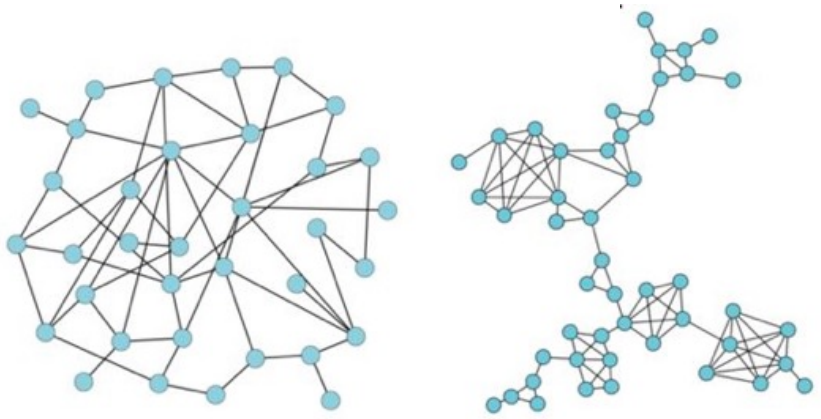
Statistical Analysis of Networks

Lecture 7 – Basic concepts

NETWORK COHESION

- it refers to a **measure** of the connectedness and togetherness among nodes in the **network**
  - the extent to which nodes (or subsets of nodes) '*stuck together*' by the relation defining edges in the network
  - if a relation is observed between two nodes, are other nodes linked to that dyads and also linked among them?



NETWORK COHESION

Density

- proportion of *possible* edges actually present in the graph (determined by # of nodes *n*
- raw measure of network *cohesion* (global)
- sometime it is preferred a more local or a different 'global' perspective *(scale)*

Different ways to define (and therefore to measure) *cohesion*:

- *local* vs *global*
- subset of nodes: ***explicit*** definition (search for cliques: a *maximal* subgraph with paths between all nodes or others more computationally tractable subsets such as *k*-cores: subgraph for which all nodes have at least degree *k*) or **implicit** definition (clusters or communities: subgraph of *well connected nodes*)

NODE CLUSTERING COEFFICIENT

Cohesion involving more one individual node

- It captures the degree to which the neighbors of a given node ***i*** with degree ***k*** link to each other (local scale)
- density of links in node *i*'s immediate neighborhood

$C_i = 2L_i/k_i\,(k_i - 1)$ with L_i # of links between the $\boldsymbol{k_i}$ neighbors of *i*

$$C_i = \frac{\text{number of pairs of neighbors of } i \text{ that are connected}}{\text{number of pairs of neighbours of } i}$$

- Assuming undirected graph without loops:

$$C_i = \frac{\sum_{j=1}^{N} \sum_{k=1}^{j-1} a_{ij} a_{ik} a_{jk}}{\binom{k_i}{2}}$$

(can be generalized to directed and weighted networks)

network's local link density:
the more densely interconnected the neighborhood of node *i*, the higher is its local clustering coefficient.

NODE CLUSTERING COEFFICIENT

$C_i = 2L_i / k_i\,(k_i - 1)$

C_i = 0 if none of the neighbors of node *i* link to each other

C_i = 1 if the neighbors of node *i* form a complete graph, i.e. they all link to each other

C_i is the probability that two neighbors of a node link to each other

- C = 0.5: a 50% chance that two neighbors of a node are linked

A possible global clustering measure in a network can be obtain by averaging the node clustering coefficients for each node

AVERAGE CLUSTERING COEFFICIENT

average clustering coefficient

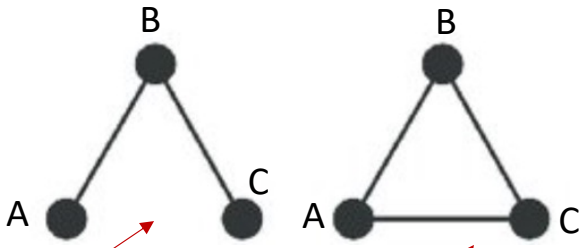
the extent of clustering of a whole network

$$\langle C \rangle = 1/n \sum C_i$$

$\langle C \rangle$ is the probability that two neighbors of a randomly selected node link to each other

GLOBAL CLUSTERING COEFFICIENT

$$C_\Delta = \frac{3 \times NumberOfTriangles}{NumberOfConnectedTriples}$$



- *connected triplet:* an ordered set of three nodes ABC such that A connects to B and B connects to C.

