CENTRALITIES

Carlo PICCARDI

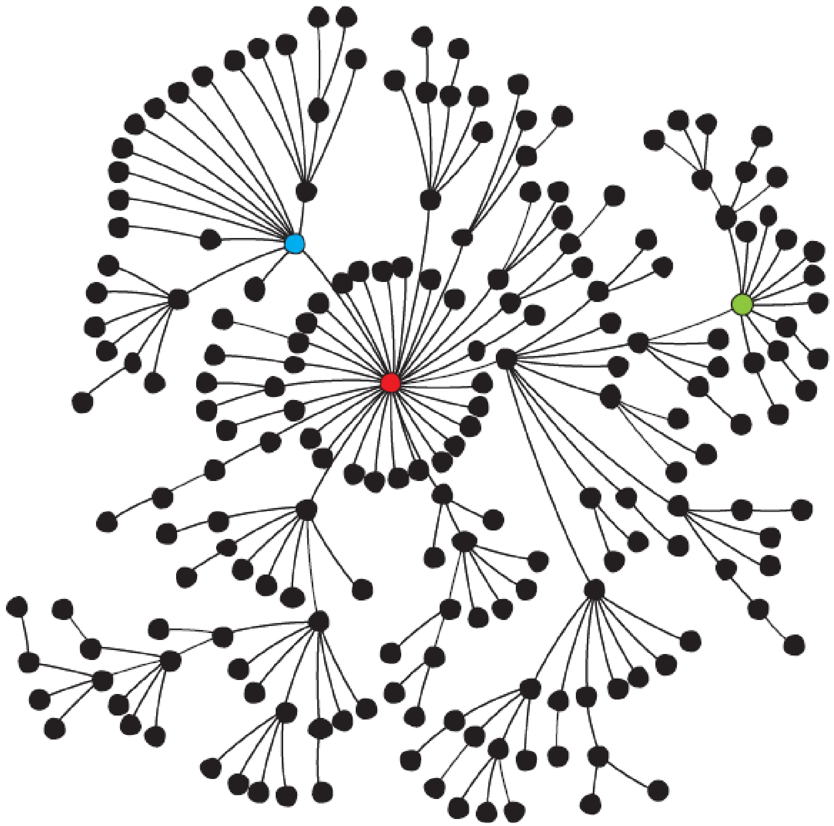
DEIB - Department of Electronics, Information and Bioengineering
Politecnico di Milano, Italy

email carlo.piccardi@polimi.it
https://piccardi.faculty.polimi.it

NODE CENTRALITY

The centrality of a node is a measure of its importance in the network.

Degree



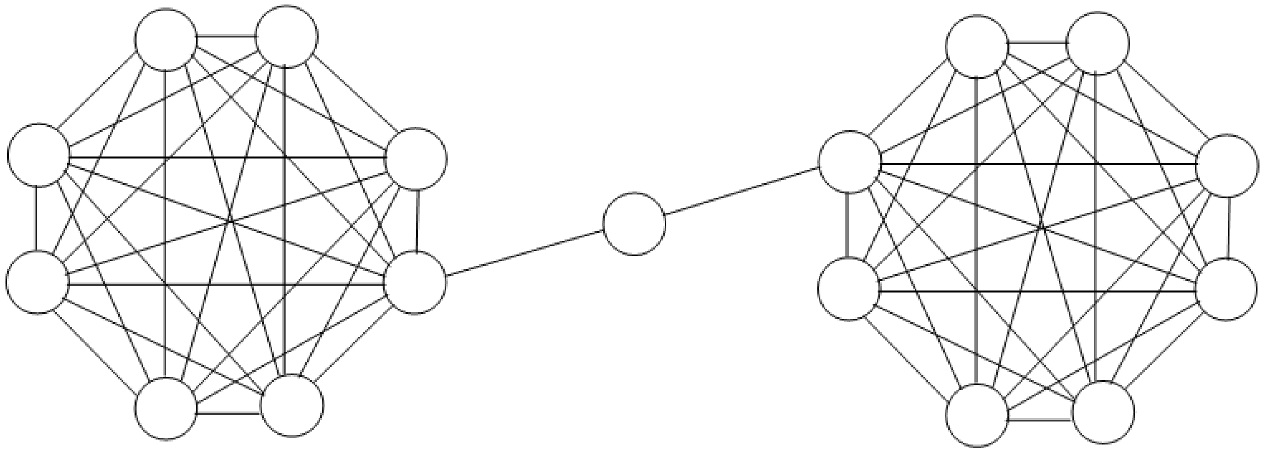
The importance of a node can trivially be captured by the number k_i of its neighbors (i.e. interactions, communication channels, sources-destinations of information, etc.).

