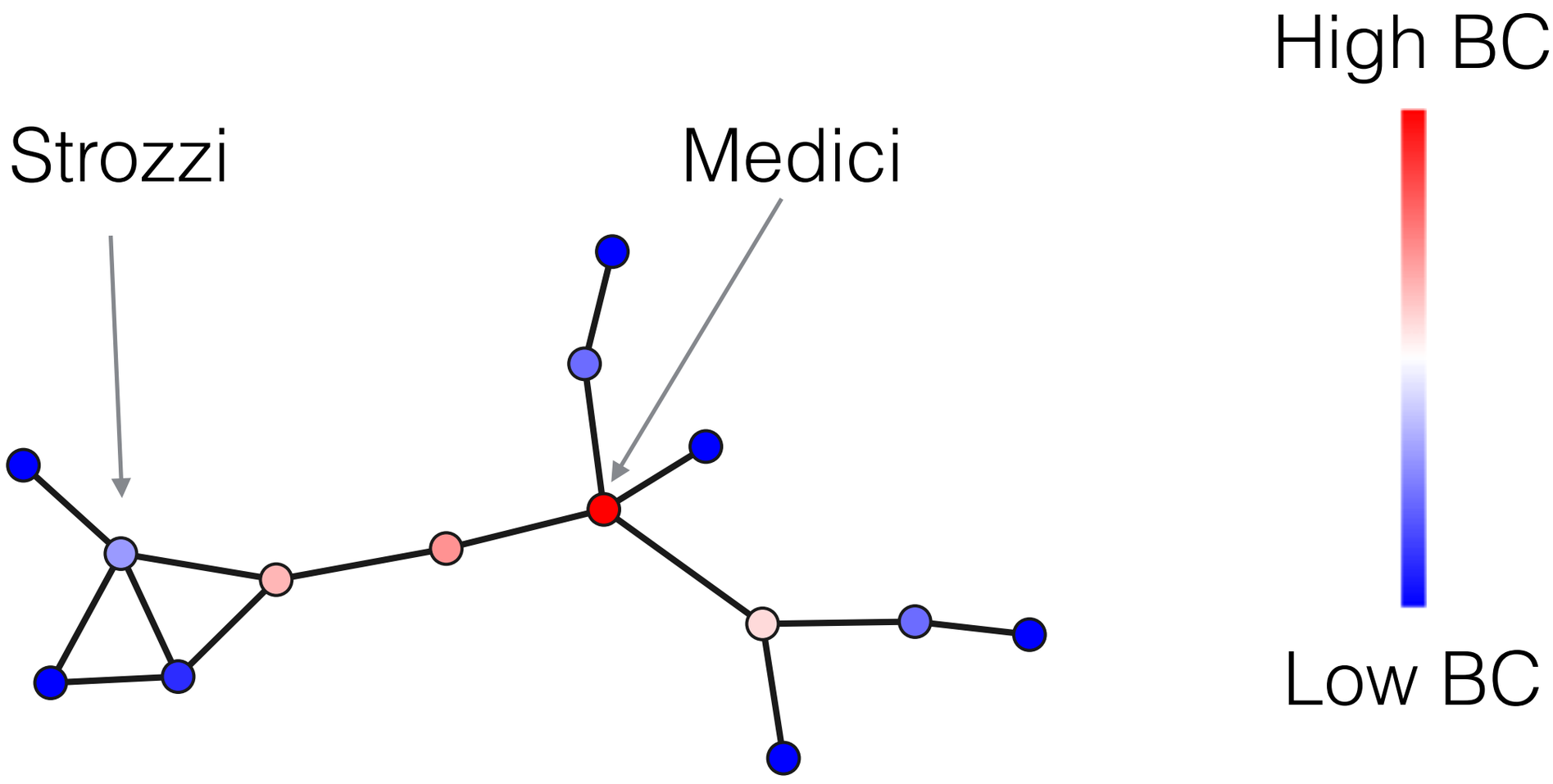
- Hard to compute for large networks (there are at $N(N\text{-}1)/2$ shortest paths...)



- See Newman, Phys. Rev. E **64**, 016132 (2001) for a fast algorithm ($O(NE)$, N = # of nodes, E = # of links)

L. Freeman "A Set of Measures Based of Centrality Based on Betweenness", Sociometry 40(1): 35- 41 (1977)

Network of alliances between Florentine families



$k_{Strozzi}=4$

$bc_{Strozzi}=34$

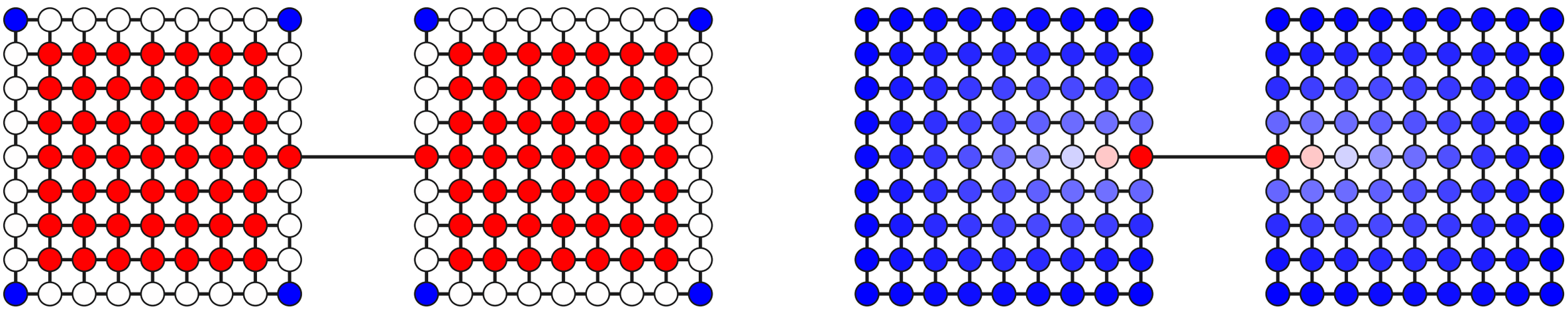
$k_{Medici}=4$

$bc_{Medici}=112$

Centralities in 2 connected grids

Degree

Betweenness



Low value

High value

Betweenness: problems

- Hard to compute for large networks
- ...however, one can *sample* sets of shortest paths, e.g. from 1% of nodes to all others
- Makes sense only if the underlying philosophy makes sense - nodes communicate with all other nodes through intermediate nodes and shortest paths
- E.g. for large social networks, this is not necessarily true

Centrality measures

- Idea: **How important is a node in the network?**

1. Degree centrality

Important nodes have many connections

2. Betweenness centrality

Important nodes work as bridges

3. Closeness centralities

Important nodes are close to other nodes

4. Eigenvector centralities

Important nodes are connected to other important nodes

Closeness centrality

=

How far is a node from all other nodes?

Centrality measures: closeness, efficiency

- **Closeness**: how far is a node from all other nodes, on average?
- $C_c(i) = \frac{1}{\langle l_i \rangle} = \frac{N-1}{\sum_{j \neq i} d_{ij}}$
- (d_{ij} = length of shortest path between i and j, i.e. their distance; $<l_i>$=avg distance from i to others. Closeness $C_c(i)$ = inverse avg distance)
- Doesn't directly work for networks with disconnected components where for some pairs $d_{ij} = \infty$
- Related to avg. path lengths & the **efficiency** (widely used in neuroscience)
- Efficiency between a pair of nodes i and j: $\epsilon_{ij} = 1/d_{ij}$
- Avg efficiency of i:

$$\langle \epsilon_i \rangle = \frac{1}{N-1} \sum_{j \neq i} \frac{1}{d_{ij}}$$

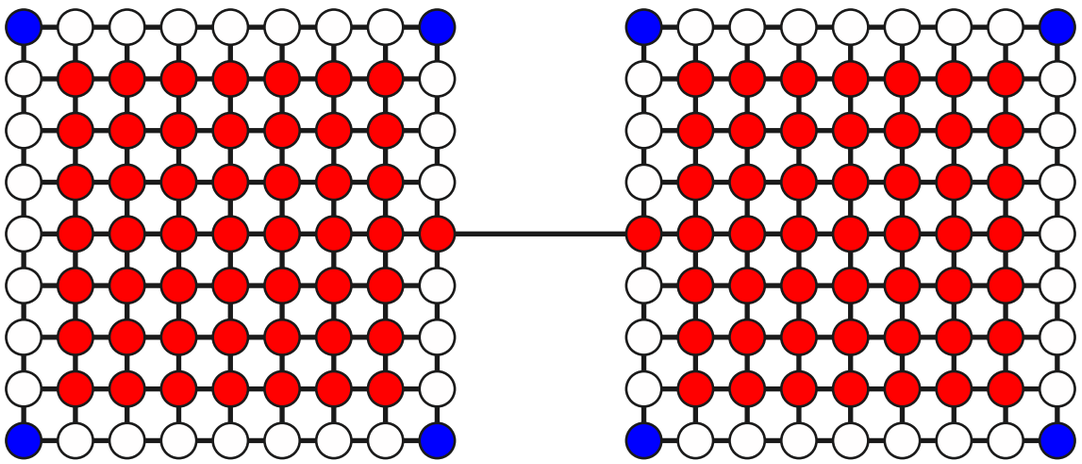
- Network efficiency (all pairs of nodes):

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

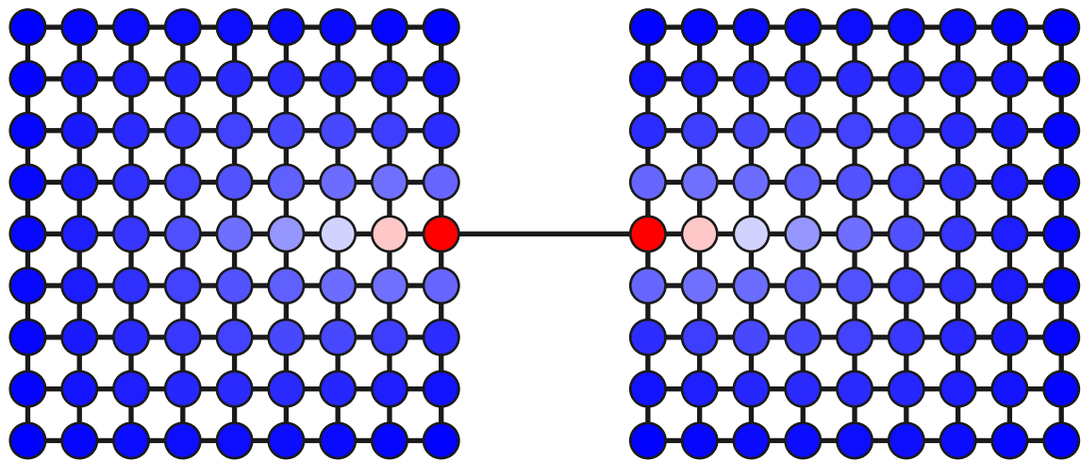
(aspiring brain scientists: see, e.g., Olaf Sporns: Networks of the Brain (MIT Press, 2010), Stam & Reijneveld: Graph theoretical analysis of complex networks in the brain, Nonlinear Biomedical Physics 2007, 1:3)

