Network Science

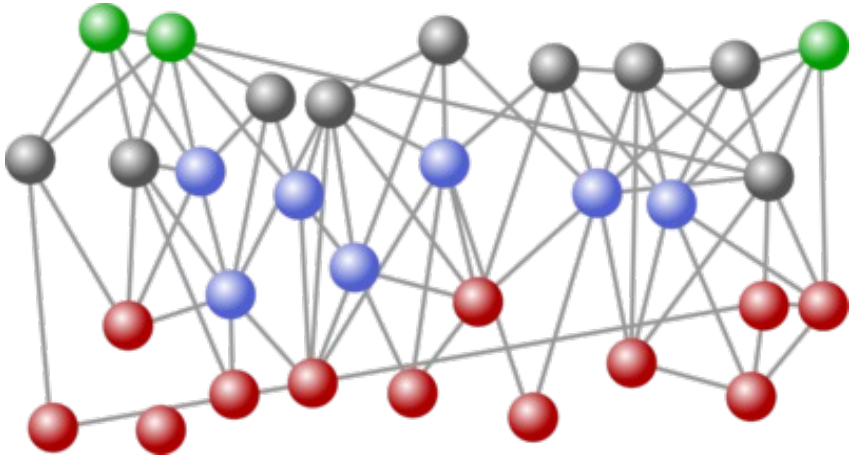
Lorenzo Cima

lorenzo.cima@phd.unipi.it; lorenzo.cima@iit.cnr.it

INTRODUCTION

Complex Systems

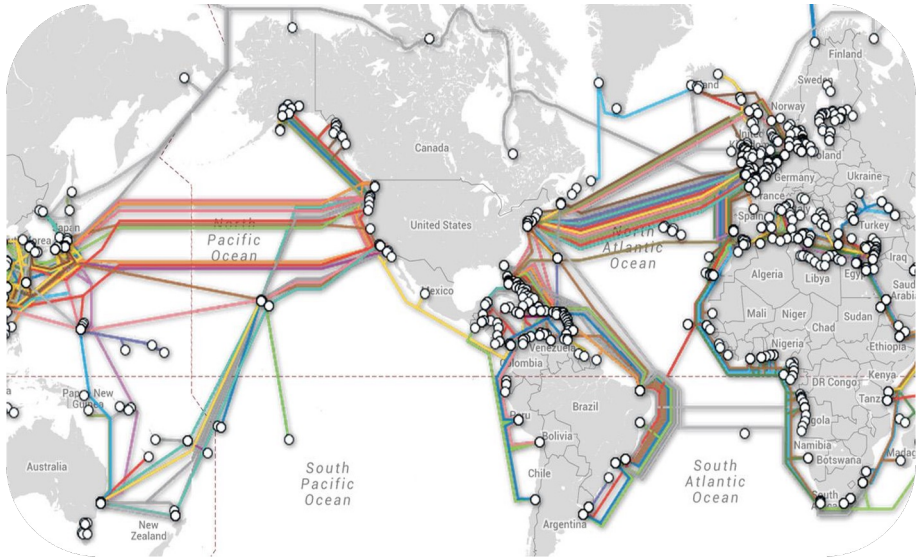
Behind each **complex system**
there is a **network**
that defines the **interactions**
between the **components**



Suggested Reading
Complexity Explained
https://complexityexplained.github.io/

Complex Systems

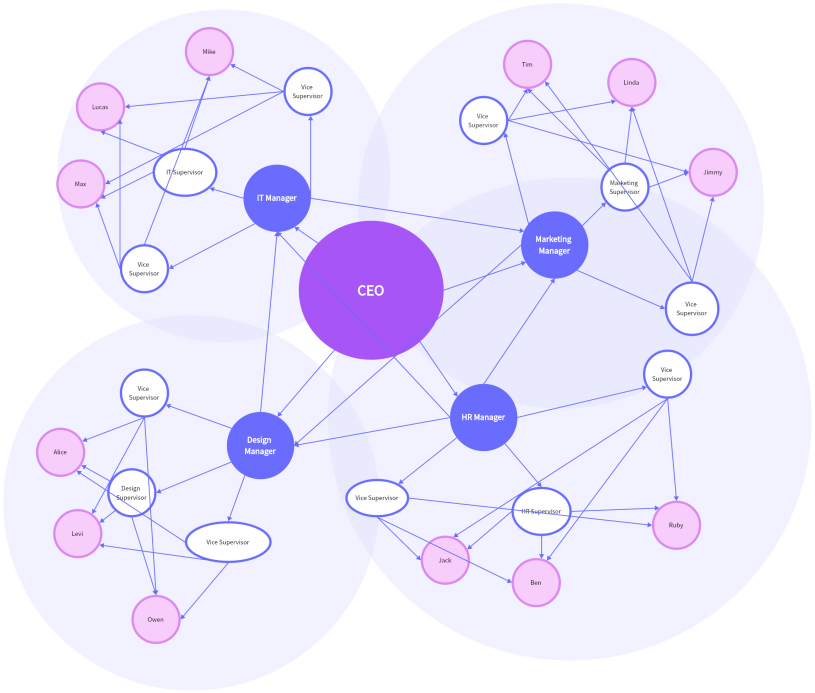
- Internet backbone



- World Wide Web

Complex Systems

- Organization structures



- Social graphs

Component of a Complex System

Network or Graph?

