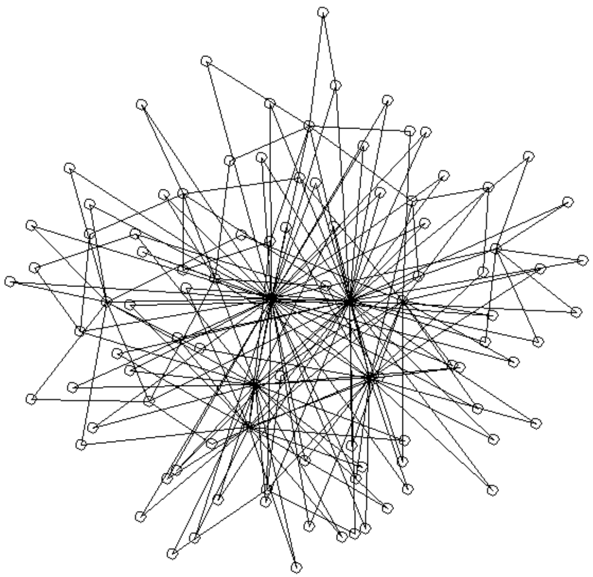
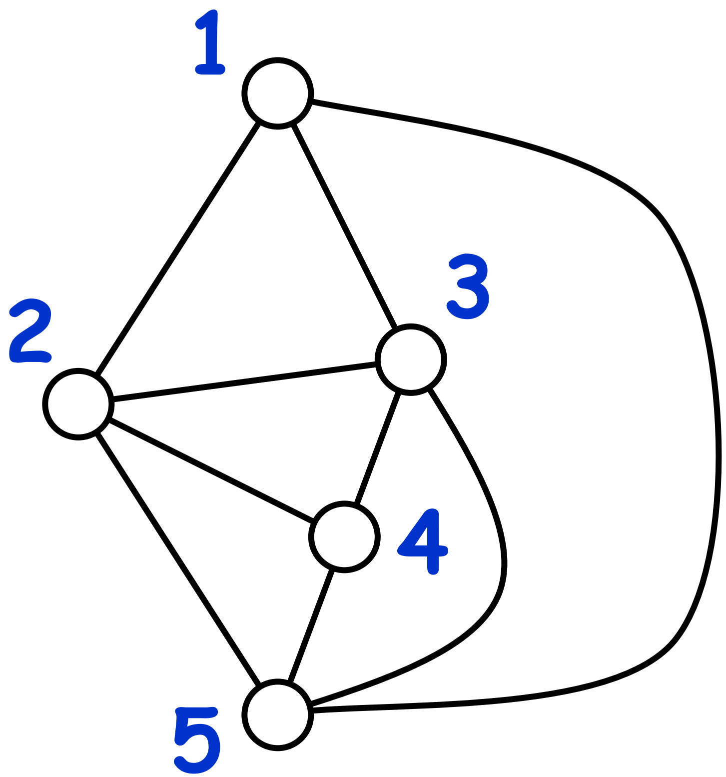
Basics of Network Analysis



Hiroki Sayama
sayama@binghamton.edu

Graph = Network

- G(V, E): graph (network)
  V: vertices (nodes), E: edges (links)



Nodes = 1, 2, 3, 4, 5

Links =
1<->2, 1<->3, 1<->5,
2<->3, 2<->4, 2<->5,
3<->4, 3<->5, 4<->5

(Nodes may have states; links may have directions and weights)

Representation of a network

- **Adjacency matrix:**

  A matrix with rows and columns labeled by nodes, where element a_{ij} shows the number of links going from node i to node j

  (becomes symmetric for undirected graph)

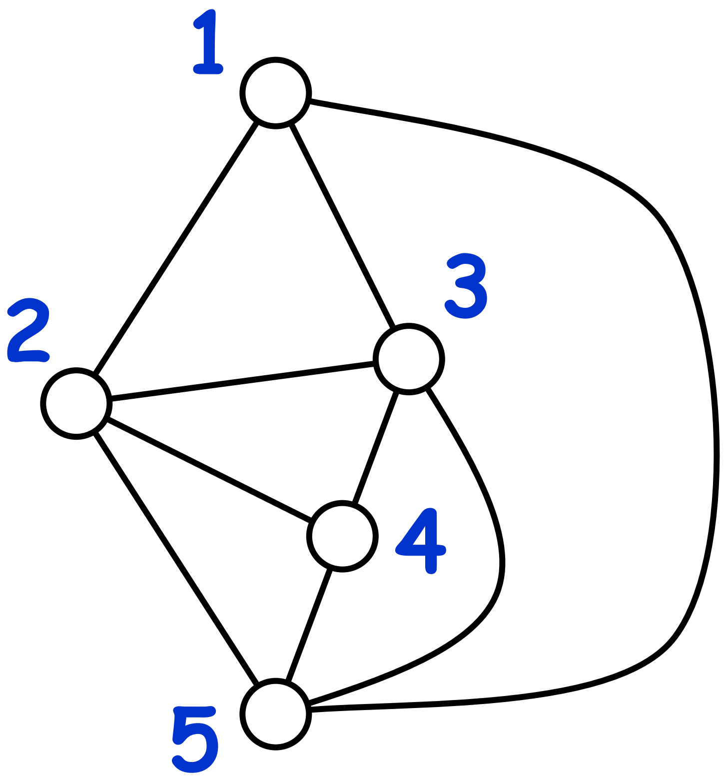
- **Adjacency list:**

  A list of links whose element "i->j" shows a link going from node i to node j

  (also represented as "i -> $\{j_1, j_2, j_3, \ldots\}$")

Exercise

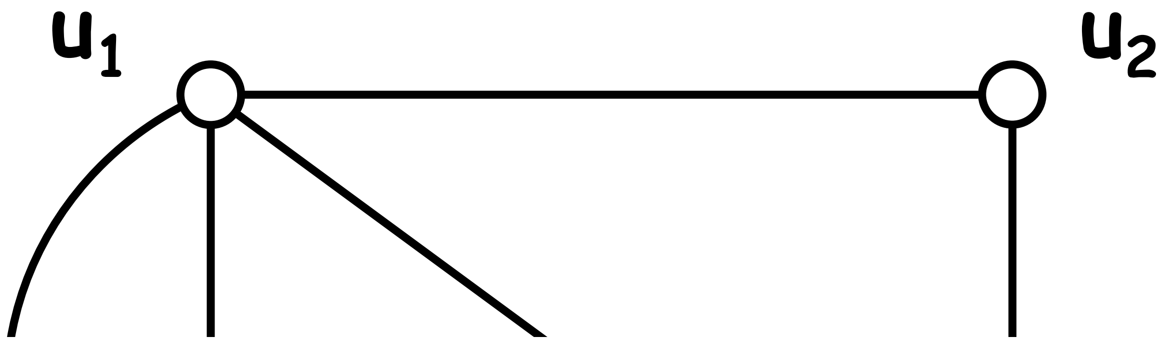
- Represent the following network in:



  - Adjacency matrix
  - Adjacency list

Degree of a node

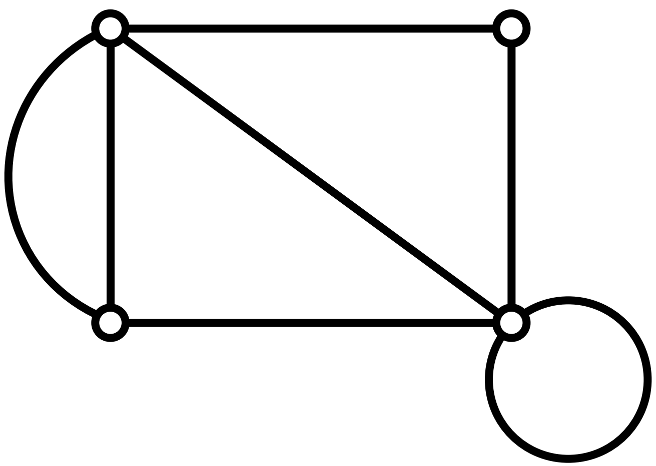
- A degree of node u, deg(u), is the number of links connected to u

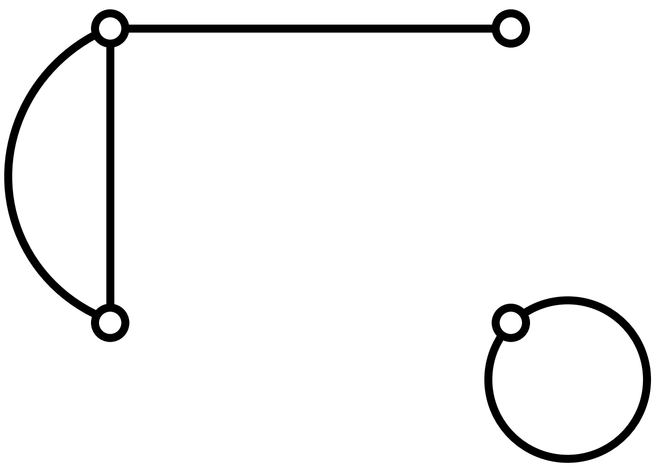


$deg(u_1) = 4$ $deg(u_2) = 2$

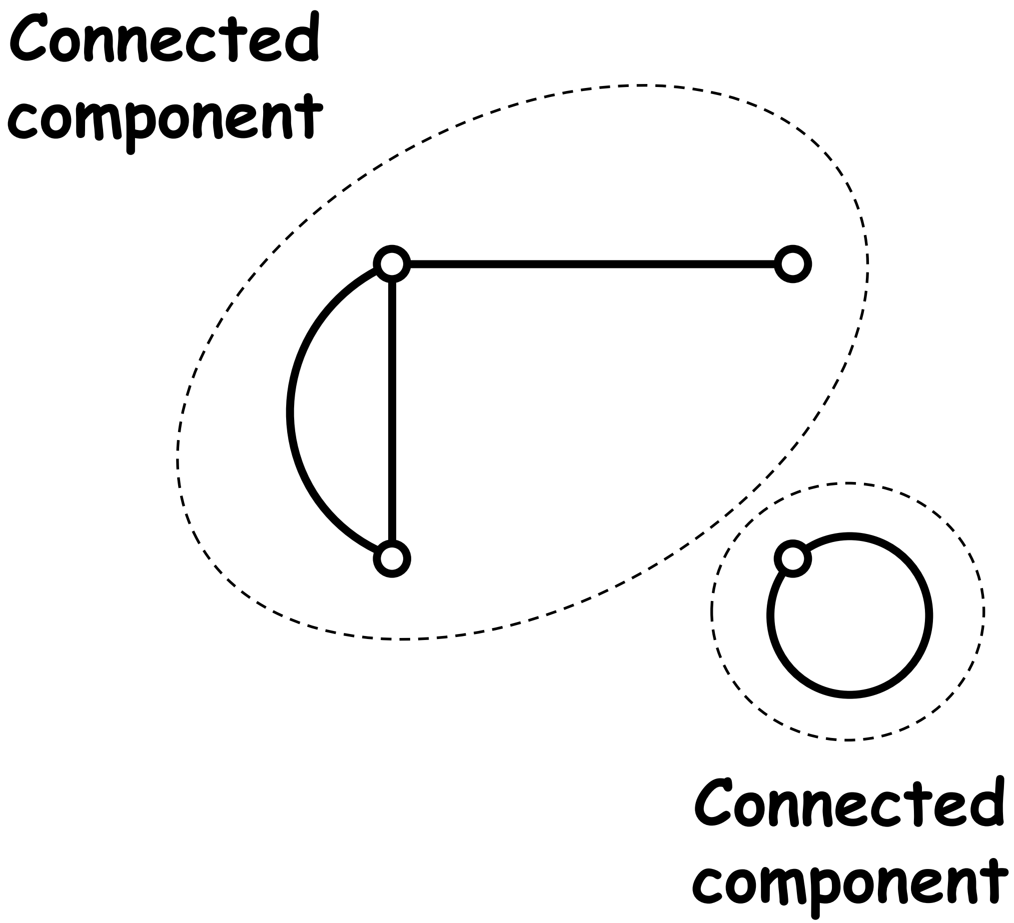
Connected graph

- A graph in which there is a path between any pair of nodes





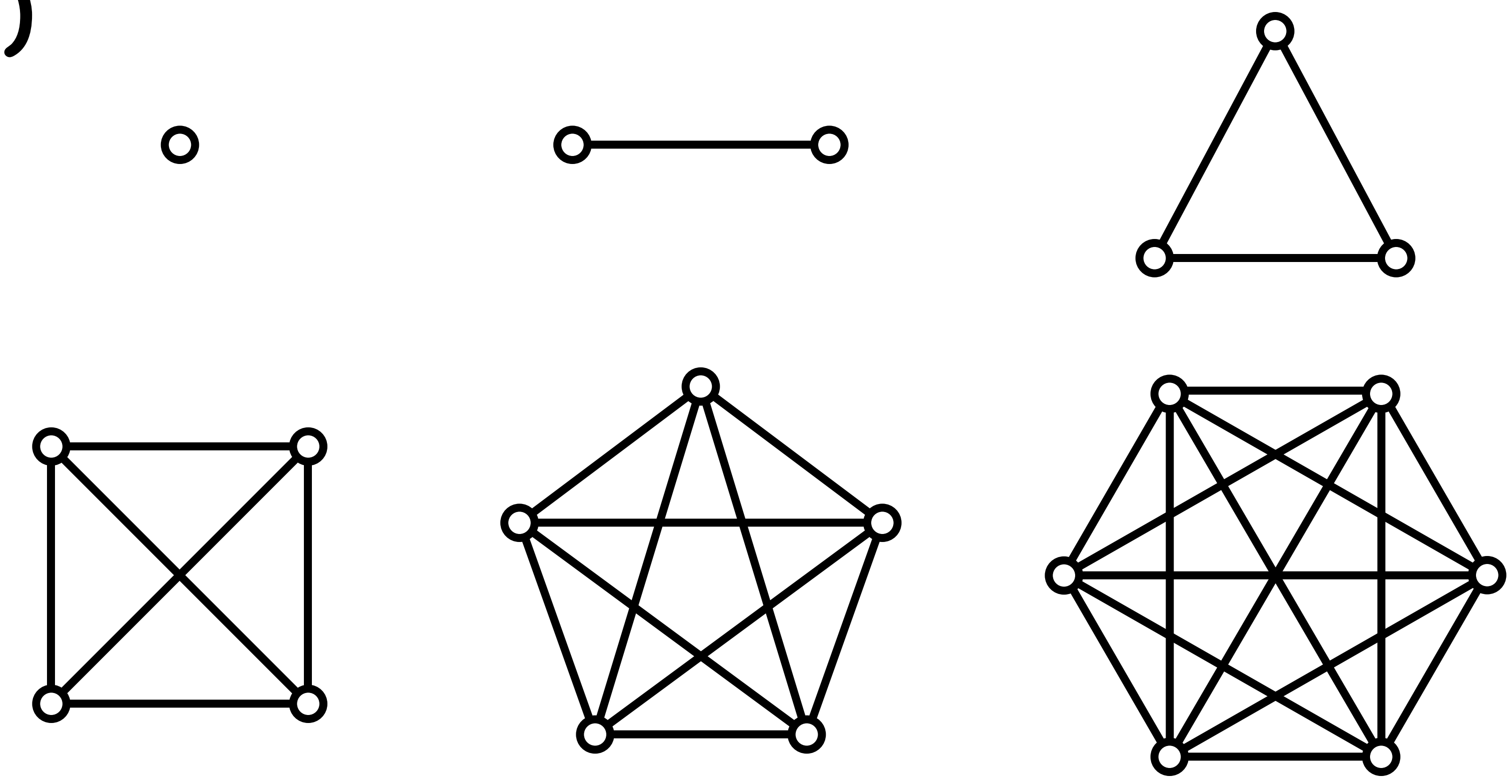