Centralities in 2 connected grids

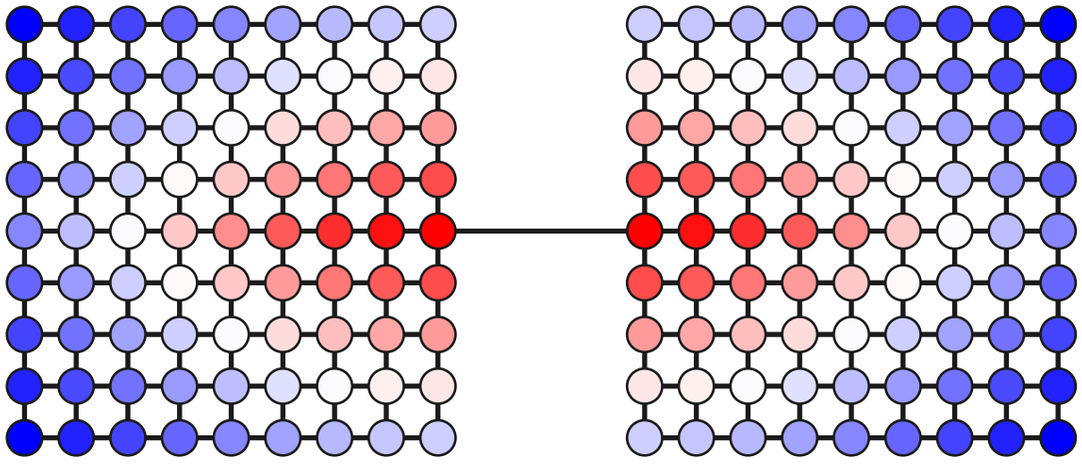
Degree



Betweenness



Closeness



Low value — High value

Katz Centrality

- Idea: Calculate the number of walks of any length from node i from any other node

- Each step is "active" with probability a

- Infinitely many walks, but can be solved with matrix inversion:

$$t_i = \sum_j ((I - aA)^{-1} - I)_{ij}$$

Katz centarality for node i

$$t_i = \sum_j \sum_k^{\infty} a^k (A^k)_{ij}$$

j k

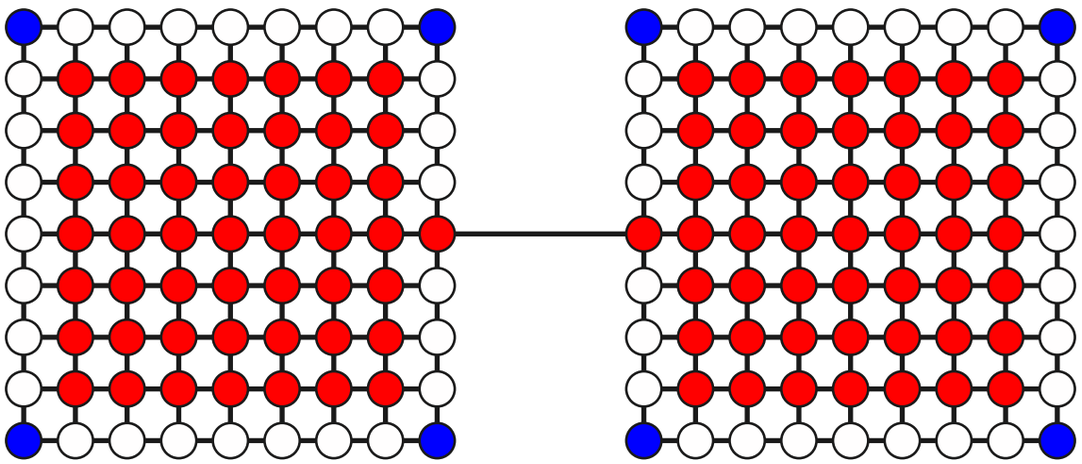
Probability that walk of length k active

Number of walks of length k from i to j

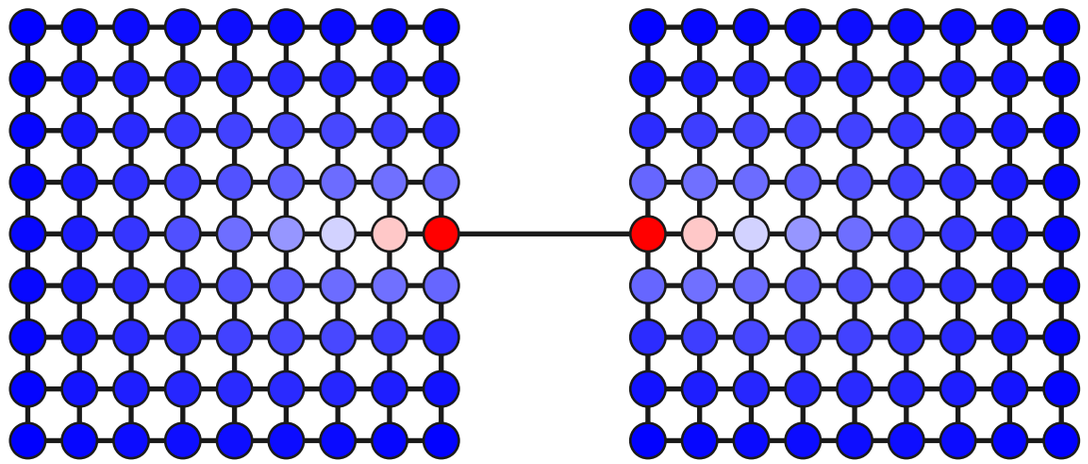
L. Katz "A New Status Index Derived From Sociometric Analysis", Psychometrica 18:1 39- 43 (1953)

Centralities in 2 connected grids

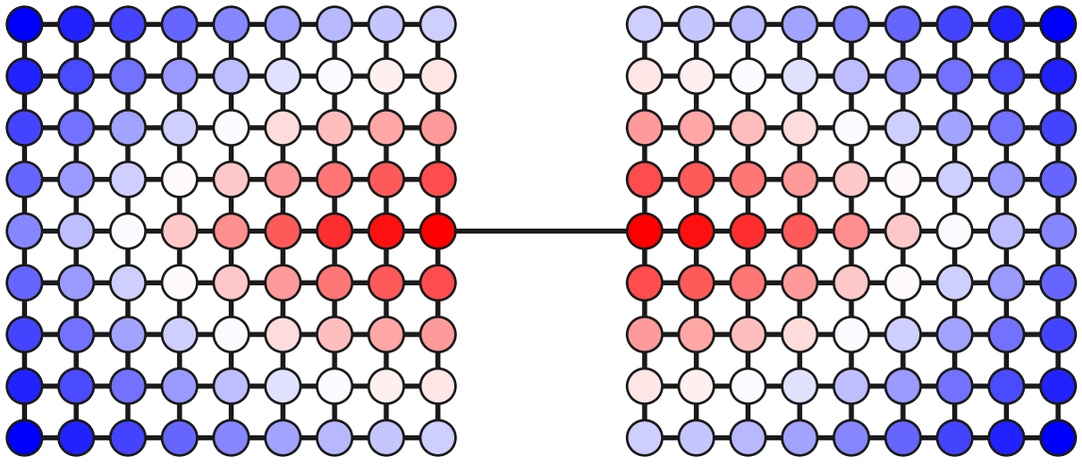
Degree



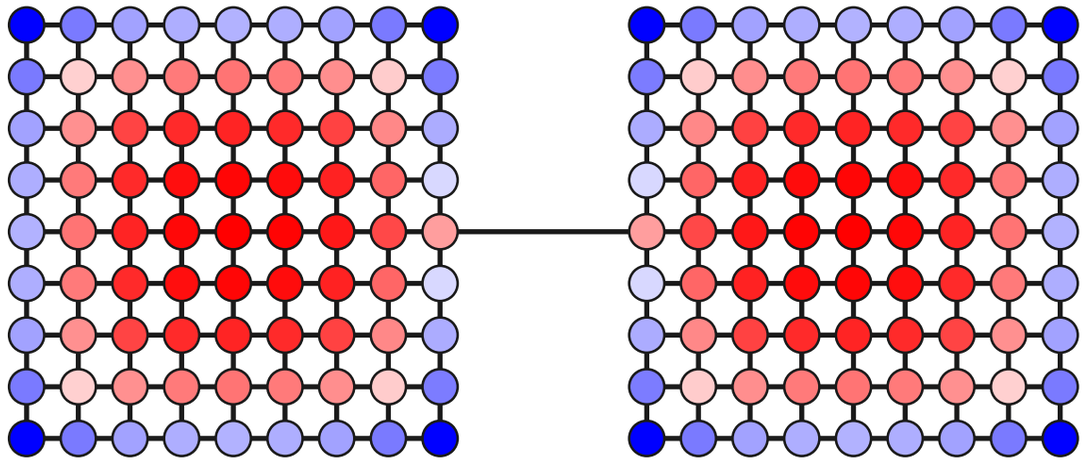
Betweenness



Closeness



Katz, a=0.2



Low value — High value

Centrality measures

- Idea: **How important is a node in the network?**

1. Degree centrality

Important nodes have many connections

2. Betweenness centrality

Important nodes work as bridges

3. Closeness centralities

Important nodes are close to other nodes

4. Eigenvector centralities

Important nodes are connected to other important nodes

Eigenvector centrality

=

How well-connected a node is to other well-connected nodes

Eigenvector centrality by iteration

- Let x_i be the centrality of node i.
- Start with $x_i(0) = 1 \; \forall i$, and define the next value as the sum of neighbors' centrality:

$$x_i(1) = \sum_{j=1}^{n} a_{ij} x_j(0)$$

In general we write this in matrix form as

$$\mathbf{x}(t) = A\mathbf{x}(t-1)$$
$$\Rightarrow \mathbf{x}(t) = A^t \mathbf{x}(0)$$

- Next write $\mathbf{x}(0)$ using the eigenvectors $\mathbf{v}_i$ of A:

$$\mathbf{x}(0) = \sum_{i=1}^{n} c_i \mathbf{v}_i$$

- Now we can write $A\mathbf{v_i} = \lambda_i \mathbf{v_i}$

$$\mathbf{x}(t) = A^t \sum_{i=1}^{n} c_i \mathbf{v}_i = \sum_{i=1}^{n} c_i \lambda_i^t \mathbf{v}_i$$
$$= \lambda_1^t \sum_{i=1}^{n} c_i \left[\frac{\lambda_i}{\lambda_1}\right]^t \mathbf{v}_i$$

where λ_1 is the largest eigenvalue.

Eigenvector centrality: 1st eigenvector

- We have

$$\mathbf{x}(t) = \lambda_1^t \sum_{i=1}^{n} c_i \left[\frac{\lambda_i}{\lambda_1}\right]^t \mathbf{v}_i$$

where λ_1 is the largest eigenvalue.

- Since $\frac{\lambda_i}{\lambda_1} < 1$ for $i > 1$ we get

$$\mathbf{x}(t) \xrightarrow{t\to\infty} \lambda_1^t c_1 \mathbf{v}_1$$

- In other words $\mathbf{x}$ is proportional to $\mathbf{v}_1$.