⇒ The "hubs" are the most central nodes.

In weighted networks, use the strength s_i.

Betweenness

The degree centrality may fail in some cases...



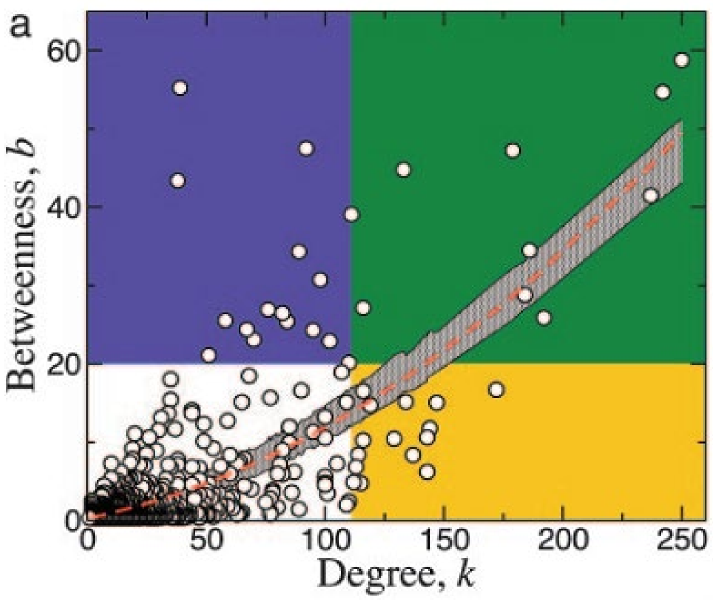
The betweenness b_i of node i is the number of shortest paths (connecting all the pairs of nodes of the network) that pass through i.

$$b_i = \sum_{j,k} \frac{\text{n. of shortest paths connecting } j,k \text{ via } i}{\text{n. of shortest paths connecting } j,k} = \sum_{j,k} \frac{n_{jk}(i)}{n_{jk}}$$

Similar definition for link betweenness.

Anomalous nodes might emerge when comparing degree and betweenness.

Example: the worldwide air transportation network (Guimerà et al., 2005)



There are anomalous cities (=nodes) with very low degree but very high betweenness.

Table 2. The 25 most central cities in the worldwide air transportation network

Rank	City	b	b/b_{ran}	Degree
1	Paris	58.8	1.2	250
2	Anchorage*	55.2	16.7	39
3	London	54.7	1.2	242
4	Singapore*	47.5	4.3	92
5	New York	47.2	1.6	179
6	Los Angeles	44.8	2.3	133
7	Port Moresby*	43.4	13.6	38
8	Frankfurt	41.5	0.9	237
9	Tokyo	39.1	2.7	111
10	Moscow	34.5	1.1	186
11	Seattle*	34.3	3.3	89
12	Hong Kong*	30.8	2.6	98
13	Chicago	28.8	1.0	184
14	Toronto	27.1	1.8	116
15	Buenos Aires*	26.9	3.2	76
16	São Paulo*	26.5	2.8	82
17	Amsterdam	25.9	0.8	192
18	Melbourne*	25.5	4.5	58
19	Johannesburg*	25.4	2.6	84
20	Manila*	24.4	3.5	67
21	Seoul*	24.3	2.1	95
22	Sydney*	23.1	3.2	70
23	Bangkok*	22.9	1.8	102
24	Honolulu*	21.1	4.4	51
25	Miami*	20.1	1.4	110

Cities are ordered according to their normalized betweenness. We also show the ratio of the actual betweenness of the cities to the betweenness that they have after randomizing the network.

*These cities are not among the 25 most connected.

Closeness centrality

A node is central if, on average, it is close (=short distance) to all other nodes: it has better access to information, more direct influence on other nodes, etc.

The average distance from i to all the other nodes is:

$$l_i = \frac{1}{N-1} \sum_j d_{ij}$$

The closeness centrality is defined as

$$c_i = \frac{1}{l_i} = \frac{N-1}{\sum_j d_{ij}}$$

If the network is directed, we must distinguish between in- and out-closeness.

If the network is weighted, several (non trivial) generalized definitions are available.