Connected components



Number of connected components = 2

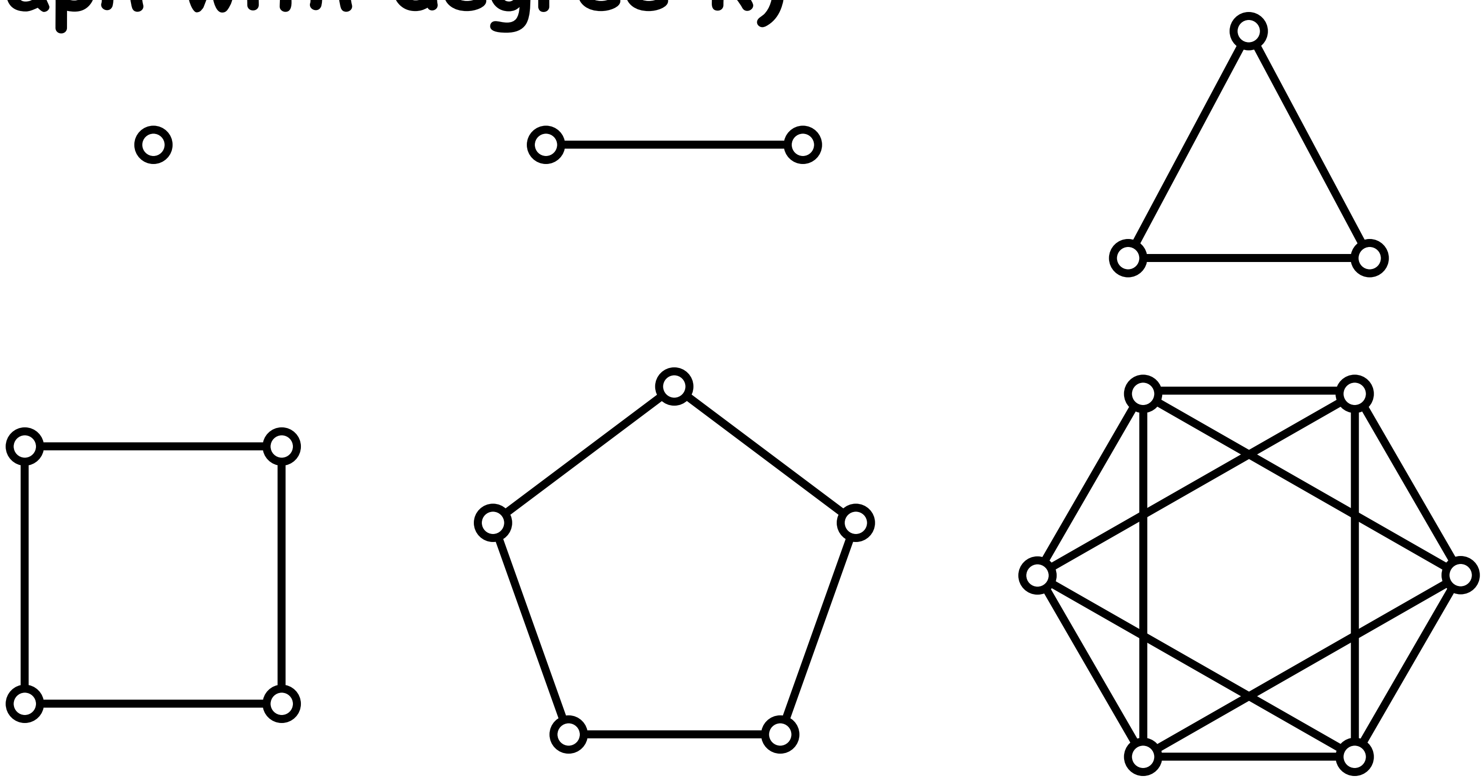
Complete graph

- A graph in which any pair of nodes are connected (often written as K_1, K_2, …)



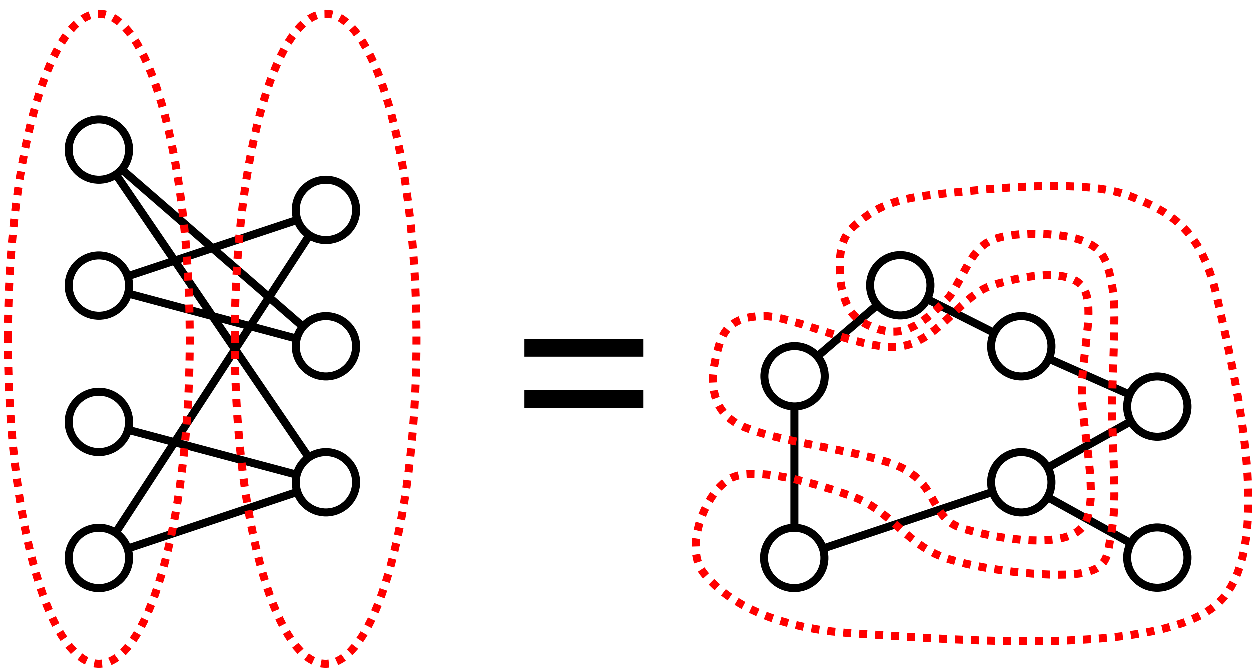
Regular graph

- A graph in which all nodes have the same degree (often called k-regular graph with degree k)

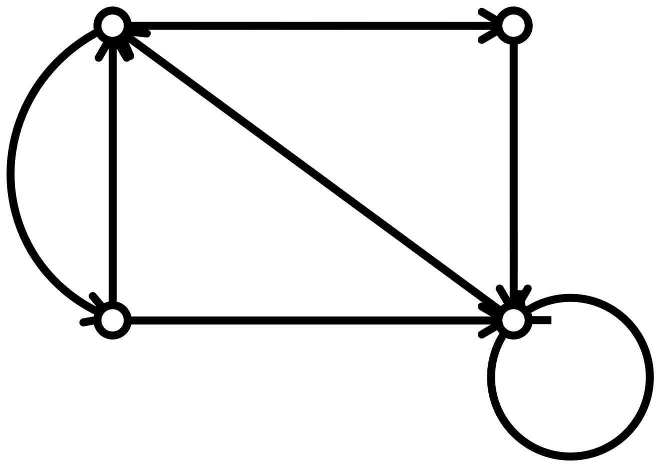


Bipartite graph