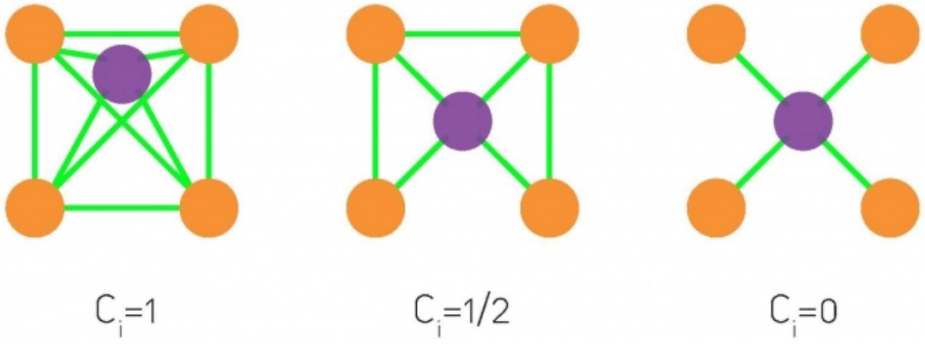
- a chain of connected nodes A, B, C, in which B connects to A and C, but A does not link to C, forms a single open triplet ABC.

N.B.: an A, B, C triangle is made of three triplets, ABC, BCA and CAB:
each triangle is counted three times in the triplet count (**3** in the numerator)

Global clustering coefficient: social network literature since 1940s where C_Δ is often called the *ratio of transitive triplets (a measure of transitivity)*

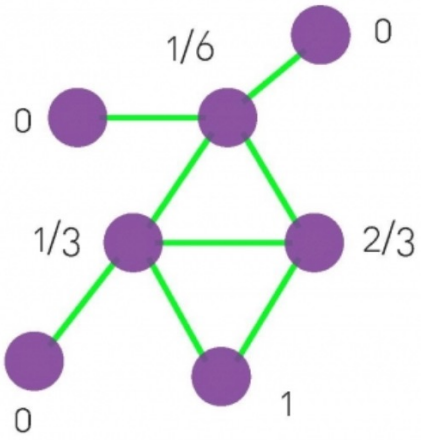
Node and Average/Global clustering coefficient

a. $C_i = 2L_i/k_i\,(k_i - 1)$



local clustering coefficient, C_i , of the central node with degree k_i = 4 (**purple**) for three different configurations of its neighborhood (the local density of links in a node's vicinity)

b.



$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

$$C_{\Delta} = \frac{3}{8} = 0.375$$

$\langle C \rangle$ and C_{Δ} measure transitivity in different ways and can differ substantially

C_{Δ} most used

CONNECTED COMPONENTS (CONNECTIVITY)

A larger global perspective than node clustering

Conneted graph (network): every node is reachable (undirected)

- Conneted component:
  - a subset of nodes with a path between any two nodes that belong to the component, but one cannot add any more nodes to it that would have the same property
- Components divide the graph into separated regions

- A connected component in a graph that contains the majority of the nodes: *giant component*
  - *Unconnected graph with a giant component: analysis is usually restricted to it*

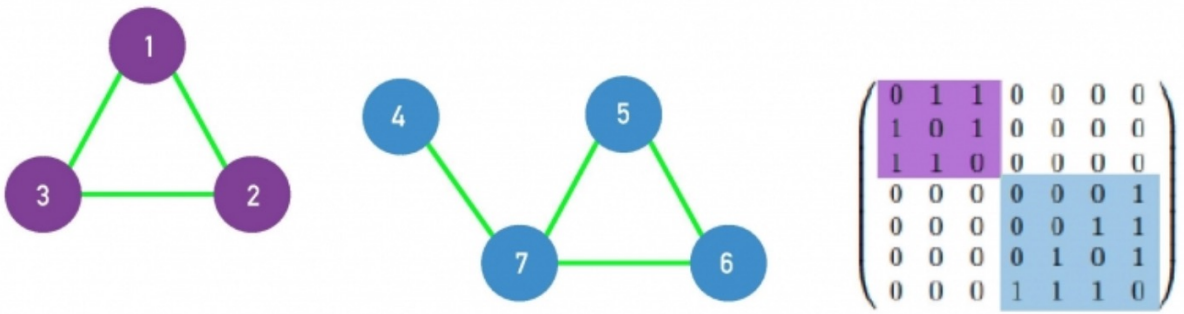
If a network consists of two components, a properly placed single link can connect them, making the network connected.

Such a link is called a *bridge*.

In general a bridge is any link that, if cut, disconnects the network.