Eigenvector centrality

The centrality γ_i is (proportional to) the sum of the centralities of the neighbors (i.e., a node is important if it relates to important nodes).

$$\gamma_i = \alpha \sum_j a_{ij} \gamma_j$$

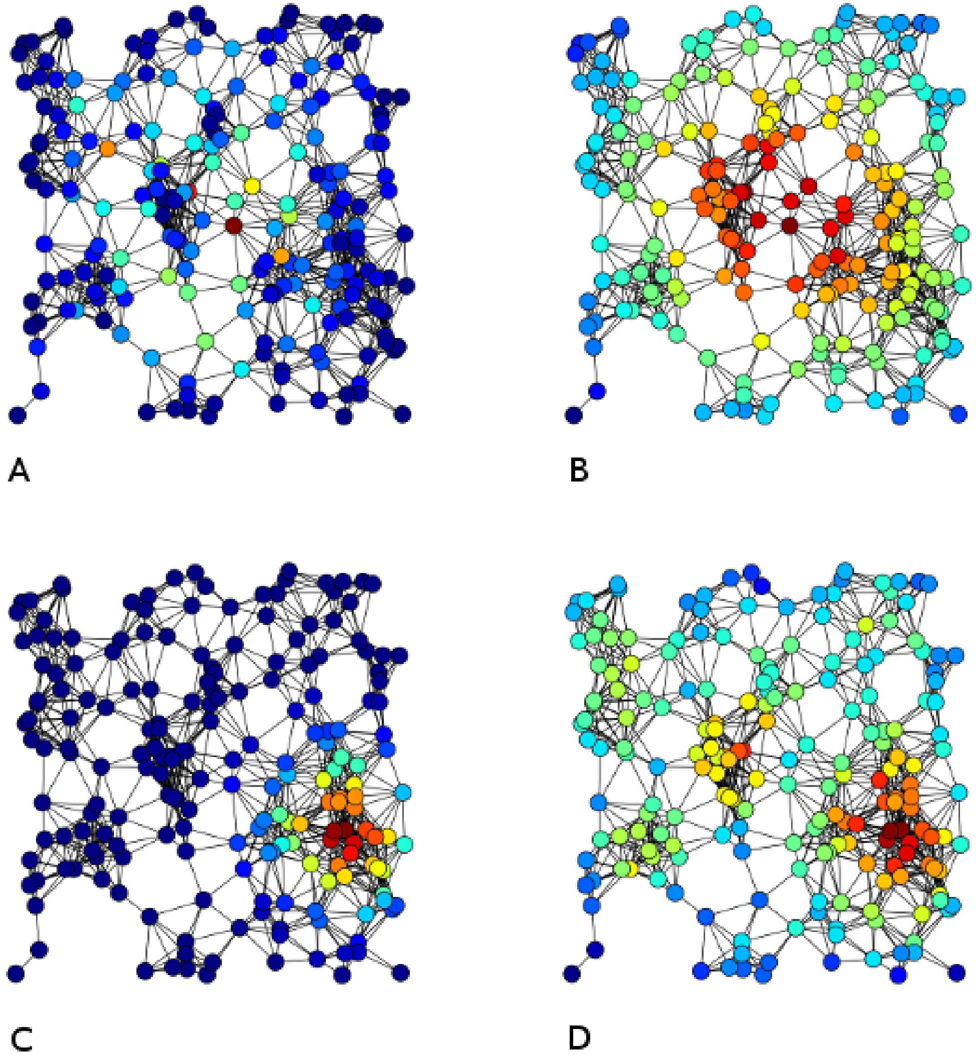
Letting $\gamma = [\gamma_1 \;\; \gamma_2 \;\; \ldots \;\; \gamma_N]^T$ and $\lambda = 1/\alpha$, we obtain the eigenvector equation

$$A\gamma = \lambda\gamma$$

If the network is connected (= A is irreducible), the centralities γ_i are given by the only solution with $\lambda > 0$, $\gamma_i > 0$ for all i (Frobenius-Perron theorem).

- applications in social networks (*who is the most influential individual?*)
- applications in web searching (with some modifications: *Google "PageRank" - which is the most important webpage*?)
- another modification is Katz (or alpha-) centrality: $\gamma_i = \alpha \sum_j a_{ij} \gamma_j + \beta$

betweenness closeness



from wikipedia.org

eigenvector degree

Authorities and Hubs

In directed networks, we can take into account the different role of in- and out-links.