- A graph whose nodes can be divided into two subsets so that no link connects nodes within the same subset



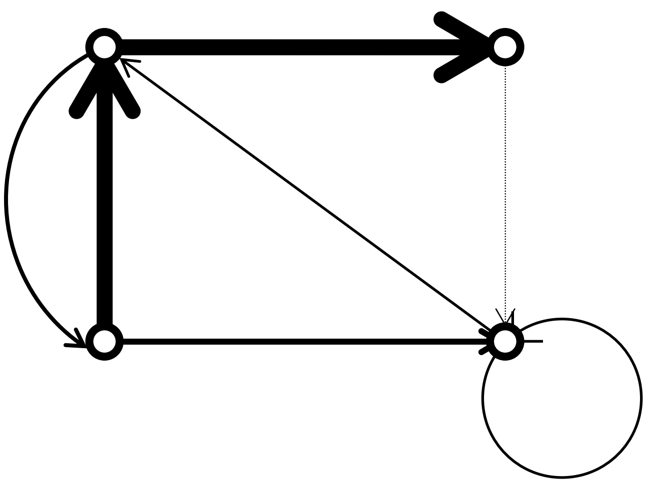
Directed graph



- Each link is directed
- Direction represents either order of relationship or accessibility between nodes

E.g. genealogy

Weighted directed graph



- Most general version of graphs
- Both weight and direction is assigned to each link