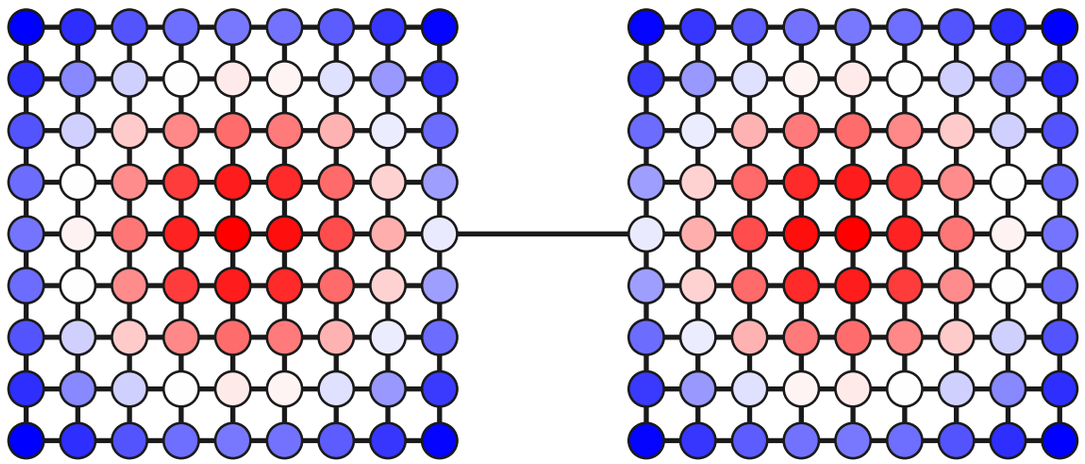
Eigenvector centrality

- Eigenvector centrality $\mathbf{x}$ is the eigenvector corresponding to the largest eigenvalue of the adjacency matrix.

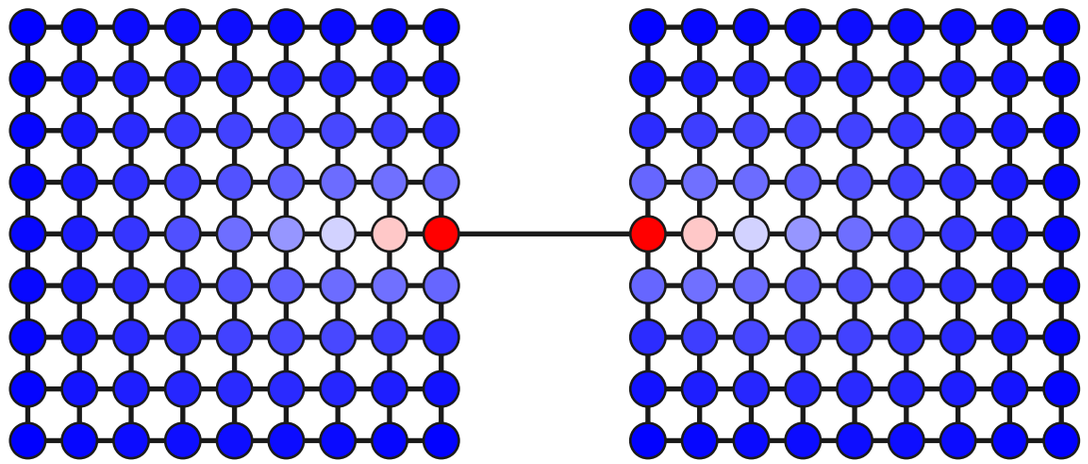
eigenvector centrality
is large, if a node
has many neighbours
who have many neighbours...

Centralities in 2 connected grids

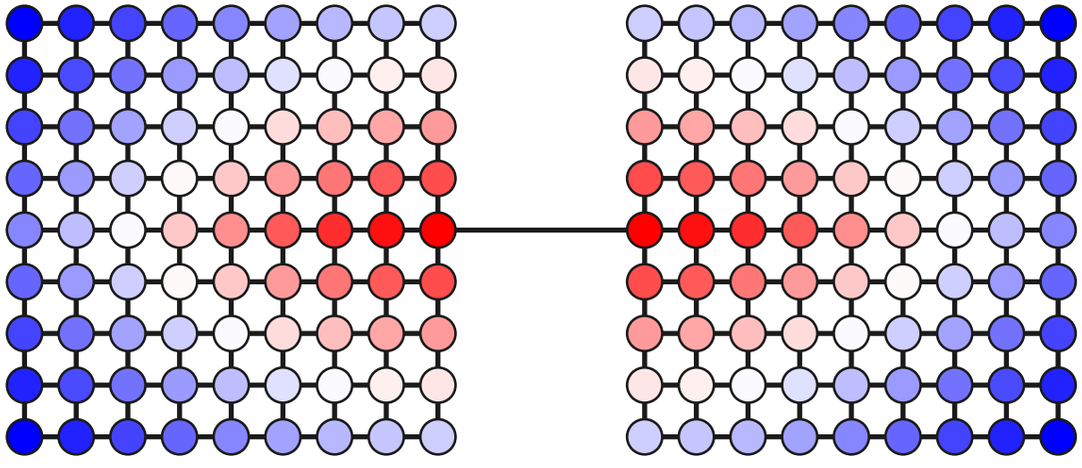
Eigenvector



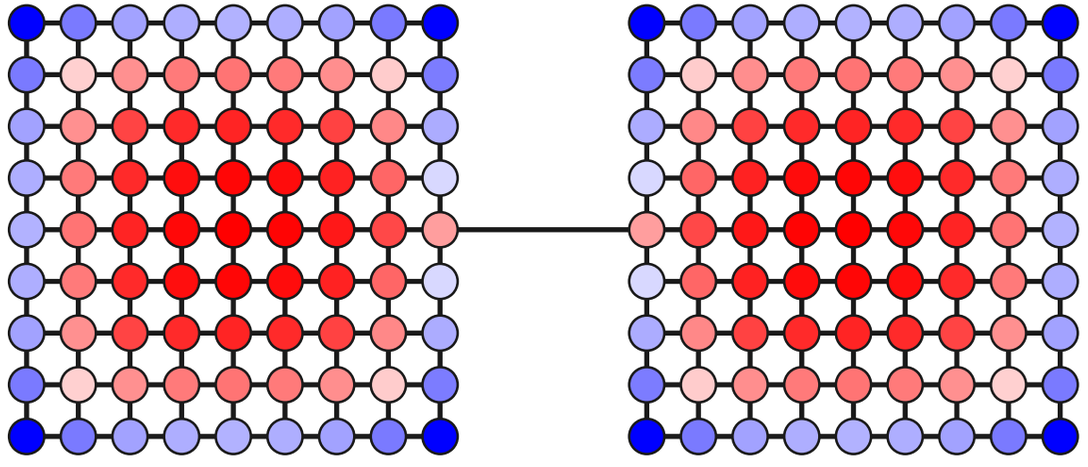
Betweenness



Closeness



Katz, a=0.2



Low value — High value

Katz Centrality

- Katz centrality can be written as*:

$$t'_i = \sum_j a A_{ij} t'_j + 1$$

- Which is similar to the eigenvector equation

- Can be solved with iteration
- Katz centrality can be also be thought of as eigenvector-like centrality
- Has solution iff $a < \frac{1}{\lambda_1}$

$$*: \; t'_i = t_i + 1$$

Katz centarality for node i

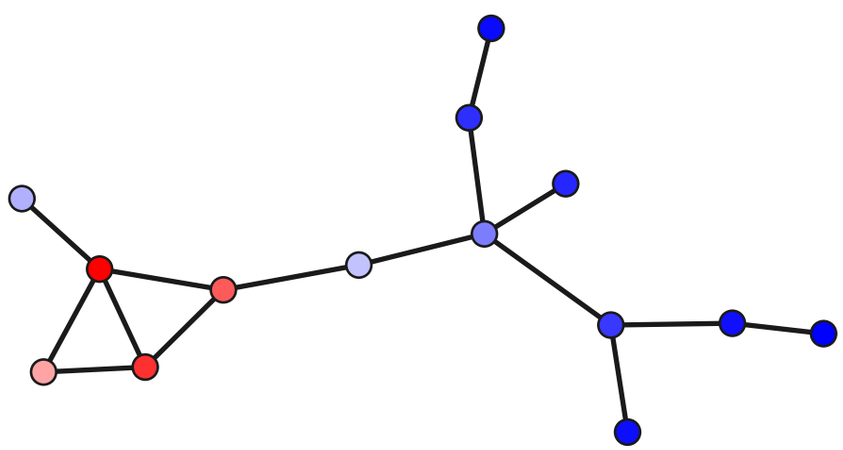
$$t_i = \sum_j \sum_k^{\infty} a^k (A^k)_{ij}$$