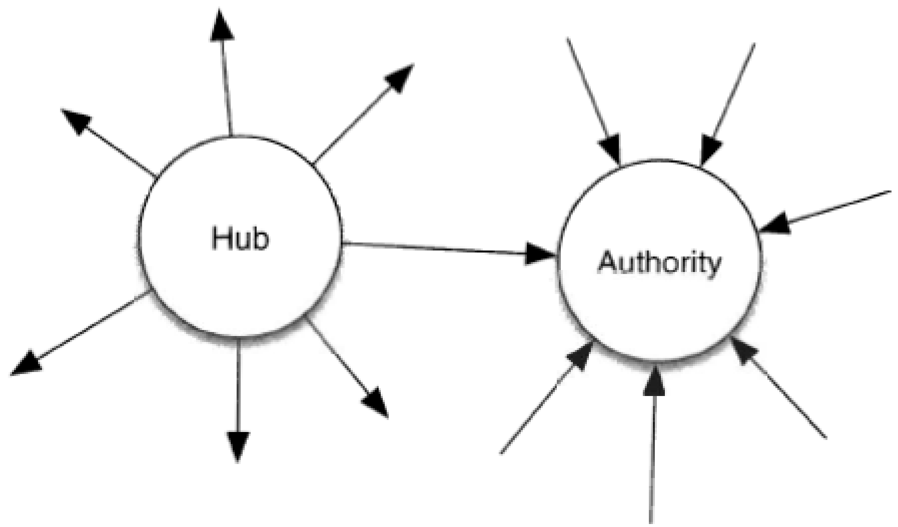
"authority" score x_i : a node with large x_i *is pointed by* highly ranked nodes

"hub" score y_i : a node with large y_i *points to* highly ranked nodes

$$x_i = \alpha \sum_j a_{ji} y_j \qquad\qquad y_i = \beta \sum_j a_{ij} x_j$$

For example, in the World Trade Network:

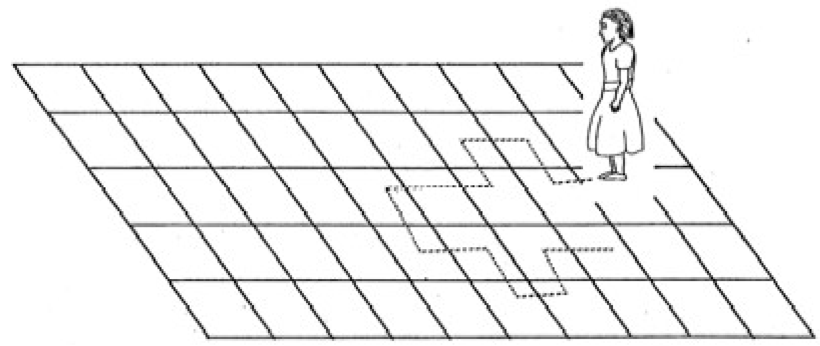
- "authorities" (= nodes with large x_i) are countries with large import flows ("consumers")
- "hubs" (= nodes with large y_i) are countries with large export flows ("producers")



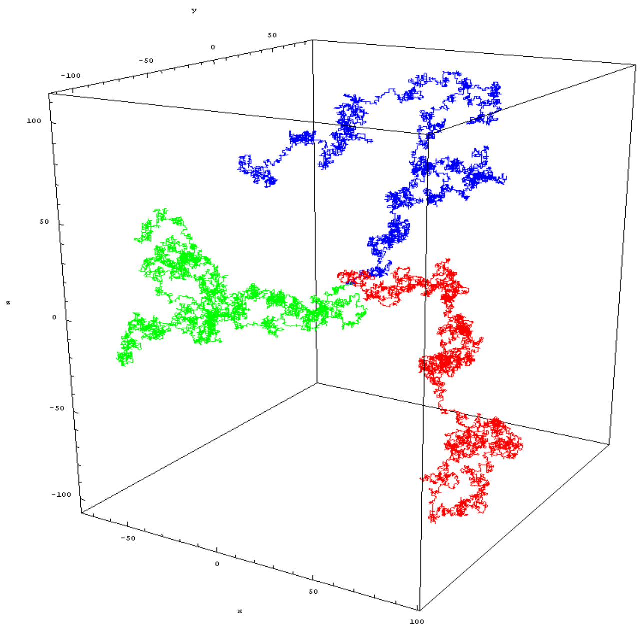
RANDOM WALKS ON NETWORKS

A random walk is a path formed by a sequence of random steps.

The term is first attributed to Karl Pearson *[Nature, 1905]*.



Applications in ecology, economics, psychology, computer science, physics, chemistry, biology, etc.



Many variants:

- discrete vs continuous time
- uniform vs non-uniform step
- Markovian vs non-Markovian process
- etc.