E.g. traffic network

Measuring Topological Properties of Networks (1): Macroscopic Properties

Network density

- The ratio of # of actual links and # of possible links
  - For an undirected graph:

    $$d = |E| \,/\, (\, |V| \,(|V| - 1) \,/\, 2 \,)$$

  - For a directed graph:

    $$d = |E| \,/\, (\, |V| \,(|V| - 1) \,)$$

Characteristic path length

- In graph theory: Maximum of shortest path lengths between pairs of nodes (a.k.a. network diameter)
- In complex network science: Average shortest path lengths
- Characterizes how large the world being modeled is
  - A small length implies that the network is well connected globally

Clustering coefficient

- For each node:
  - Let n be the number of its neighbor nodes
  - Let m be the number of links among the k neighbors
  - Calculate c = m / (n choose 2)

  Then C = <c> (the average of c)
- C indicates the average probability for two of one's friends to be friends too
  - A large C implies that the network is well connected locally to form a cluster

Degree distribution

P(k) = Prob. (or #) of nodes with degree k

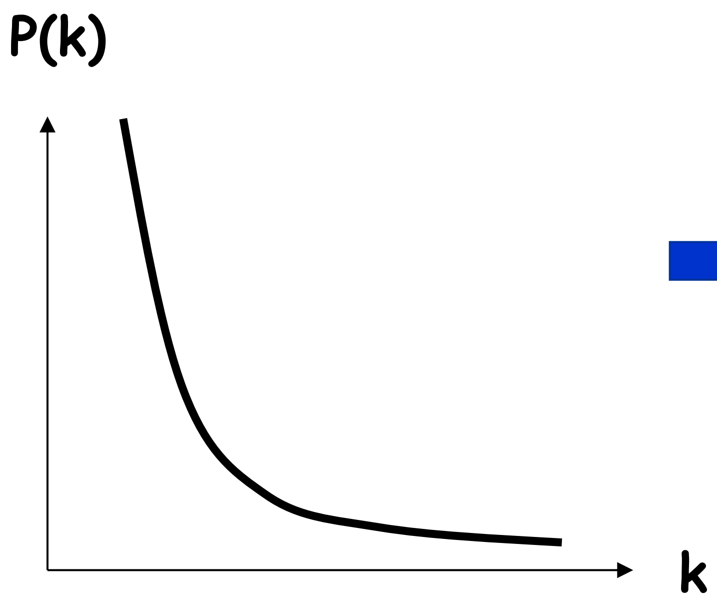
- Gives a rough profile of how the connectivity is distributed within the network

$\Sigma_k P(k) = 1$ (or total # of nodes)