Probability that walk of length k active

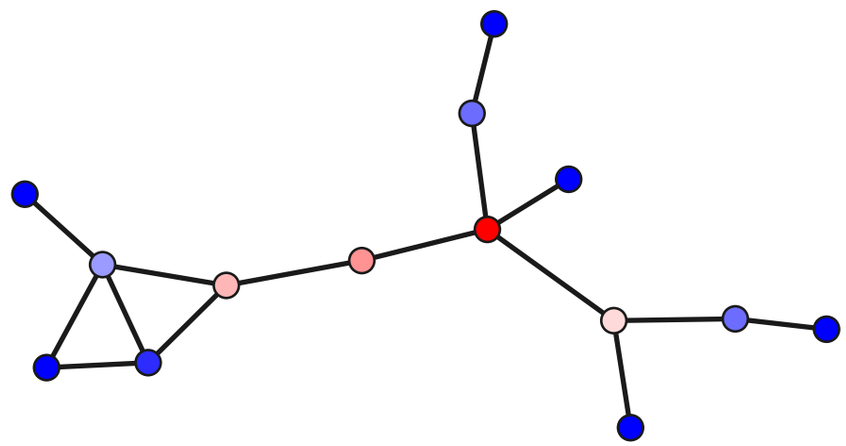
Number of walks of length k from i to j

Centralities in Florentine families

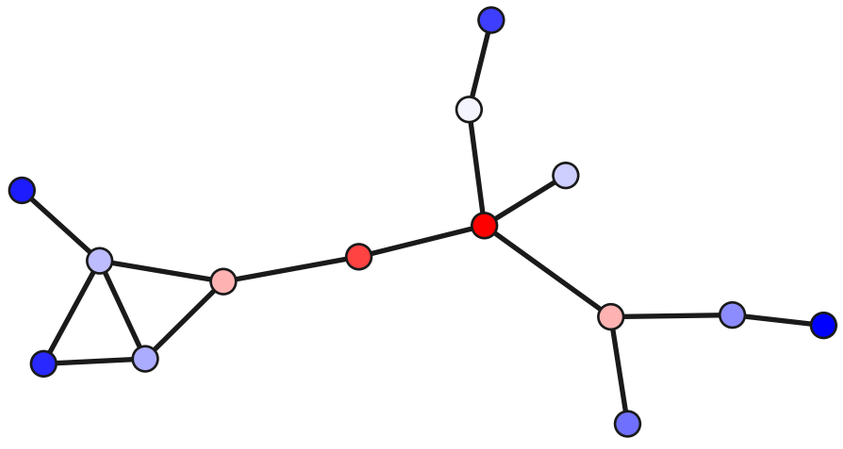
Eigenvector



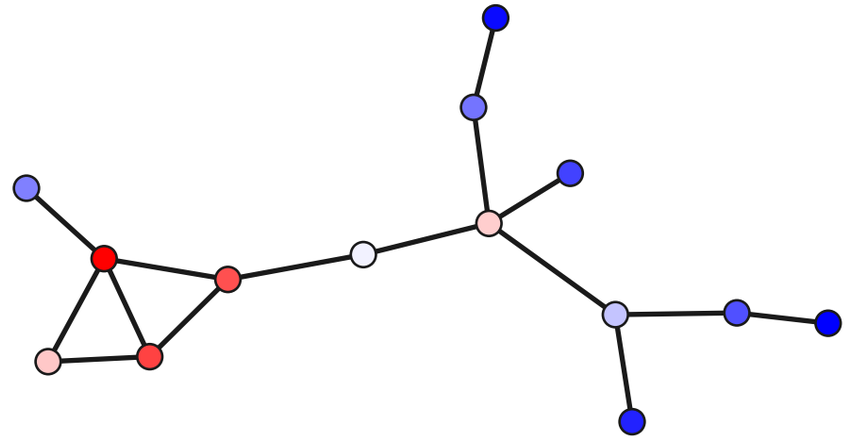
Betweenness



Closeness



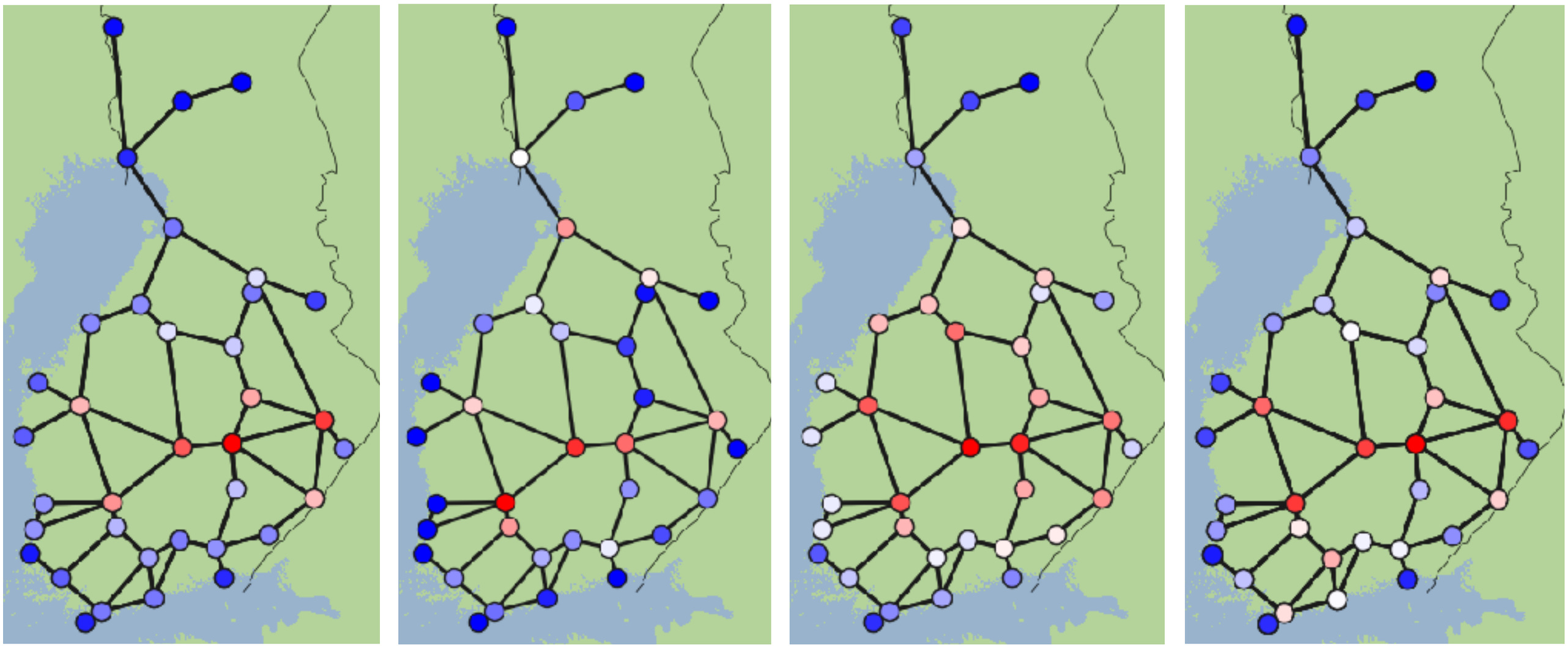
Katz, a=0.3



Low value — High value

Centralities in Finnish railroads

Eigenvector | Betweenness | Closeness | Katz, a=0.2



Low value — High value

PageRank

≈

eigenvector centrality

generalized 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/
To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine -- the first such detailed public description we know of to date. Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Keywords

World Wide Web, Search Engines, Information Retrieval, PageRank, Google

PageRank

- Idea: Random walker that jumps to random neighbour with probability d, and to any random node with probability 1-d

- PageRank is the probability x_i that the walker is found in node i (after infinite time)

- Doesn't get stuck to small loops with no exit etc.

- Depends only on node degree in undirected graphs

PageRank for node i

With probability 1-d the walker jumps…

… and with probability 1/N it arrives to node i

$$x_i = (1-d)\frac{1}{N} + d\sum_j A_{ji} x_j \frac{1}{k_j^{out}}$$

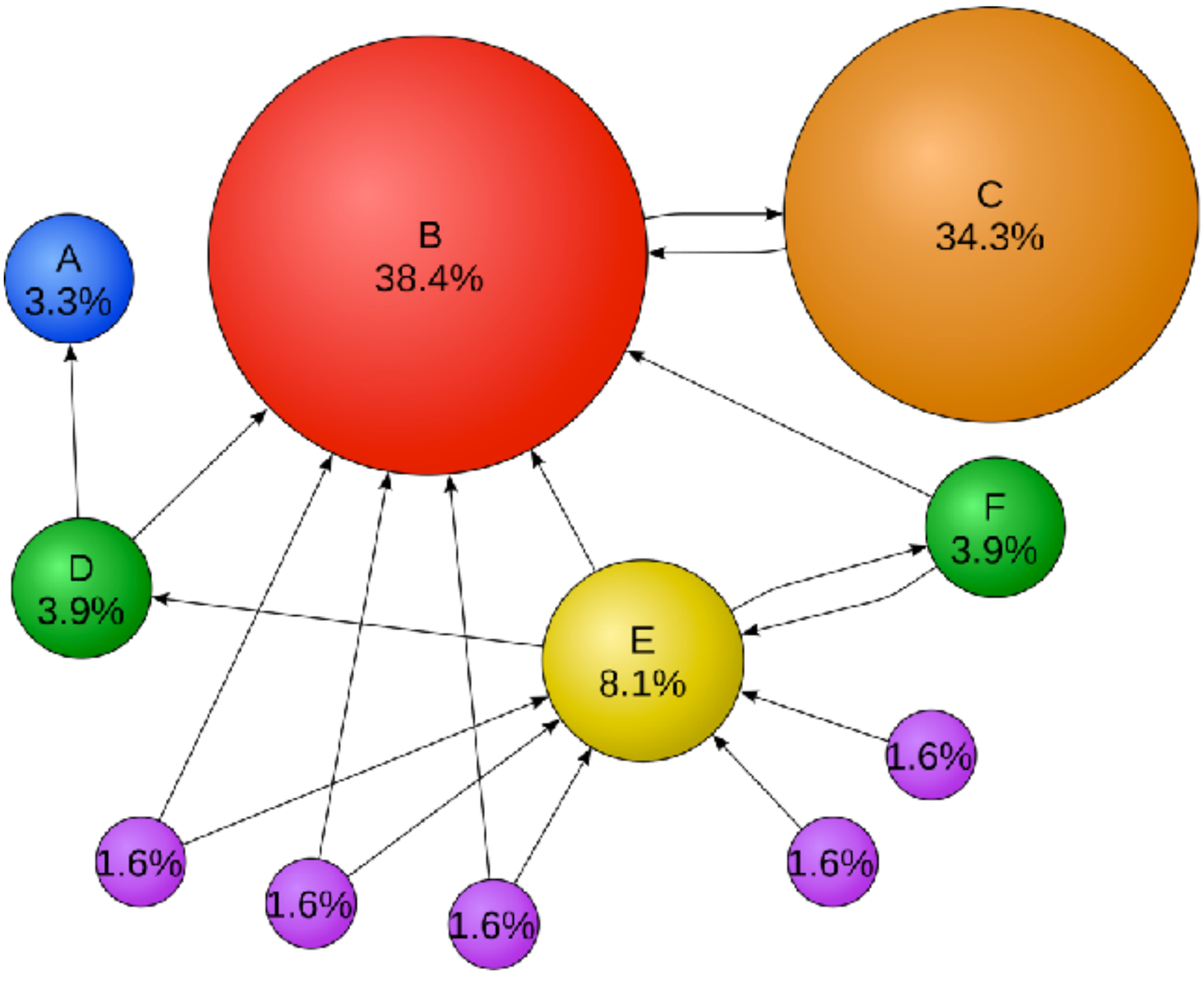
With probability d the walker moves…

… from a neighbour j …

… which has the walker with probability x_j …

… to i which is one of k_j^{out} neighbours.

PageRank example



Centrality measures

- Idea: **How important is a node in the network?**

1. Degree centrality

Important nodes have many connections

2. Betweenness centrality

Important nodes work as bridges

3. Closeness centralities

Important nodes are close to other nodes

4. Eigenvector centralities

Important nodes are connected to other important nodes

Part 2:
Measures for *sets of nodes*

Degree correlations

- If a vertex has degree k, what degrees do its neighbours have?
- One can estimate the conditional probability

$$P\left(k'|k\right)$$

- In practice, this probability is hard to estimate

- Usually, the average nearest-neighbour degree k_{nn} is used instead:

$$k_{nn}(k) = \frac{1}{N(k)} \sum_{i,k_i=k} \left[\frac{1}{k} \sum_{j \in \Gamma_i} k_j \right]$$

”nn”=”nearest neighbour”

“pick all nodes of degree k, calculate avg nearest-neighbour degree for each, and average over these values to get $k_{nn}(k)$”

Assortativity

- If $k_{nn}(k)$ is an **increasing** function, high-degree nodes tend to connect to each other and the network is **assortative**

- If $k_{nn}(k)$ is **decreasing**, hubs avoid each other, and the network is **disassortative**

- Social networks tend to be **assortative**

- Biological networks tend to be **disassortative**

- Alternative measure: the Pearson correlation coefficient between degrees of linked nodes:

$$r = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\sqrt{\langle k_i^2 \rangle - \langle k_i \rangle^2} \sqrt{\langle k_j^2 \rangle - \langle k_j \rangle^2}}$$