Random walks on networks

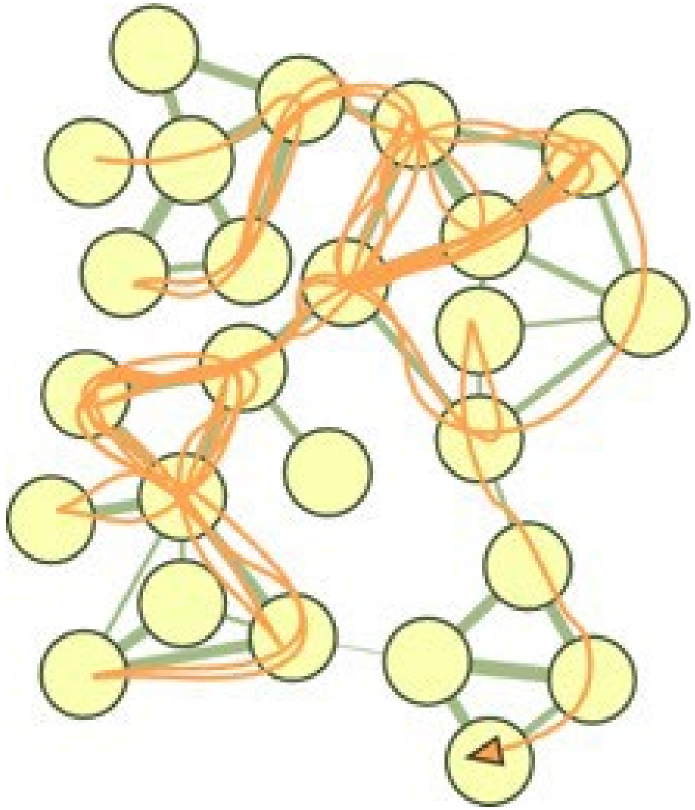
In a binary (unweighed) network, the random walker in node i chooses an out-link $i \to j$ with uniform probability:

$$p_{ij} = \frac{a_{ij}}{k_i^{out}}$$

In a weighted network, the out-link is chosen with probability proportional to its weight:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} = \frac{w_{ij}}{s_i^{out}}$$

$P = [p_{ij}]$ is the $N \times N$ transition matrix.



Random walks and Markov chains

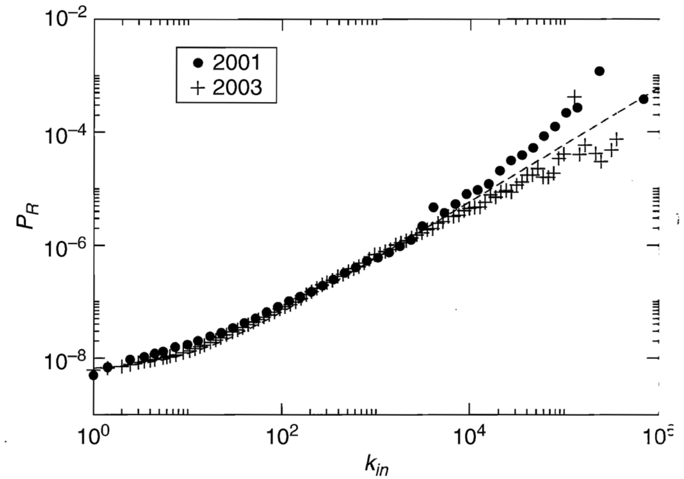
$\pi_{i,t}$ = state probability = probability of being in node i at time t ($\sum_i \pi_{i,t} = 1 \ \forall t$)

$\pi_t = (\pi_{1,t} \quad \pi_{2,t} \quad \cdots \quad \pi_{N,t})$ evolves according to the Markov chain equation

$$\pi_{t+1} = \pi_t P \quad , \quad \pi_{i,t+1} = \pi_{1,t} p_{1i} + \pi_{2,t} p_{2i} + \cdots + \pi_{N,t} p_{Ni}$$

If the network is strongly connected ⇒
⇒ the transition matrix $P = [p_{ij}]$ is irreducible ⇒
⇒ there exists a unique stationary state probability distribution $\pi = \pi P$, which is strictly positive ($\pi_i > 0$ for all i).

π_i = fraction of time spent on node i
= centrality of node i



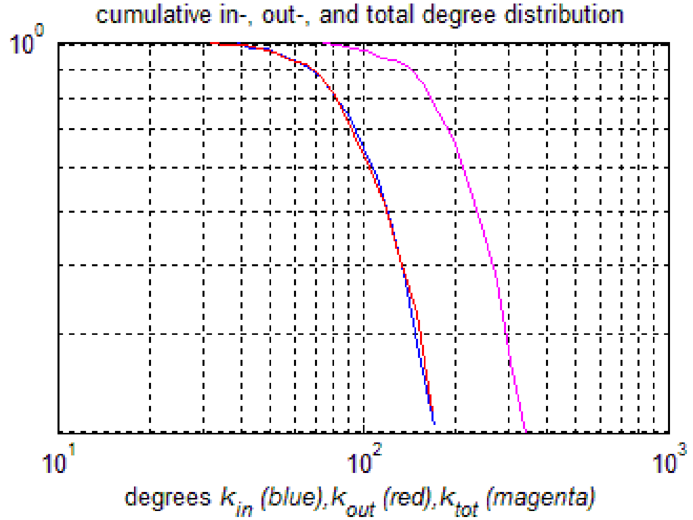
In undirected networks, π_i is the (rescaled) node strength $\pi_i = s_i / \sum_j s_j$