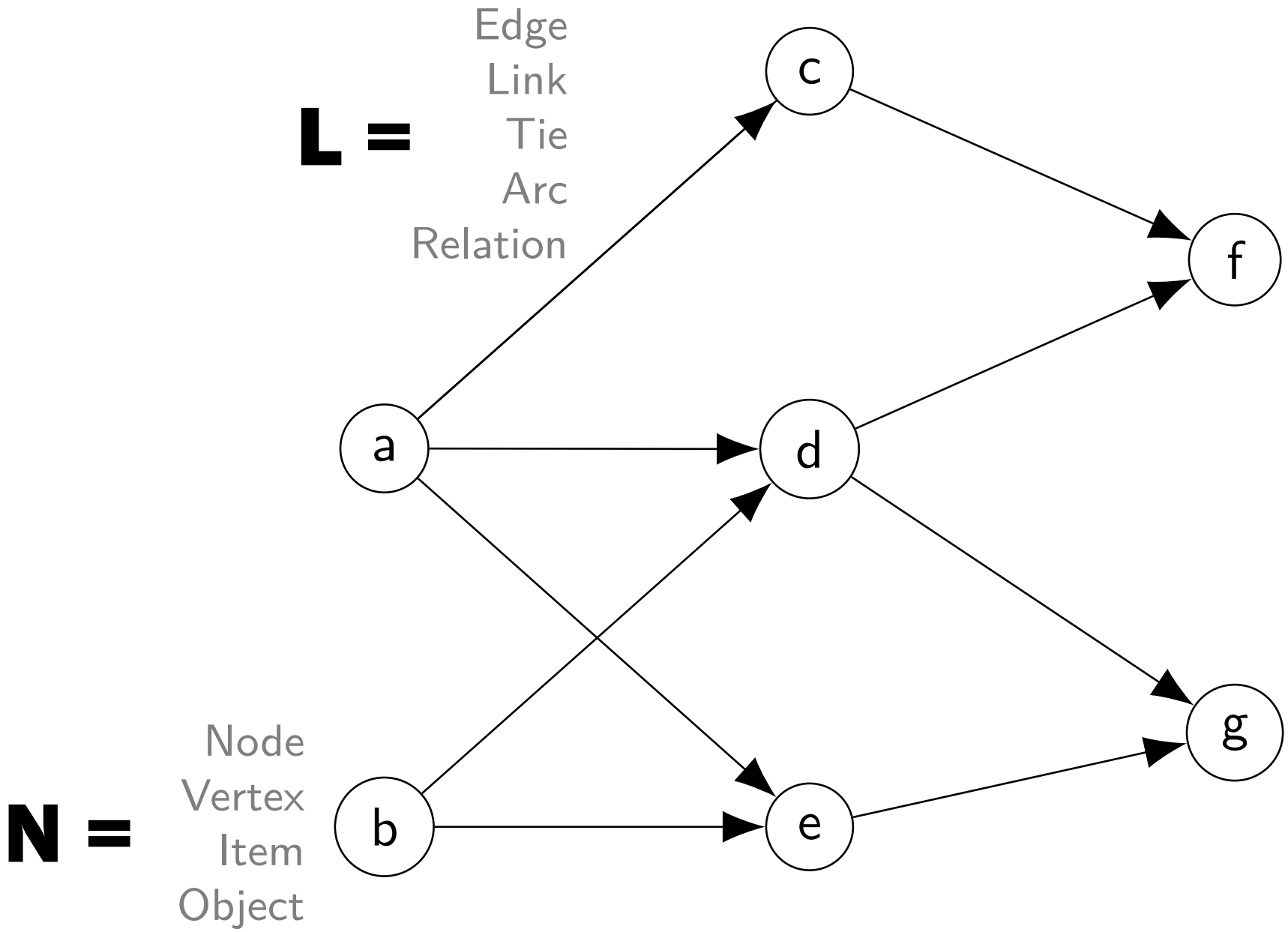
- **Network:** refers to real systems

  (www, social network)