- r positive: assortative mixing

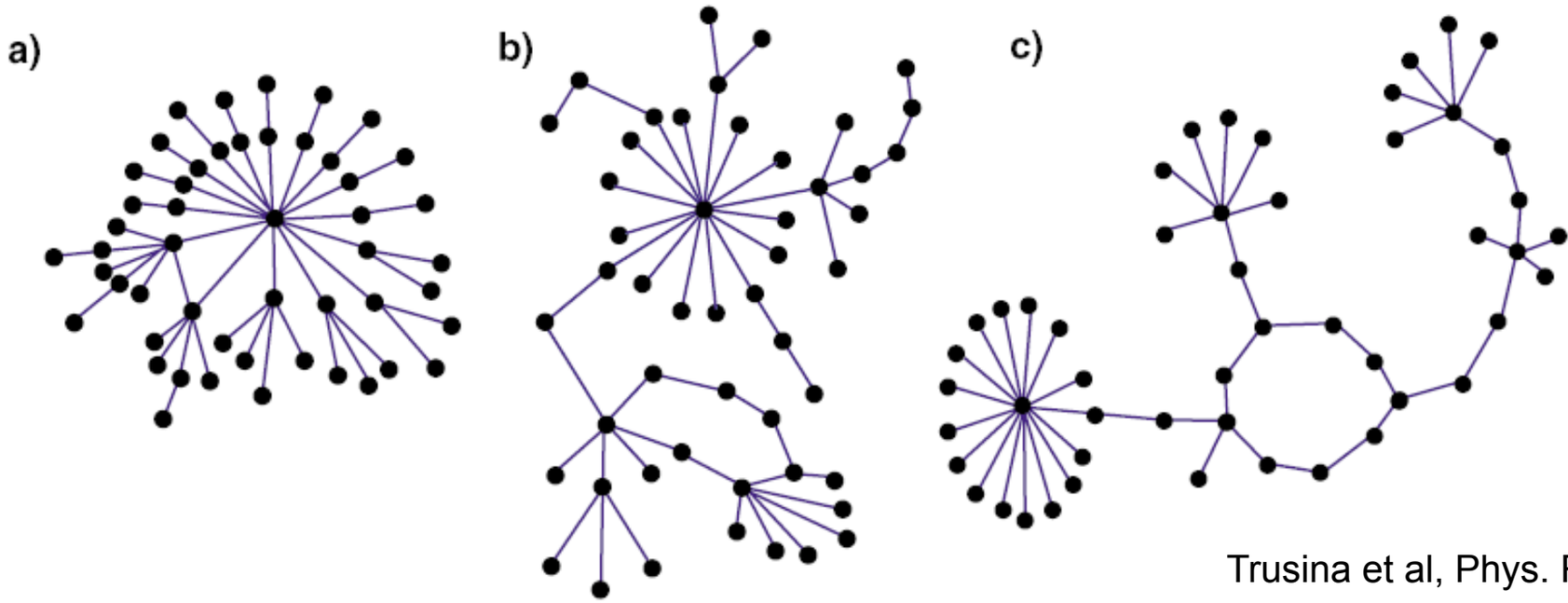
- r negative: disassortative mixing

Degree correlations: examples

maximally assortative

no correlations

maximally disassortative

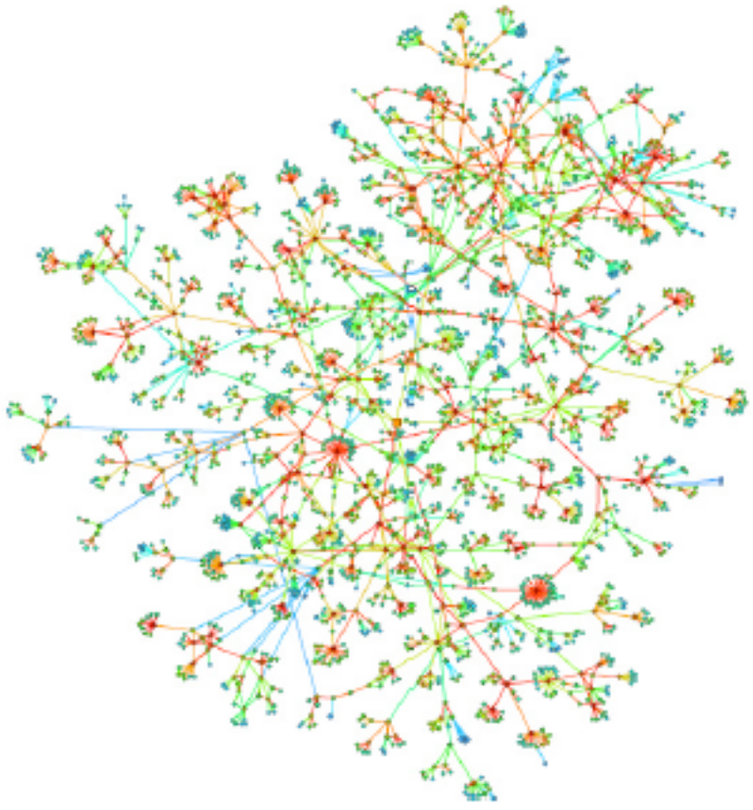


Trusina et al, Phys. Rev. Lett. 92, 2004

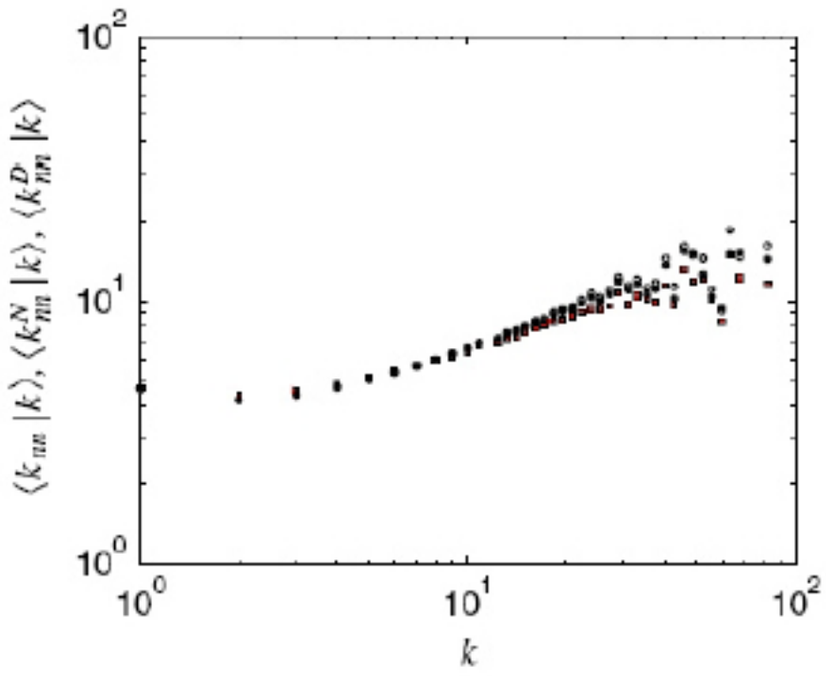
Three networks with exactly the same degree sequence,
but wired differently

Degree correlations: examples

Onnela, J.-P. et al. New Journal of Physics 9, 179+ (2007)



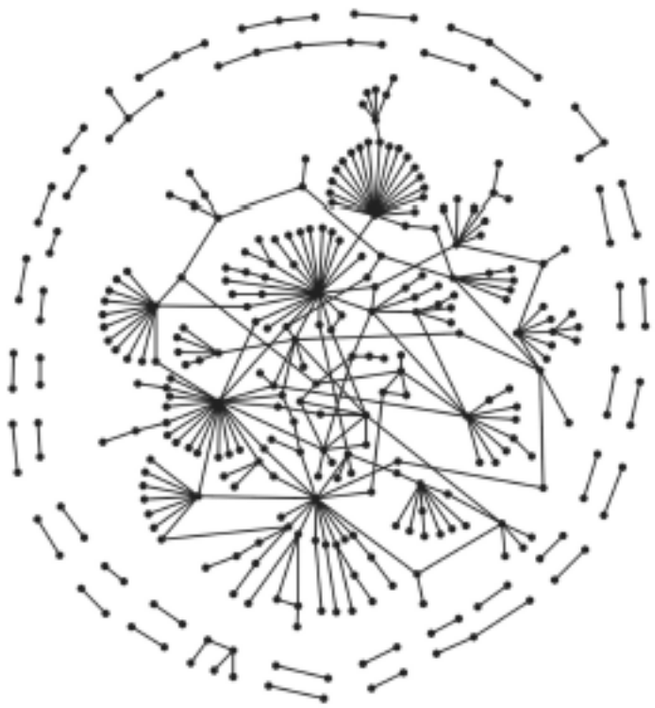
- Social networks are typically assortative: those with many friends know others with many friends.

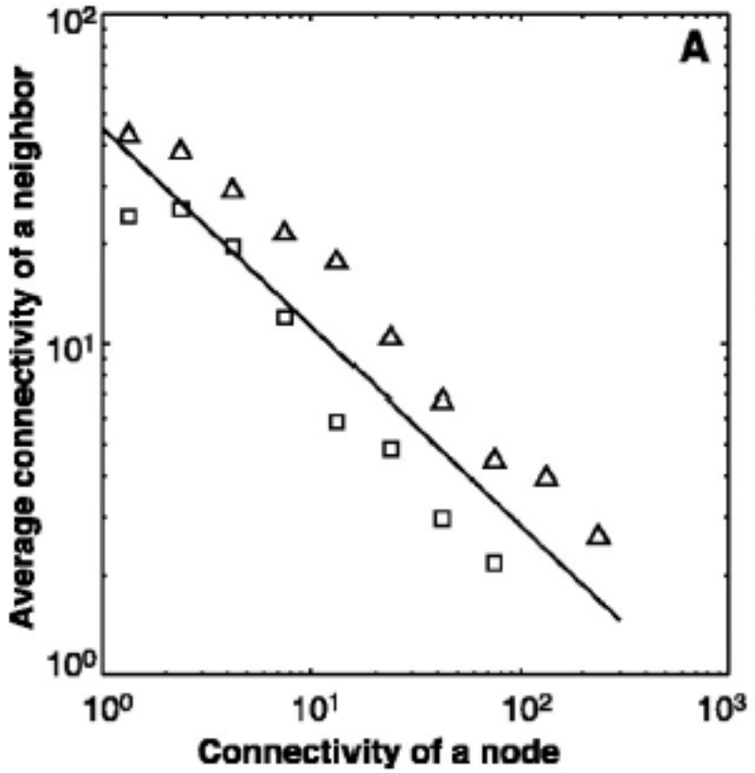


- The average mean degree $k_{nn}(k)$ for
  - (red squares) unweighted
  - (green circles) weighted by the number of calls
  - (black dots) weighted by the total call duration

Degree correlations: examples

Sergei Maslov and Kim Sneppen, Science 296, 910–913 (2002)





- Biological networks are typically disassortative: most neighbors of highly connected nodes have low degree

- The average mean degree $k_{nn}(k)$ for
  - (triangles) physical interaction network
  - (squares) regulatory network

The rich-club coefficient

- How well connected are high-degree vertices among themselves?