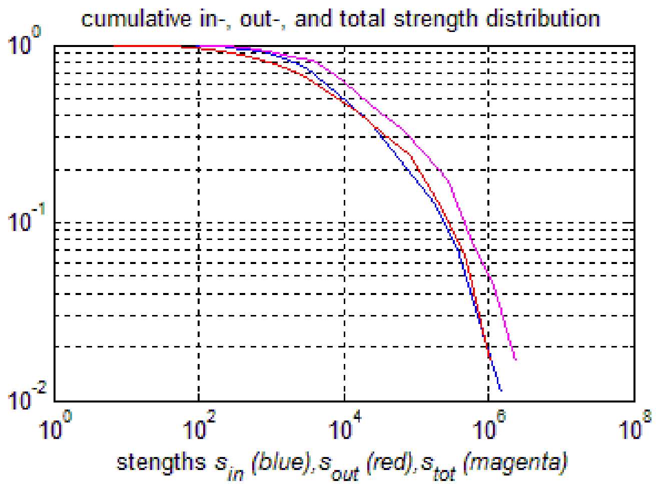
In directed networks, π_i turns out to be mostly correlated to the in-strength s_i^{in} *(example: WWW).*

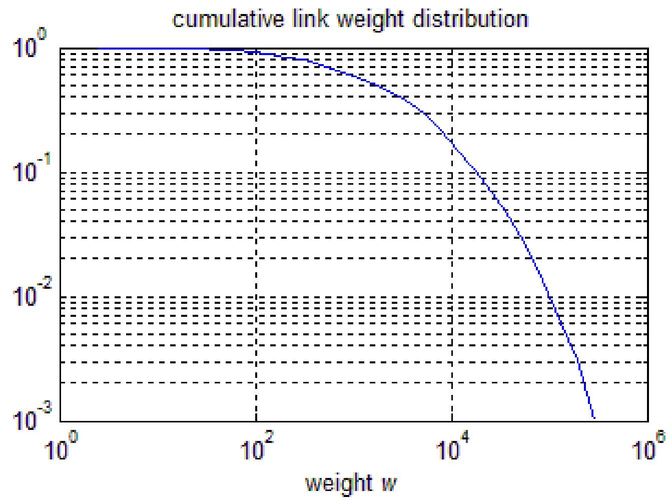
Example: the World Trade Network (2008)

The trading system can be modelled as a directed, weighted network: w_{ij} is the export flow (million US dollars) from country i to country j

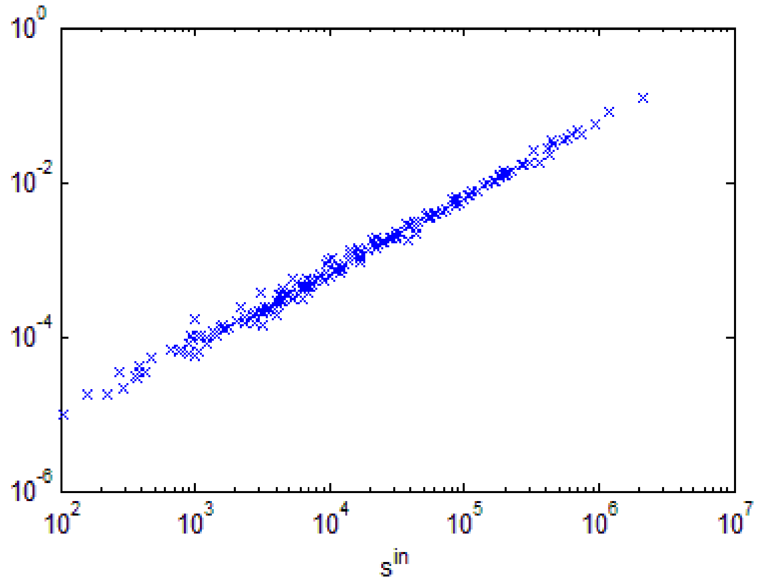
- The strongly connected component includes $N = 181$ countries (94% of the total).
- The network is extremely dense ($\frac{L}{N(N-1)} = 0.65$) ...
- ... and very heterogeneous (multi-scale) in node degrees, node strengths, and link weights.