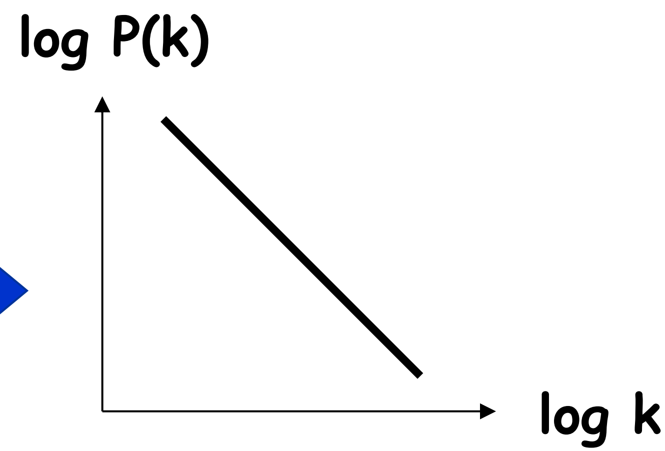
Power law degree distribution

- $P(k) \sim k^{-\gamma}$

A few well-connected nodes,
a lot of poorly connected nodes



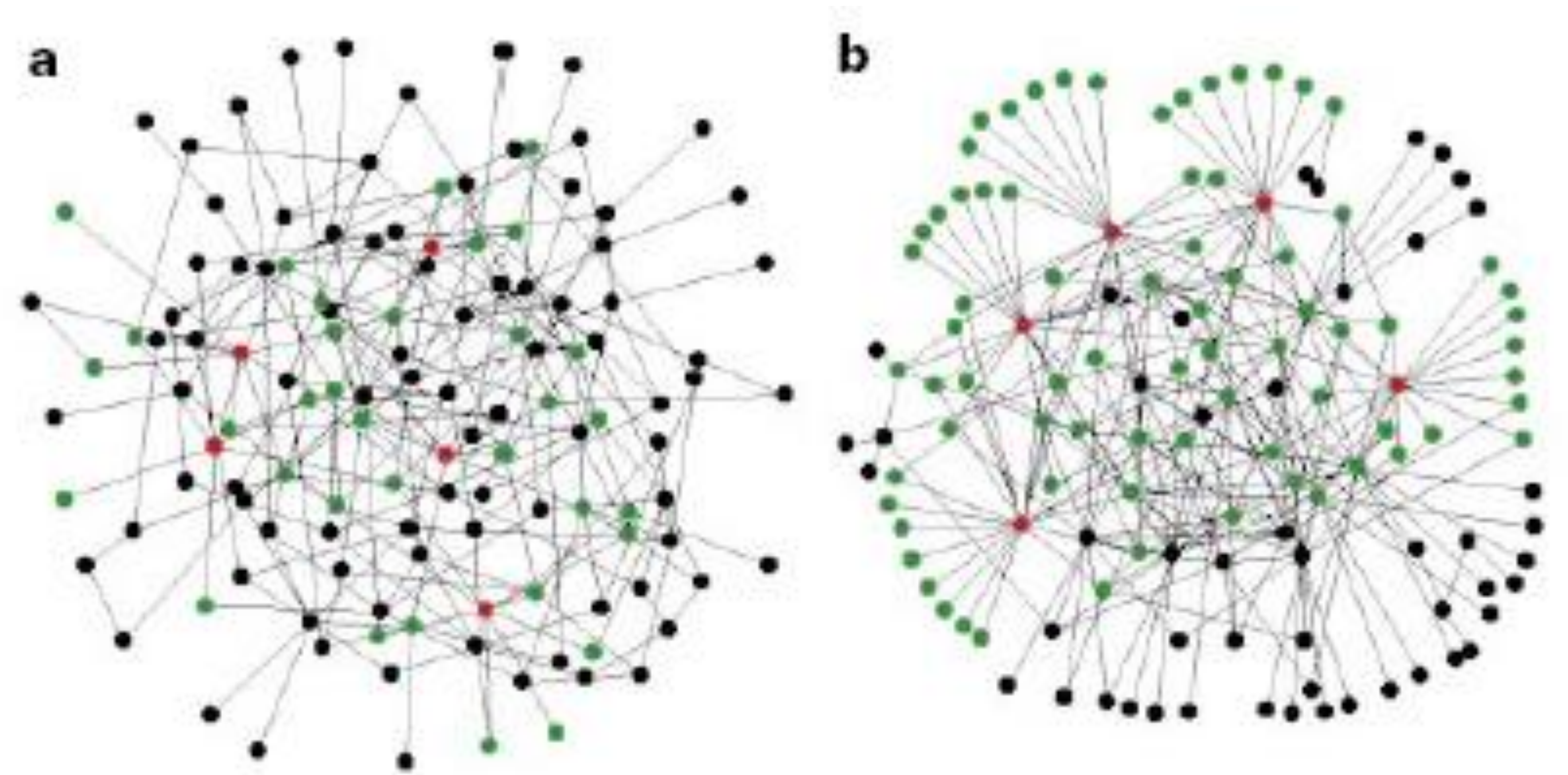
Scale-free network



Linear in log-log plot

-> No characteristic scale
(Scale-free networks)