- The rich-club coefficient:

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

- ($N_{>k}$ = # of nodes with degree higher than k; $E_{>k}$ = # of links between these)

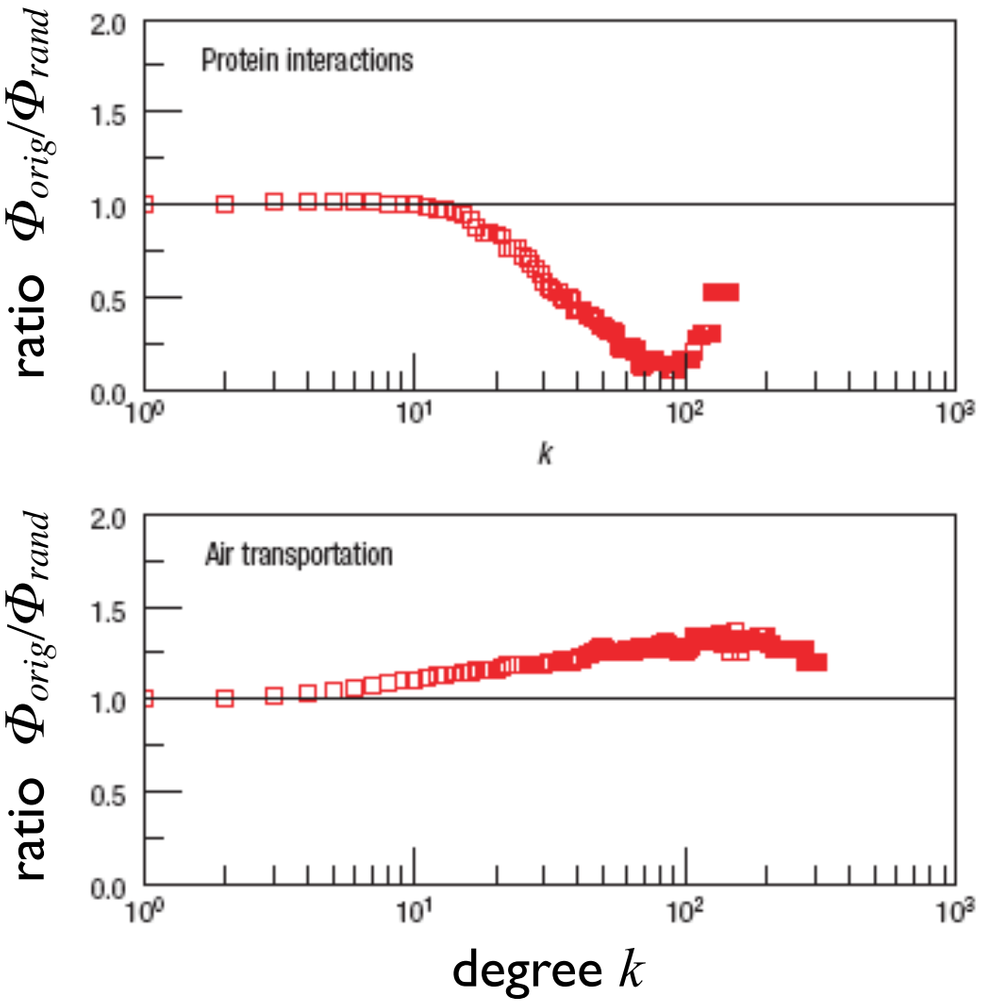
- Values should be compared to some random reference or null model

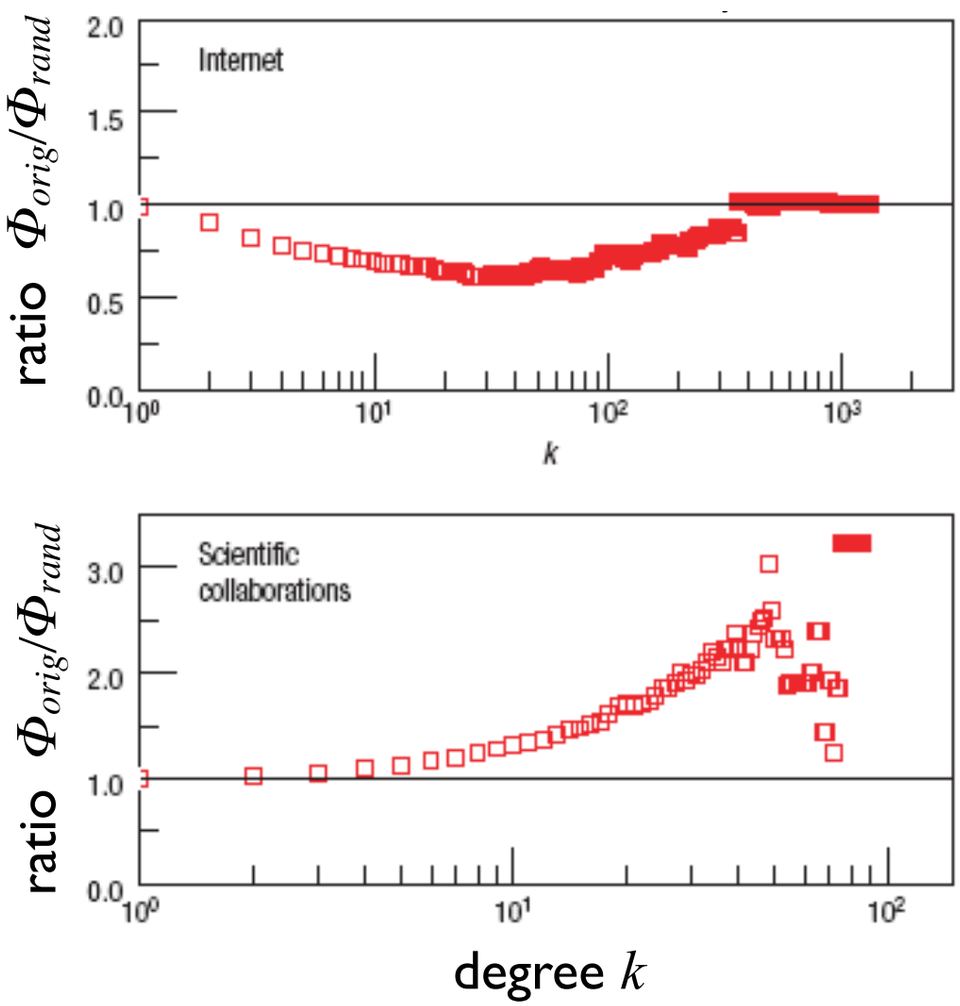
- Usually, the ***configuration model*** is used
  - Take as many nodes as in the original network, with exactly the same degrees, and connect randomly
  - Empirical networks: exchange endpoints of randomly chosen pairs of links until the whole network has been rewired, see lecture 2

- A network can have a dense rich-club and be disassortative at the same time!

The rich-club coefficient

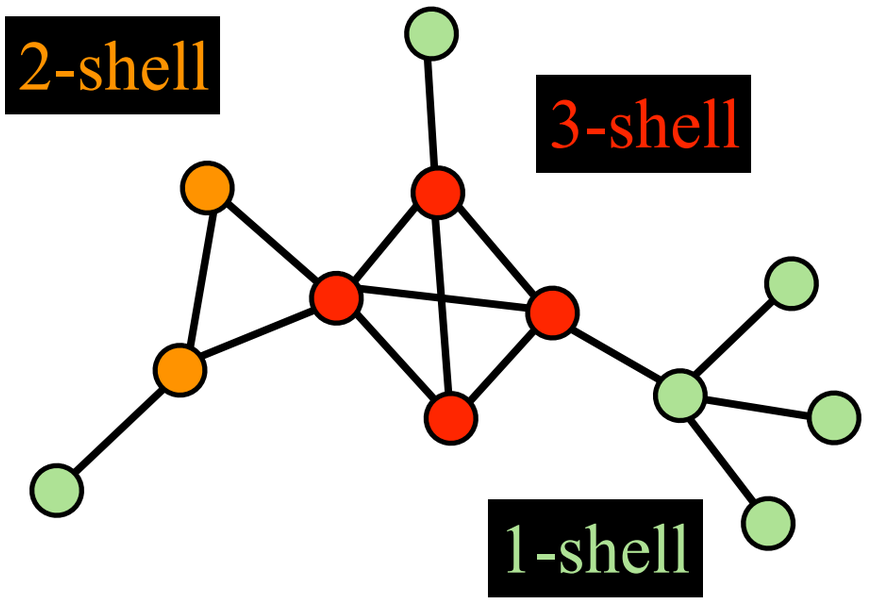
Colizza *et al,* Nature Physics **2**, 2006

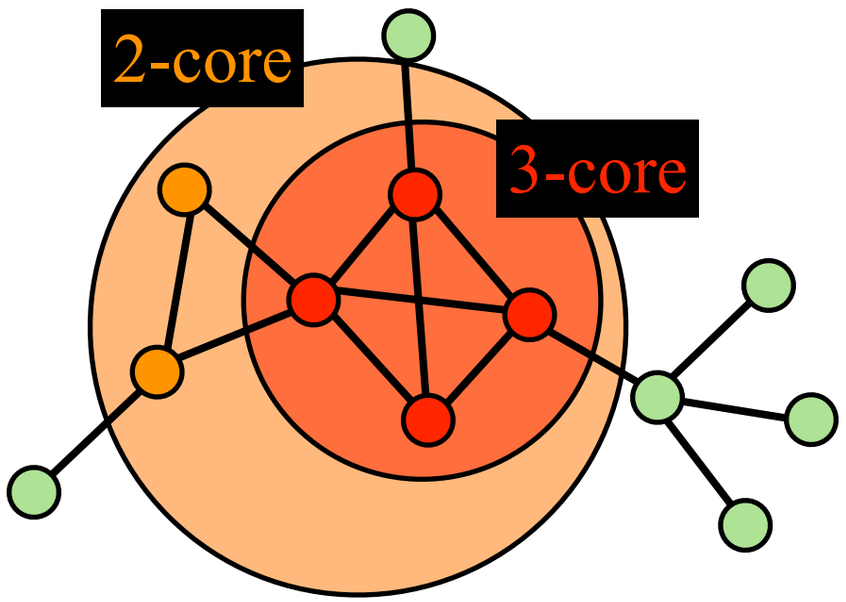




k-cores and *k*-shells

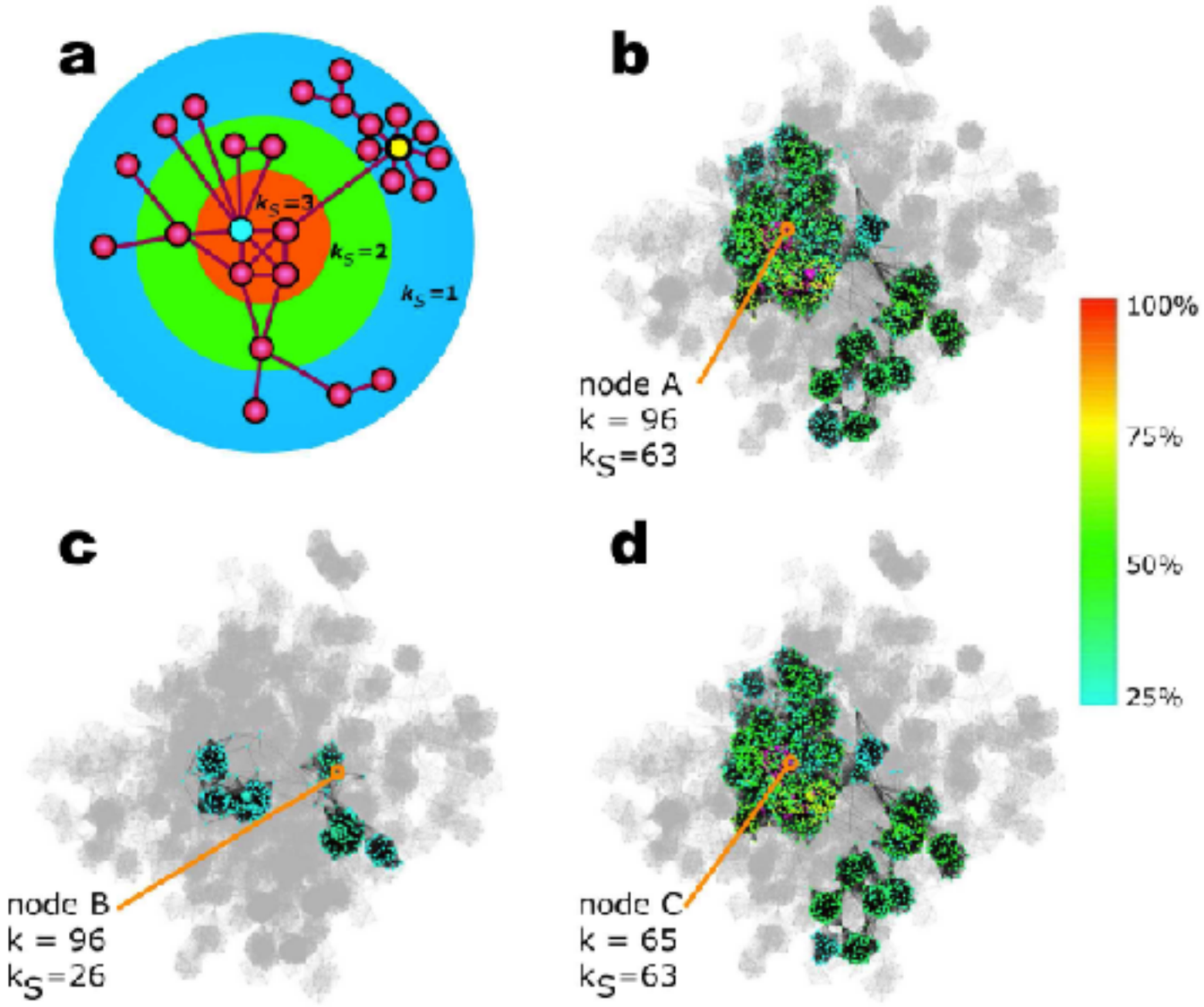
- Idea: to find dense "cores" of networks and measure the importance of nodes
  - Remove all nodes with degree $k=1$, and then all nodes which because of the removal now have $k\leq 1$ ($k=1$ or $k=0$), repeating the recursion until no more nodes can be removed
- Removed nodes form the 1-shell, and the remaining nodes the 2-core
- Continue by recursively removing nodes of degree $k\leq 2$, like in the first step, for getting the 2-shell and 3-core
- ...and so on, until all nodes removed