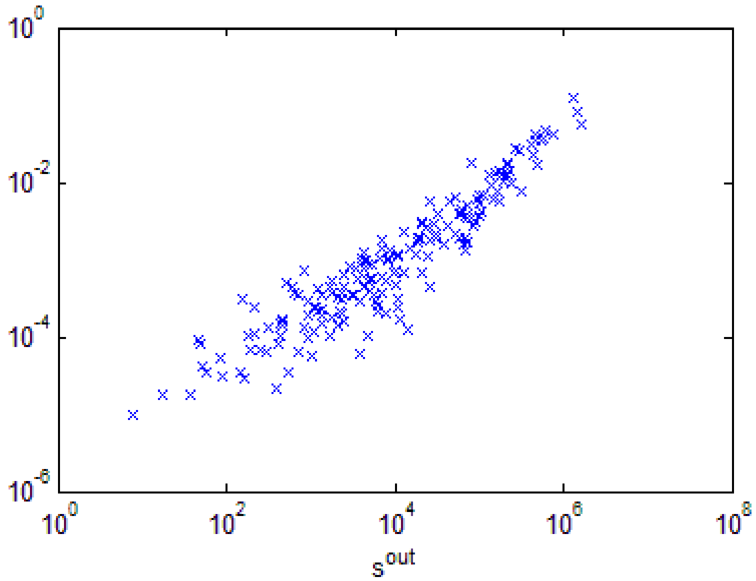






In the WTN, centrality π_i strongly correlates with the in-strength s_i^{in}...





...but it also fairly correlates with the out-strength s_i^{out} (because the latter correlates with s_i^{in}).

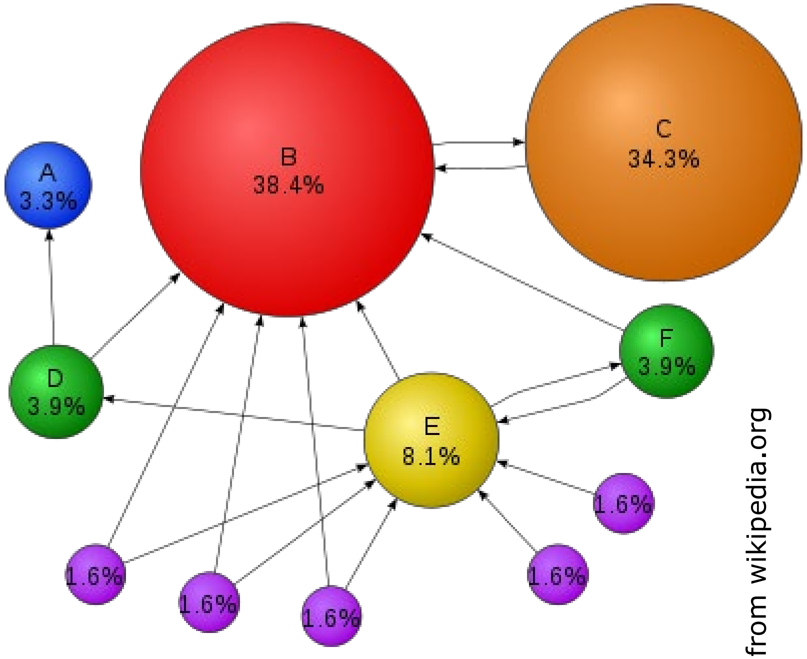
PageRank ("Google") centrality

Most directed networks are not connected.

The solution to $\pi = \pi P$ is not unique or non positive, or the Markov chain might be not even well defined.

Teleportation: at each time step, the random walker has probability $\gamma > 0$ to jump to a randomly selected node.

$$p_{ij} \rightarrow p'_{ij} = (1-\gamma)\frac{w_{ij}}{s_i^{out}} + \gamma\frac{1}{N}$$



from wikipedia.org

The network becomes complete (all-to-all) thus connected $\Rightarrow$ there exists a unique strictly positive solution to $\pi = \pi P'$.

π_i = PageRank of node i

A proper value for γ ?

Not too large (the network would be heavily modified) nor too small (π_i too sensitive to γ).

The standard (Google) value is $\gamma = 0.15$.

The WWW is an extremely heterogeneous network with self-organized structure.

Google ranking exploits the network structure for retrieving information.

Examples of PageRank values *(in logarithmic scale 1-10 – e.g. checkpagerank.net):*

RISULTATI

apple.com Google PageRank: 9/10

bbc.co.uk Google PageRank: 9/10

repubblica.it Google PageRank: 8/10

polimi.it Google PageRank: 6/10

Technical problem: (re)computing π_i in a network with trillions of nodes.