CONNECTED COMPONENTS (CONNECTIVITY)

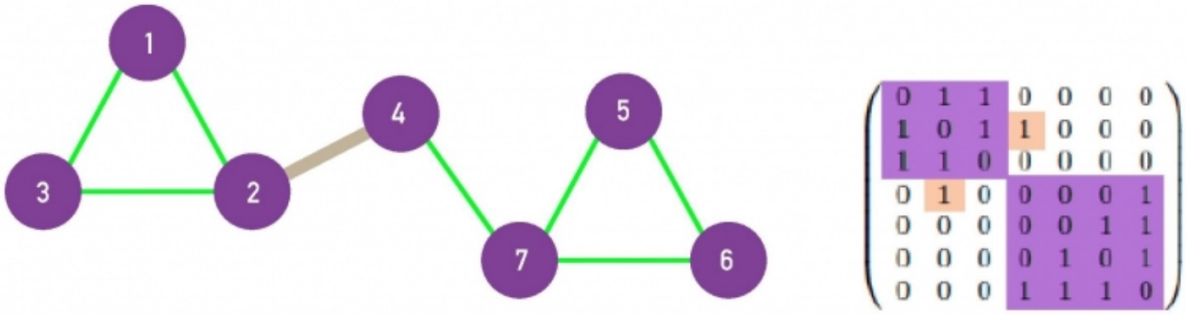
a.



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

Adjacency matrix is block diagonal

b.



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

- The addition of a single link, called a *bridge*, shown in grey, turns a disconnected network into a single connected component. Now there is a path between every pair of nodes in the network. Consequently the adjacency matrix cannot be written in a block diagonal form.

FINDING CONNECTED COMPONENTS

Challenging question in large networks

Mathematical and algorithmic tools can help to identify the connected components of a graph.

- linear algebra tools can be used to decide if the adjacency matrix is block diagonal, helping us to identify the connected components

- in practice, for large networks the components are more efficiently identified using the **Breadth-First Search** algorithm

FINDING CONNECTED COMPONENTS

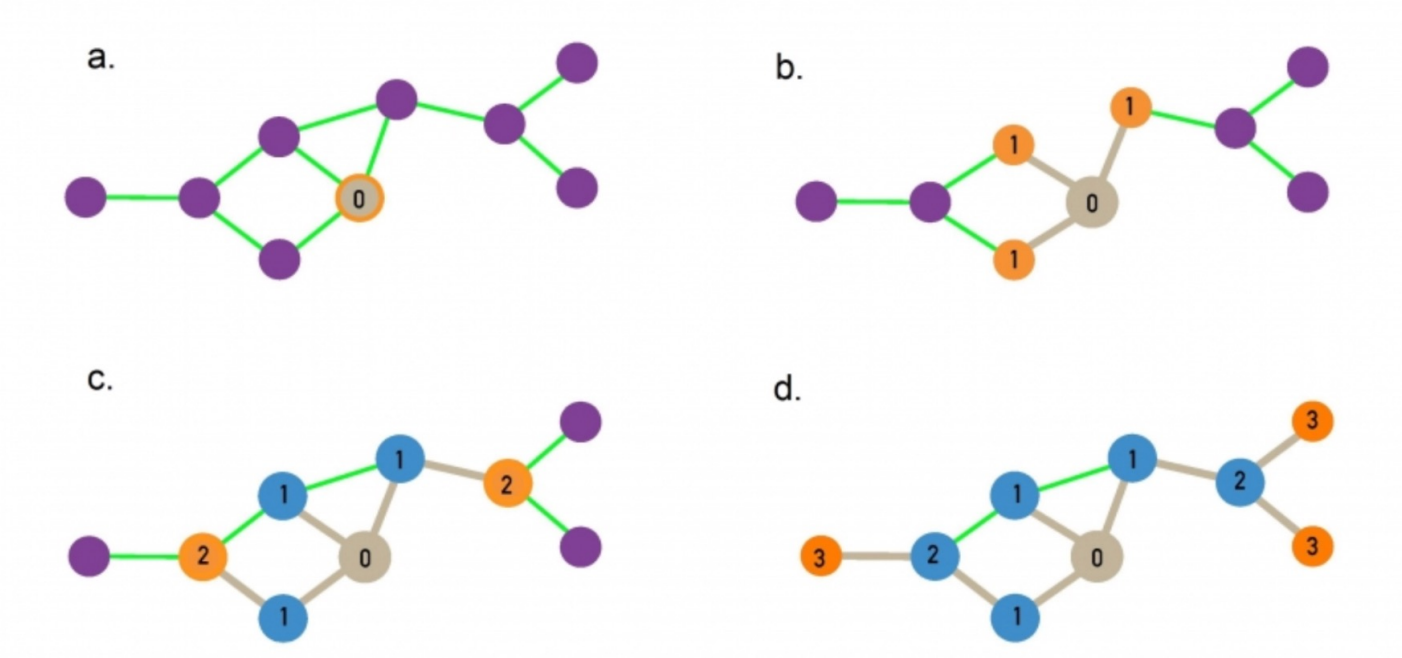
Finding the Connected Components of a Network by Breadth-First Search (BFS) Algorithm

1. Start from a randomly chosen node ***i*** and perform a BFS. Label all nodes reached this way with $n = 1$.

2. If the total number of labeled nodes equals N, then the network is connected. If the number of labeled nodes is smaller than N, the network consists of several components. To identify them, proceed to step 3.