k-cores and *k*-shells

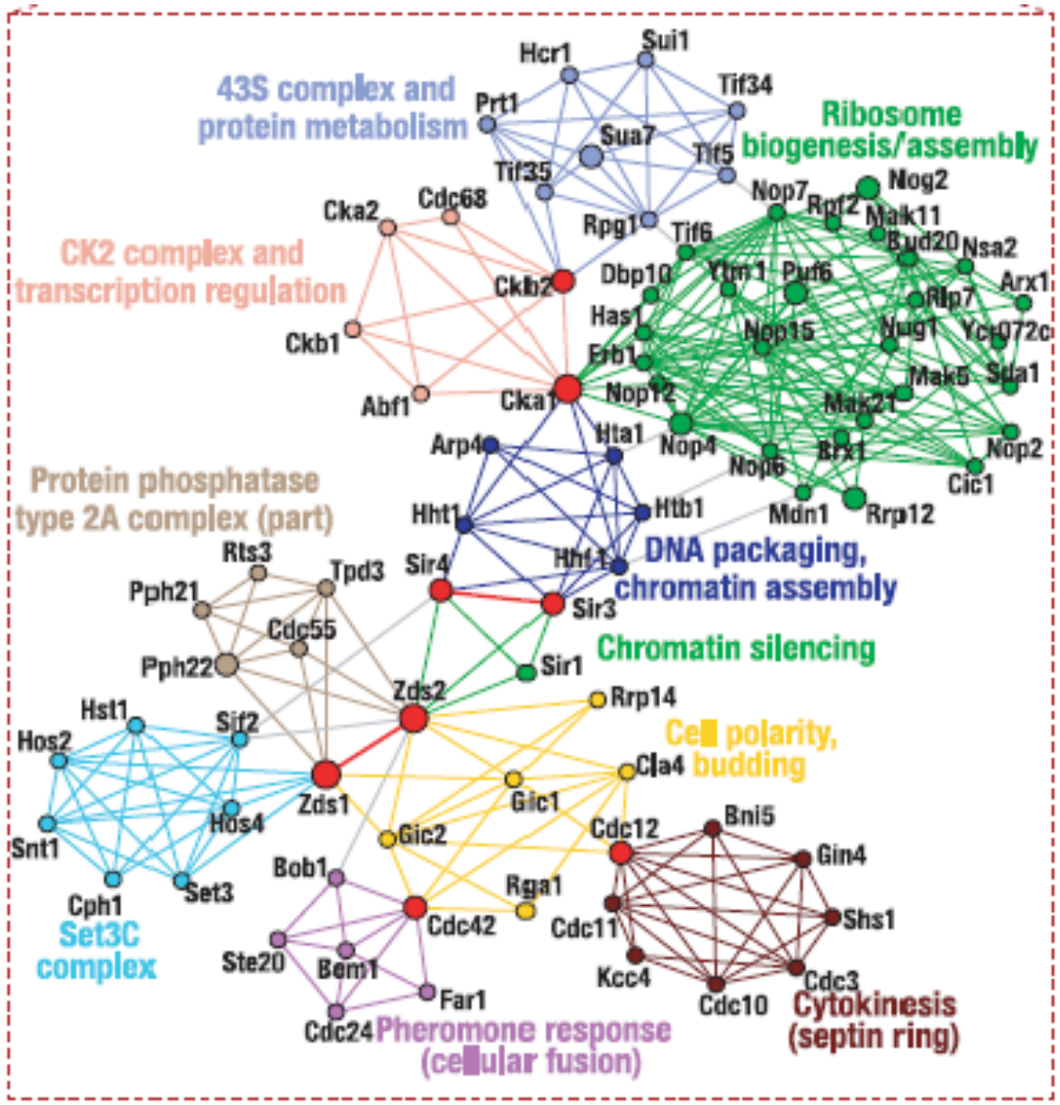
- The k-shell index k_s of a node measures its centrality
- A high value of the k-shell index indicates that the node is positioned centrally in the network, within densely connected sets of nodes
- It has been argued that such nodes are influential in spreading processes (disease, rumours, etc)



outcome of a spreading process when initiated on nodes with different degrees and k-shell indices

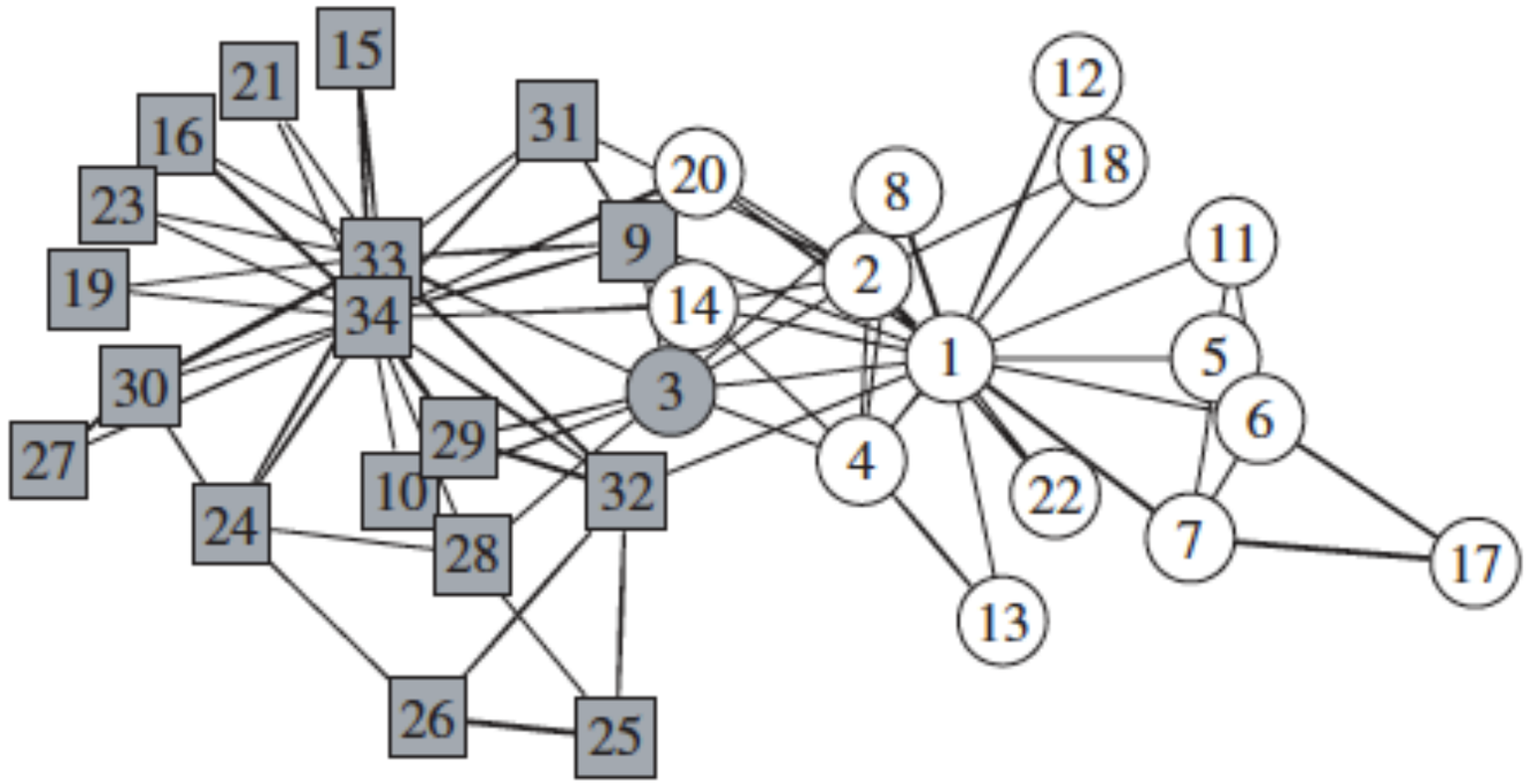
Kitsak *et al.*, Identifying influential spreaders in complex networks, Nature Physics **6**, 888 (2010)

Communities, clusters, modules



- sets of densely connected nodes, joined by sparse links
- ubiquitous in real-world networks
- e.g. social groups, functional modules in biological networks
- to be discussed in lecture 8

Example: community structure in Zachary's Karate club network



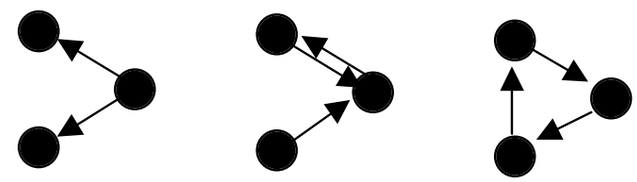
Subgraphs, motifs

- subgraph: any subset of nodes and the links joining them

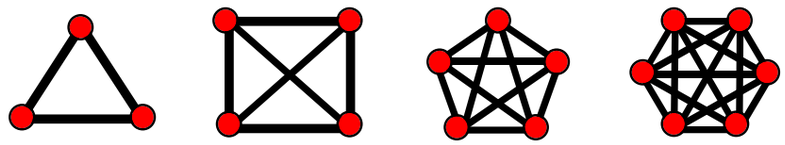
- k-clique: subgraph of k fully connected nodes

- **motif:** subgraph that occurs frequently

directed subgraphs of order 3



cliques



Counting motifs: z-score

number of times the subgraph M occurs in an empirical network

average number of times the subgraph M in randomized reference ensemble

$$Z_M = \frac{n_M - \left\langle n_M^{\text{rand}} \right\rangle}{\sigma_{n_M}^{\text{rand}}}$$

standard deviation of n_M in reference ensemble

- count subgraphs in original data
- repeat:
  i) *rewire original data with configuration model,*
  ii) *count subgraphs*
- compare original count to reference ensemble values