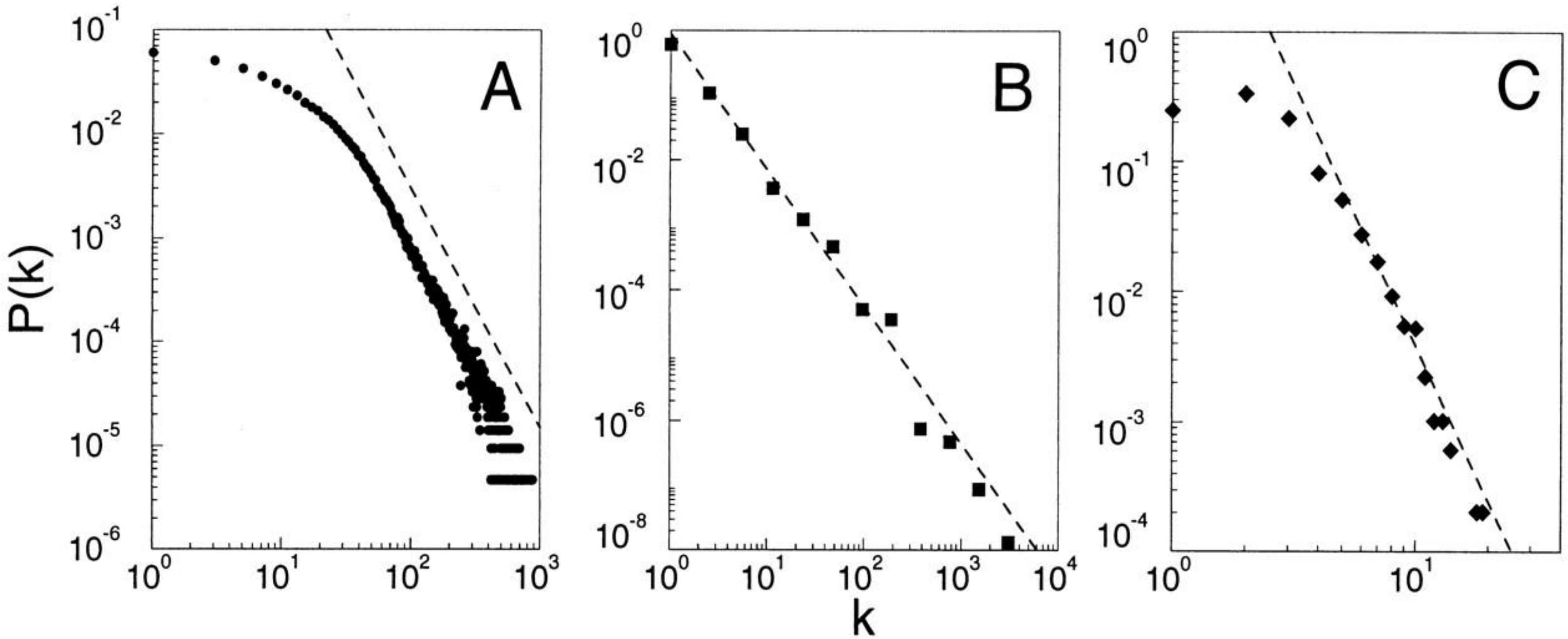
How it appears



Random

Scale-free

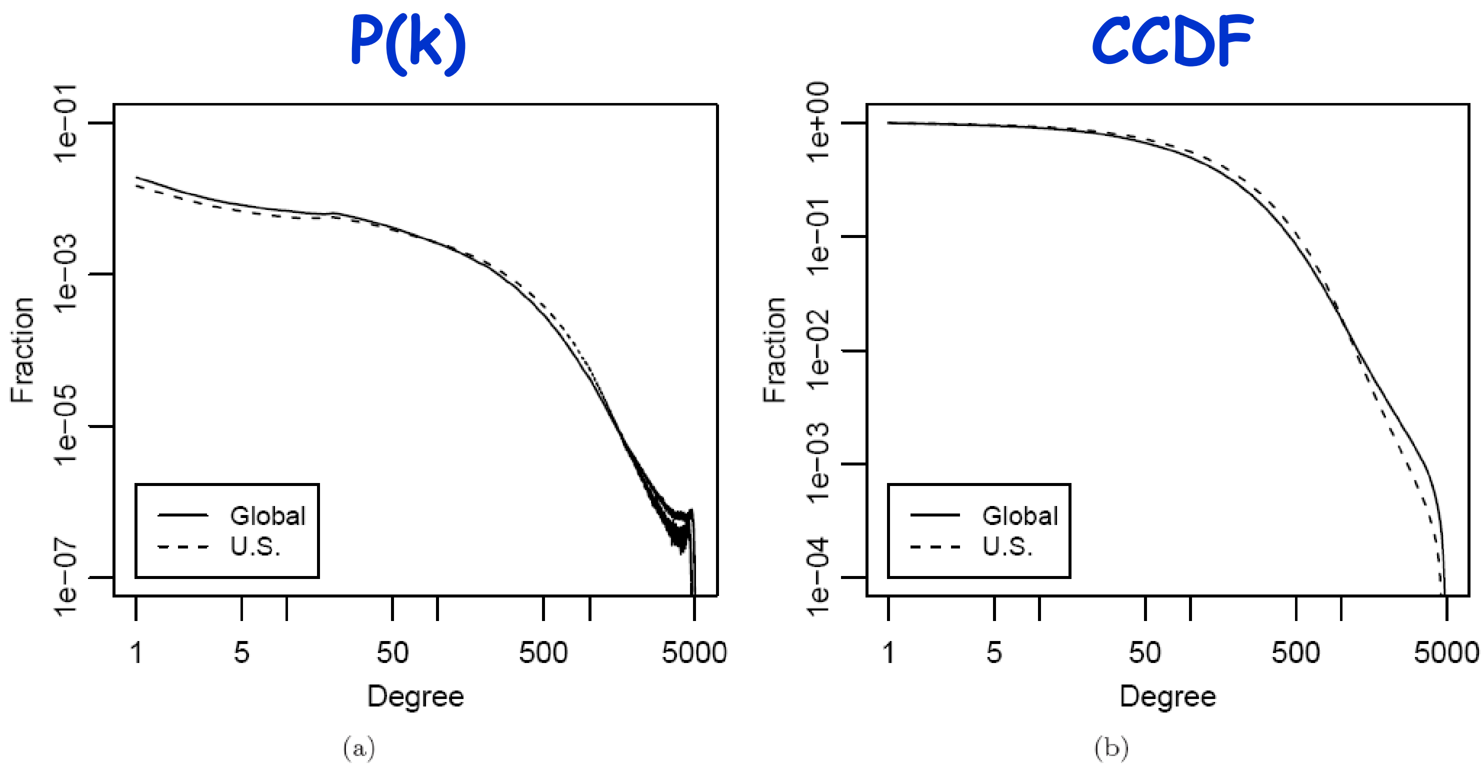
Degree Distributions of Real-World Complex Networks



Actors WWW Power grid

A Barabási, R Albert Science 1999;286:509-512

Degree distribution of FB



- http://www.facebook.com/note.php?note_id=10150388519243859
- http://arxiv.org/abs/1111.4503

Measuring Topological Properties of Networks (2): Centralities

Centrality measures (“B,C,D,E”)

- Degree centrality
  - How many connections the node has
- Betweenness centrality
  - How many shortest paths go through the node
- Closeness centrality
  - How close the node is to other nodes
- Eigenvector centrality

Degree centrality

- Simply, # of links attached to a node

$$C_D(v) = deg(v)$$

or sometimes defined as

$$C_D(v) = deg(v) / (N-1)$$

Betweenness centrality

- Prob. for a node to be on shortest paths between two other nodes

$$C_B(v) = \Sigma_{s \neq v, t \neq v} \frac{\#sp(s,e,v)}{\#sp(s,e)}$$

  - s: start node, e: end node
  - #sp(s,e,v): # of shortest paths from s to e that go though node v
  - #sp(s,e): total # of shortest paths from s to e
  - Easily generalizable to “group betweenness”

Closeness centrality

- Inverse of an average distance from a node to all the other nodes

$$C_C(v) = \frac{n-1}{\Sigma_{w \neq v} d(v,w)}$$

- d(v,w): length of the shortest path from v to w
- Its inverse is called "farness"
- Sometimes "Σ" is moved out of the fraction (it works for networks that are not strongly connected)
- NetworkX calculates closeness within each connected component

Eigenvector centrality

- Eigenvector of the largest eigenvalue of the adjacency matrix of a network

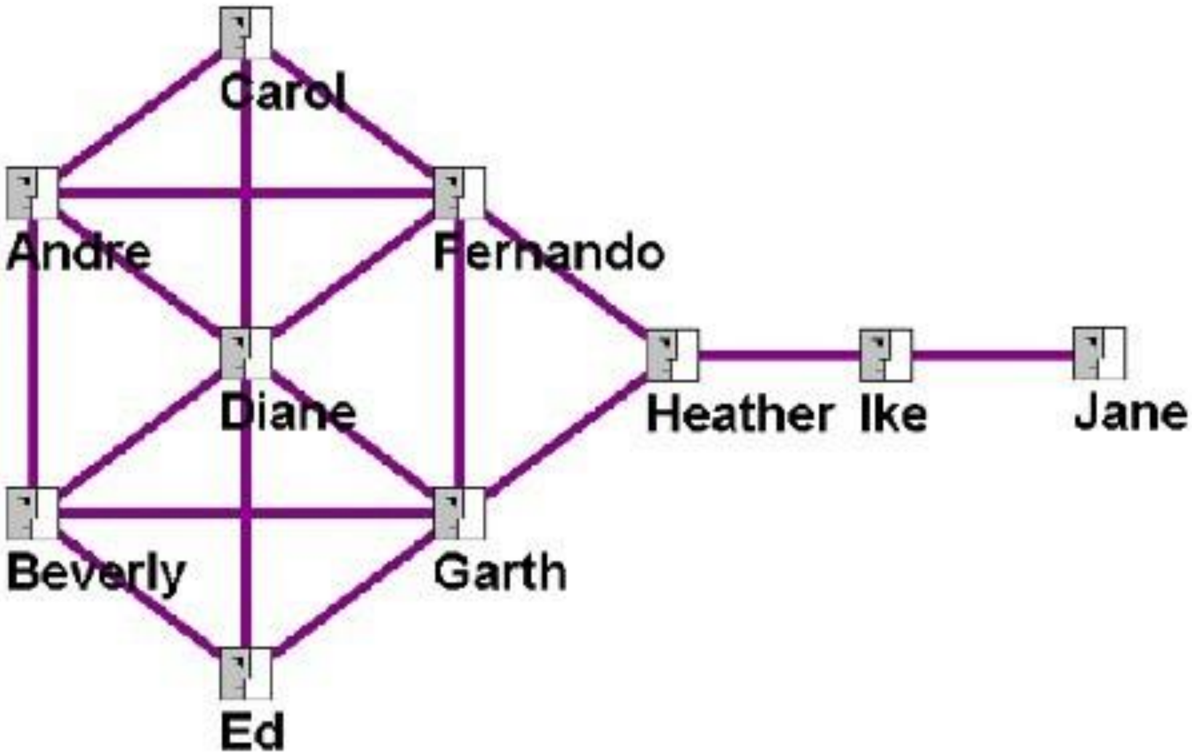
$$C_E(v) = (v\text{-th element of } x)$$

$$Ax = \lambda x$$

- λ: dominant eigenvalue
- x is often normalized ($|x| = 1$)

Exercise

- Who is most central by degree, betweenness, closeness, eigenvector?