Centrality name	Characteristics of a central node	Equation
Degree (DC)	Connected to many other nodes [3]	$DC_i = d_i = \sum_{j\neq i} A_{ij}$
Eigenvector (EC)	Connected to many other nodes and/or to other high-degree nodes [40]	$EC_i = \frac{1}{\lambda_1}\sum_j A_{ji} v_j$
Katz (KC)	Connected to many other nodes and/or connected to other high-degree [41]	$KC_i = \alpha \sum_j A_{ji} v_j + \beta$
PageRank (PR)	Connected to many other nodes and connected to other high-degree nodes [42]	$PR_i = \alpha \sum_j A_{ji} \frac{v_j}{k_j} + \beta$
Leverage (LC)	Has a higher degree than its neighbours [43]	$LC_i = \frac{1}{d_i}\sum_{j\in h(i)} \frac{d_i - d_j}{d_i + d_j}$
H-index (HC)	Connected to many other high-degree nodes [44]	$HC_i = \max_{1\le h \le d_i} \min(\lvert\mathcal{N}_{\ge h}(i)\rvert, h)$
Laplacian (LAPC)	Removal of this node would greatly impair the network [45,46]	$LAPC_i = d_i^2 + d_i + 2\sum_{j\in\mathcal{N}(i)} d_j$
Shortest-path Closeness (CC)	Low average shortest path length to other nodes in the network [47]	$CC_i = \frac{N}{\sum_j l_{ij}}$
Subgraph (SC)	Involved in many closed short-range walks [48]	$SC_i = [e^A]_{ii}$
Participation coefficient (PC)	Connections distributed across different topological modules [24]	$PC_i = 1 - \sum_{m=1}^{M}\left(\frac{d_i(m)}{d(i)}\right)^2$
Total Communicability (TCC)	Can be easily reached by a walk from any other node [21]	$TCC_i = \sum_j [e^A]_{ji}$
Random-walk Closeness (RWCC)	Can be easily reached by a random-walk from any other node [49,50]	$RWCC_i = \frac{N}{\sum_j H_{ji}}$
Information (IC)	Can be easily reached by paths from other nodes [51]	$IC_i = \left(C_{ii} + \frac{\sum_j C_{jj} - 2\sum_j C_{ij}}{N}\right)^{-1}$
Shortest-path Betweenness (BC)	Lies on many shortest topological paths linking other node pairs [3]	$BC_i = \sum_{p\neq i, p\neq q, q\neq i} \frac{g_{pq}(i)}{g_{pq}}$
Communicability betweenness (CBC)	Takes part in many walks between pairs of other nodes [52]	$CBC_i = \frac{1}{\acute{C}}\sum_p\sum_q \frac{G_{piq}}{G_{pq}}, p\neq q, q\neq i$
Random-walk Betweenness (RWBC)	Takes part in many random walks between pairs of other nodes [53]	$RWBC_i = \frac{\sum_{p<q} I_i^{(pq)}}{\frac{1}{2}N(N-1)}$
Bridging (BridC)	Forms key links between high degree nodes [54]	$BridC_i = BC_i \times Bc_i$

A = adjacency matrix; d_i = degree of node i; λ_1 = leading eigenvalue of A; v = leading eigenvector of A; α = penalty on distant connections to a node's centrality score; β = preassigned centrality constant; $h(i)$ = the neighbours of node i; $\mathcal{N}_{\ge h}(i)$ = neighbours of node i which have at least a degree of h; N = number of nodes in a network; l_{ij} = length of the shortest between nodes i and j; e^A = matrix exponential of A; M = number of modules in a network; $d_i(m)$ = neighbours of node i which are part of module m; H = the matrix of mean-first passage times between nodes in a network; $C = (L+J)^{-1}$ where L is the Laplacian of A and J is a $N\times N$ matrix with all elements equal to one; g_{pq} = the number of shortest-paths between nodes p and q; $g_{pq}(i)$ = the number of shortest-paths between nodes p and q which pass through i; G_{pq} = number of walks between nodes p and q; G_{piq} = number of walks between nodes p and q involving node i; $\acute{C} = (N-1)^2-(N-1)$ which is a normalisation term; $I_i^{(pq)}$ = current flowing through nodes p and q which passes through node i; $Bc_i = d_i^{-1}/\sum_{j\in\mathcal{N}(i)} d_j^{-1}$. All measures here are defined for unweighted networks, see S1 Text for information on weighted versions.

https://doi.org/10.1371/journal.pone.0220061.t001

Many other **centrality measures** have been proposed…