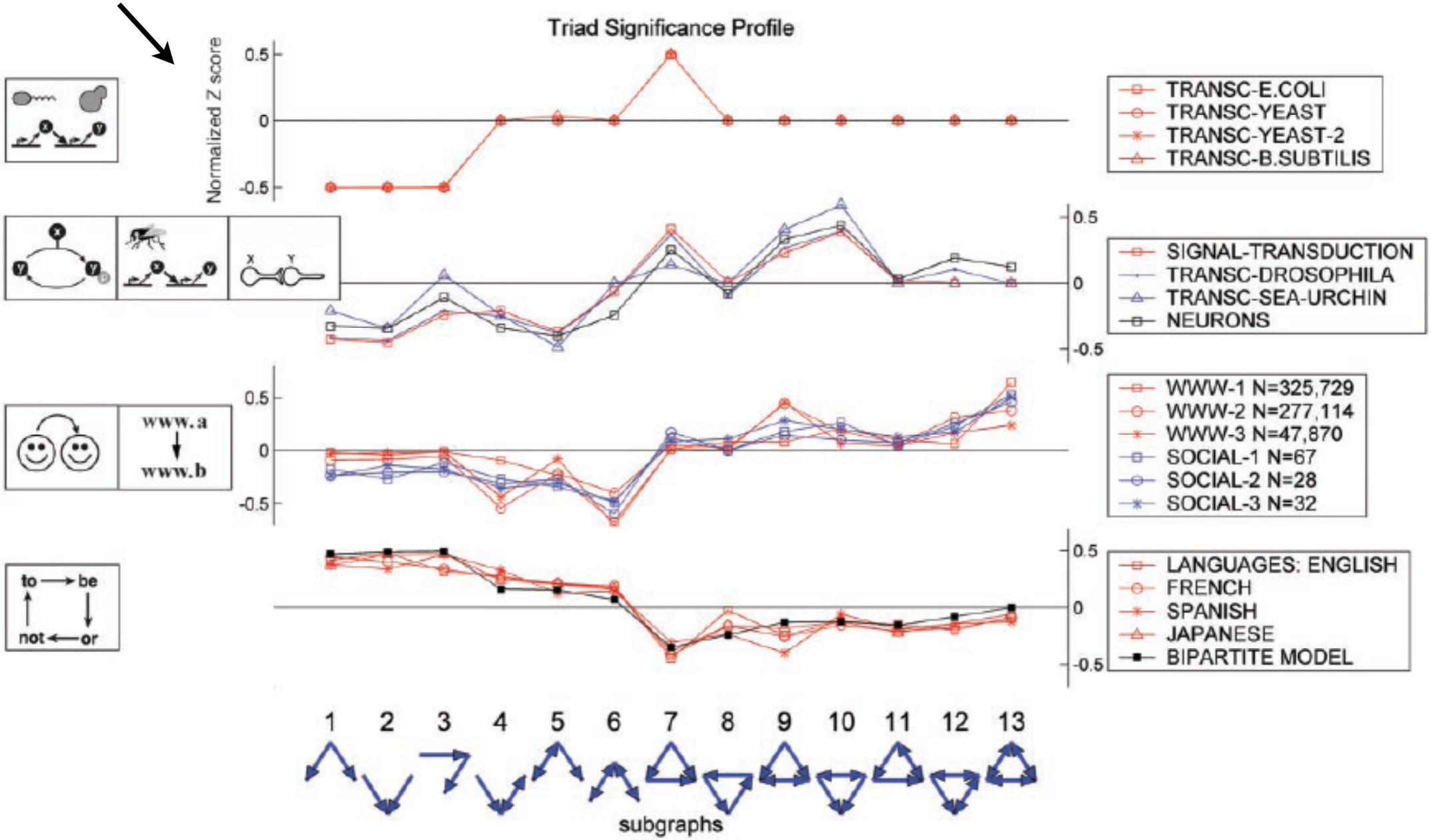
- subgraph frequency in configuration model depends on network size!

- for comparing different networks, one needs to normalize z-scores:

$$\mathrm{SP}_i = \frac{z_i}{\sqrt{\sum_i z_i^2}}$$

Motifs: "superfamilies" of networks

normalized z-scores



Milo et al., Science 303, 1538 (2004)

Limitations of Motif Analysis

- Larger systems will in general have less variance in motif counts
  - Comparing z-scores of two systems of different size is problematic
  - Larger system will generally have larger z-scores
- Large z-value means that the null hypothesis is wrong
  - Comparing the exact values of very large z-scores problematic
- Z-score assumes a normal distribution and not necessarily a good indicator for effect size

Spatial networks

- networks whose nodes "live" in an Euclidean space

- e.g. transport networks, power grids, river systems, brain connections

- every link is characterized by the Euclidean distance d_E between its endpoint nodes (or the distance one has to travel to cross the link, e.g. for roads)

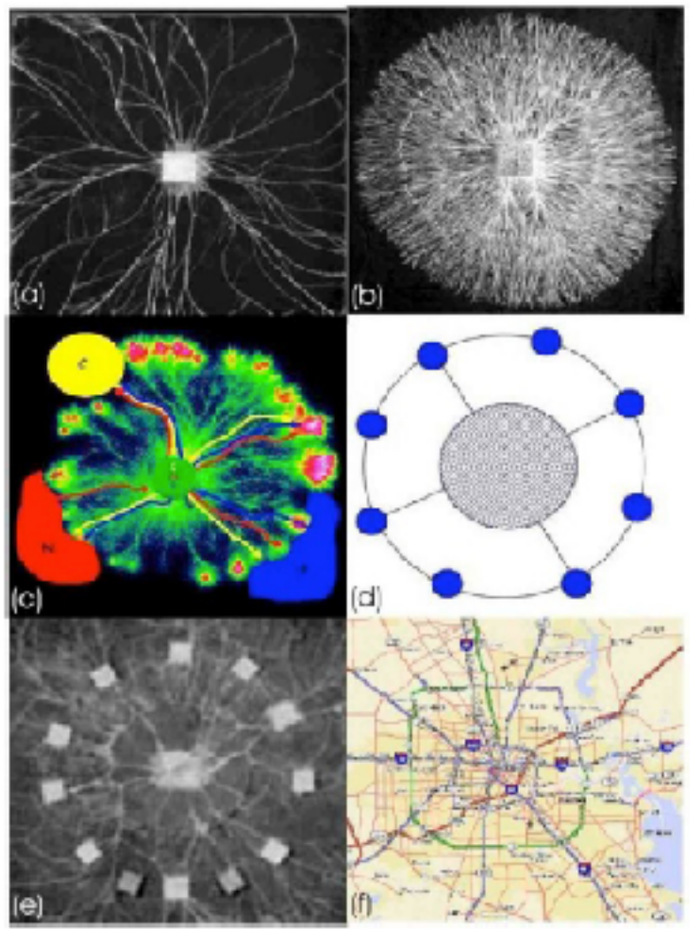


FIG. 57 Examples of hub-and-spoke structures with rings. (a c,e): Typical fungi networks, in (c) a schematic representation of the nutrient flow is shown. (d) The model studied in [19, 112, 171] with spokes radiating from a hub. (f) Road network in Houston showing an inner hub with a complicated structure. From [171].

see M. Barthélemy, Spatial Networks, Physics Reports **499**:1-101 (2011)

Spatial networks: route factor (detour index)

graph distance

$$Q(i,j) = \frac{d_R(i,j)}{d_E(i,j)}$$

Euclidean distance

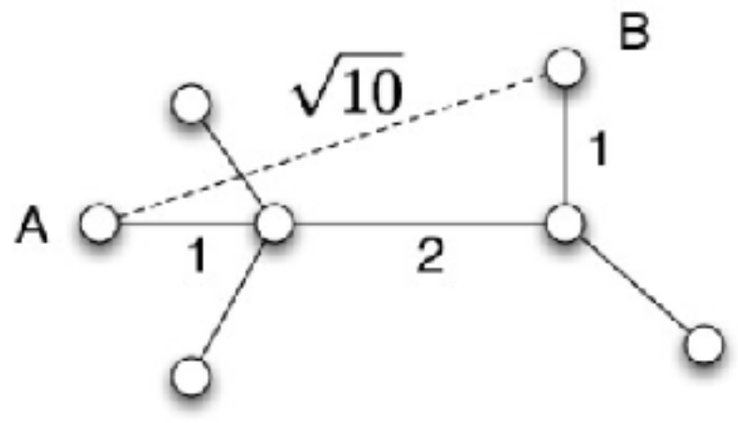


FIG. 5 Example of a detour index calculation. The 'as crow flies' distance between the nodes A and B is $d_E(A,B) = \sqrt{10}$ while the route distance over the network is $d_R(A,B) = 4$ leading to a detour index equal to $Q(A,B) - 4/\sqrt{10} \simeq 1.265$.

- the average route factor for a single node

$$\langle Q(i) \rangle = \frac{1}{N-1} \sum_j Q(i,j)$$

measures its accessibility

- the network average measures its efficiency

$$\langle Q \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} Q(i,j)$$

(the closer to one = the more efficient)

Spatial networks: wiring cost

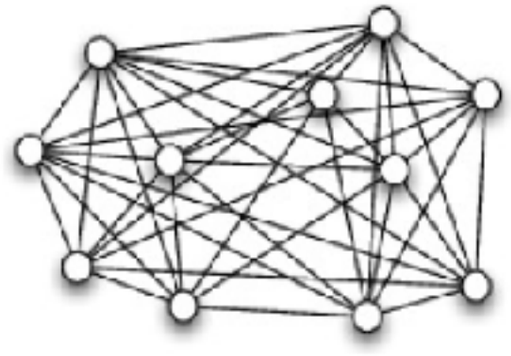
- The Minimum Spanning Tree is the tree linking all nodes that minimizes the length of links,

$$\ell_T = \sum_{e \in E} d_E(e)$$

- One can estimate the "wiring cost" of a network by comparing the sum of link lengths to that of the MST:

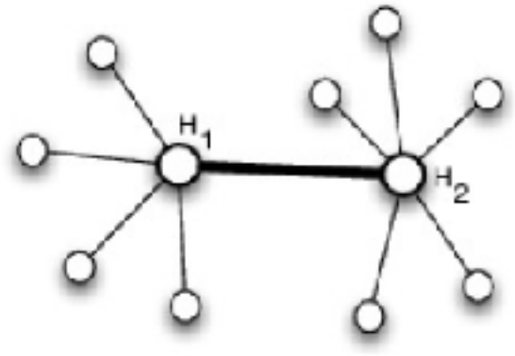
$$C = \frac{\ell_T}{\ell_T^{MST}}$$

low route factor, high cost



Point-to-point

high route factor, low cost



Hub-and-spoke

Further reading

- L.F. Da Costa et al: *Characterization of Complex Networks: A survey of measurements*, Advances in Physics **56**, 167 - 242 (2007), preprint at http://arxiv.org/abs/cond-mat/0505185

- M.E.J. Newman, *The Structure and Function of Networks*, SIAM Review 45, 167-256 (2003), preprint at http://arxiv.org/abs/cond-mat/0303516

- Ed Bullmore & Olaf Sporns: *Complex brain networks: graph theoretical analysis of structural and functional systems*, Nature Reviews Neuroscience **10**, 186-198 (2009)

- C.J. Stam & J.C. Reijneveld: *Graph theoretical analysis of complex networks in the brain*, Nonlinear Biomed Phys. 2007 Jul 5;1(1):3

- M. Barthélemy, *Spatial Networks*, Physics Reports **499**:1-101 (2011)