Which centrality to use?

- To find the most popular person
- To find the most efficient person to collect information from the entire organization
- To find the most powerful person to control information flow within an organization
- To find the *most important* person (?)

Measuring Topological Properties of Networks (3): Mesoscopic Properties

Degree correlation (assortativity)

- Pearson's correlation coefficient of node degrees across links

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \, \sigma_Y}$$

  - X: degree of start node (in / out)
  - Y: degree of end node (in / out)

Assortative/disassortative networks

Network	n	r
Physics coauthorship (a)	52 909	0.363
Biology coauthorship (a)	1 520 251	0.127
Mathematics coauthorship (b)	253 339	0.120
Film actor collaborations (c)	449 913	0.208
Company directors (d)	7 673	0.276
Internet (e)	10 697	−0.189
World-Wide Web (f)	269 504	−0.065
Protein interactions (g)	2 115	−0.156
Neural network (h)	307	−0.163
Marine food web (i)	134	−0.247
Freshwater food web (j)	92	−0.276
Random graph (u)		0
Callaway *et al.* (v)		$\delta/(1+2\delta)$
Barabási and Albert (w)		0

Social networks are assortative

Engineered / biological networks are disassortative

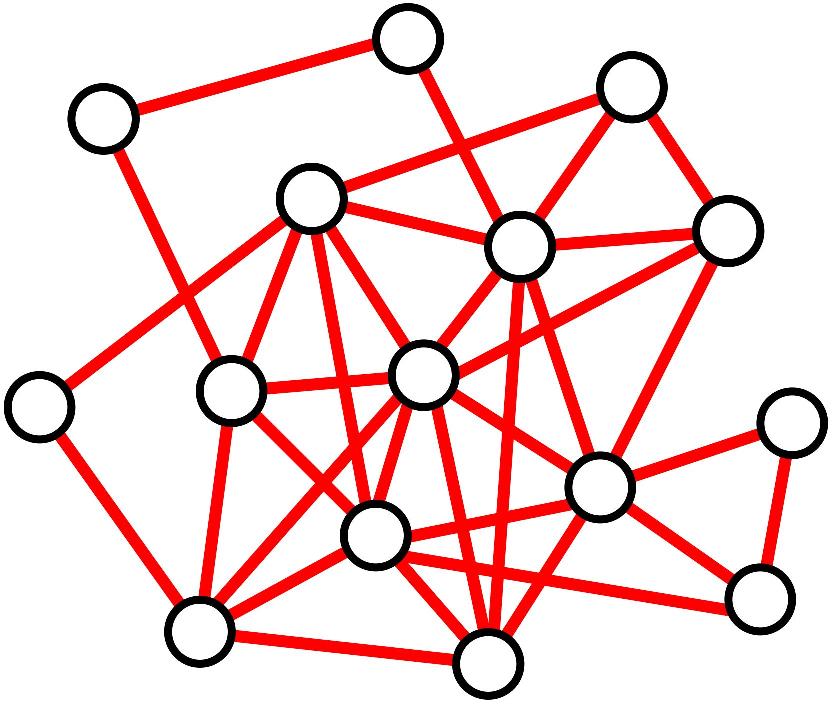
(from Newman, M. E. J., Phys. Rev. Lett. 89: 208701, 2002)

K-cores

- A connected component of a network obtained by repeatedly deleting all the nodes whose degree is less than k until no more such nodes exist
  - Helps identify where the core cluster is
  - All nodes of a k-core have at least degree k
  - The largest value of k for which a k-core exists is called "degeneracy" of the network

Exercise

- Find the k-core (with the largest k) of the following network



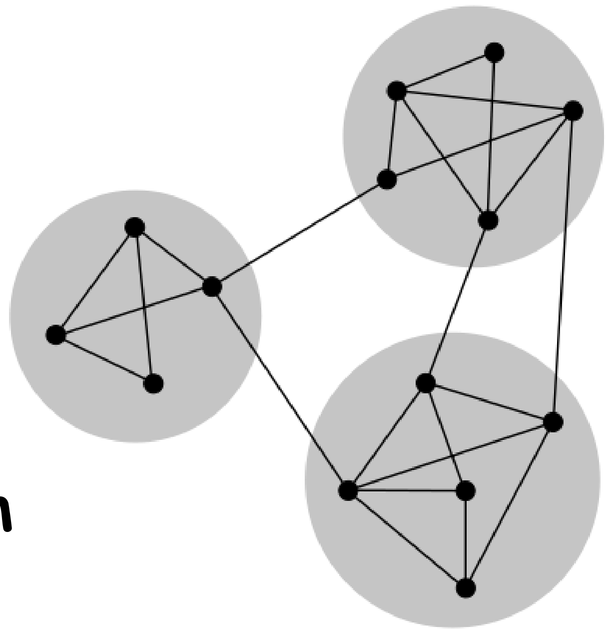
Coreness (core number)

- A node's coreness (core number) is c if it belongs to a c-core but not (c+1)-core

- Indicates how strongly the node is connected to the network
- Classifies nodes into several layers
  - Useful for visualization

Community

- A subgraph of a network within which nodes are connected to each other more densely than to the outside
  - Still defined vaguely…
  - Various detection algorithms proposed
    - K-clique percolation
    - Hierarchical clustering
    - Girvan-Newman algorithm
    - Modularity maximization (e.g., Louvain method)



(diagram from Wikipedia)

Modularity

- A quantity that characterizes how good a given community structure is in dividing the network

$$Q = \frac{|E_{in}| - |E_{in-R}|}{|E|}$$

- $|E_{in}|$: # of links connecting nodes that belong to the same community
- $|E_{in-R}|$: Estimated $|E_{in}|$ if links were random

Community detection based on modularity

- The Louvain method
  - Heuristic algorithm to construct communities that optimize modularity
    - Blondel et al. J. Stat. Mech. 2008 (10): P10008
- Python implementation by Thomas Aynaud available at:
  - https://bitbucket.org/taynaud/python-louvain/