3. Increase the label $n \rightarrow n + 1$. Choose an unmarked node j, label it with n. Use BFS to find all nodes reachable from j, label them all with n. Return to step 2.

APPLYNG THE BFS ALGORITHM



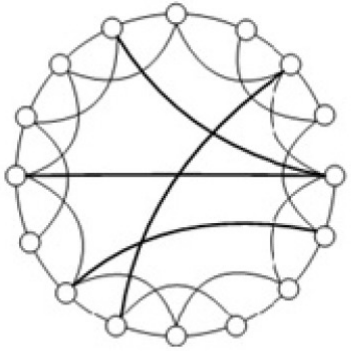
Applying the BFS Algorithm

- Starting from the orange node, labeled ”0”, we identify all its neighbors, labeling them ”1”.
- Next we label ”2” the unlabeled neighbors of all nodes labeled ”1”, and so on, in each iteration increasing the label number, until no node is left unlabeled. The length of the shortest path or the distance d_{0i} between node 0 and any other node i in the network is given by the label of node i. For example, the distance between node 0 and the leftmost node is $d = 3$.

CONNECTED COMPONENTS

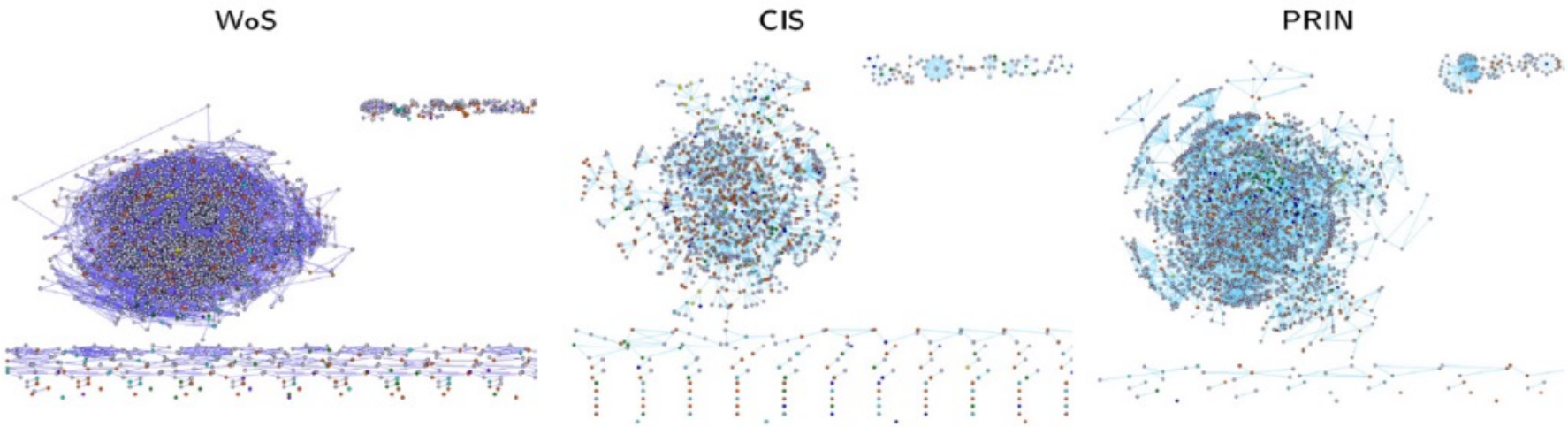
- Giant component of many types of different networks (relations and size) share an ‘interesting’ property

- ‘Small word’ property: high global clustering coefficient and small average path length (Watts and Strogatz, 1998)
  - *average path length* = the average distance between all pairs of nodes in the network (measured only for node pairs that are in the same component)



- small average path length and high clustering

CO-AUTHORSHIP NETWORK OF ITALIAN STATISTICIANS



	WoS	CIS	PRIN
#. of authors	5291	1525	2839
#. of authors per pub (St. Dev.)	12.6 (61.5)	2.4 (0.7)	2.8 (1.6)
#. of pub per author (St. Dev.)	6.1 (8.8)	7.9 (8.6)	14.8 (12.3)
#. of statisticians	481	581	556
#. of isolated	26	60	7
#. of edges	427,238	2534	9379
#. of internal edges	403	631	999
Density	0.031	0.002	0.002
Average degree	161.5	3.3	6.6
Largest distance	16	19	17
Average path length (ℓ)	5.47	7.15	6.52
Clustering coefficient (Γ)	0.91	0.30	0.54
# of components ≥ 1	77	54	20
Giant component (%)	91.7	87.7	94.9
E–I index	0.76	0.03	0.24