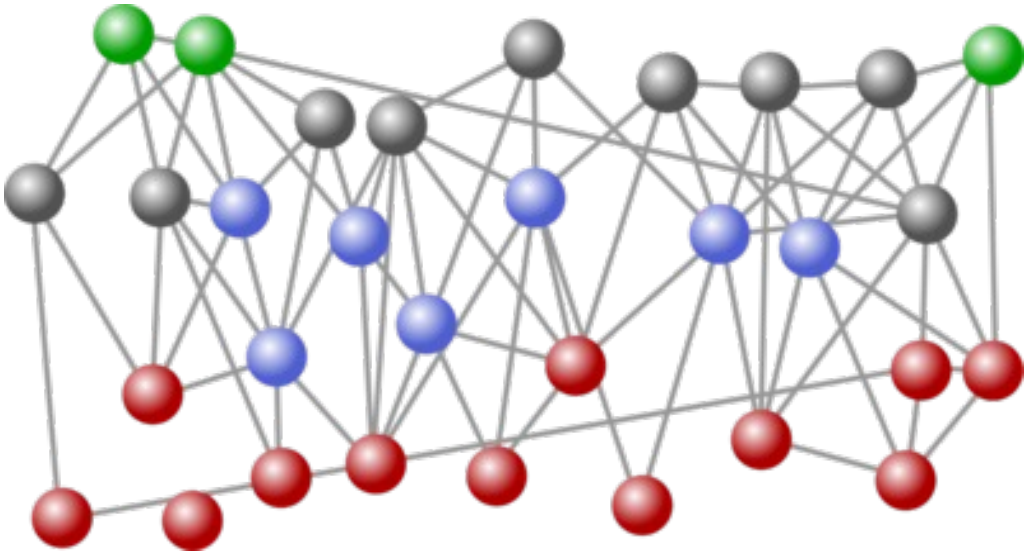
- **Graph:** mathematical representation of a network

  (web graph, social graph)

Component of a Complex System



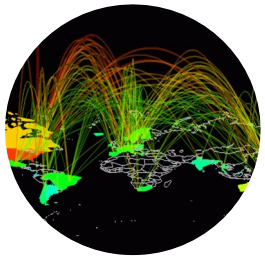
Real World Networks



Type: Social
Nodes: Individuals
Links: Social relationship

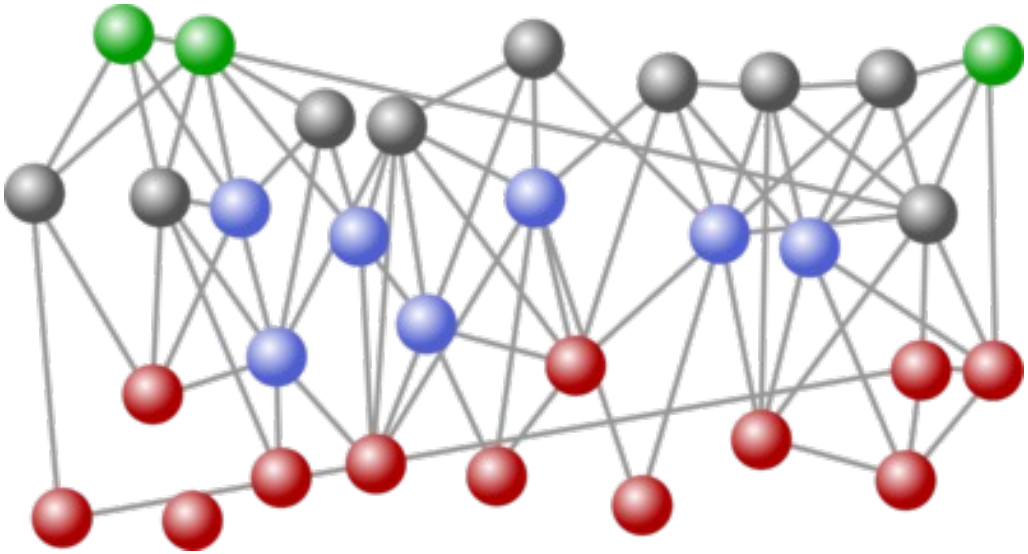
Type: Scientific Collaborations
Nodes: Researchers
Links: Co-Authorships

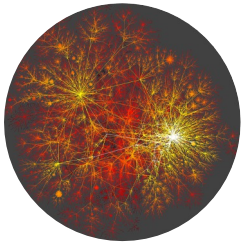
Type: Actor connectivity
Nodes: Actors
Links: Cast jointly



Type: Communication
Nodes: Phones, Airports..
Links: Phone calls, Flights..

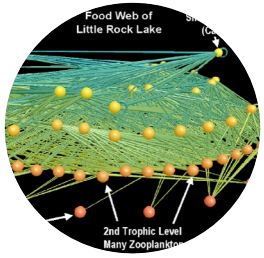
Real World Networks





Type: Technological
Nodes: PC, Routers
Links: Physical lines

Type: Scientific Citation
Nodes: Papers
Links: Citations



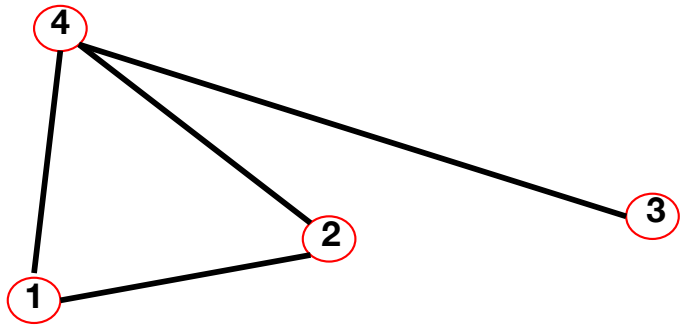
Type: Biological
Nodes: Species
Links: Trophic interactions

Type: Mobility
Nodes: Individuals, Cars...
Links: Co-Location...

Graph Types

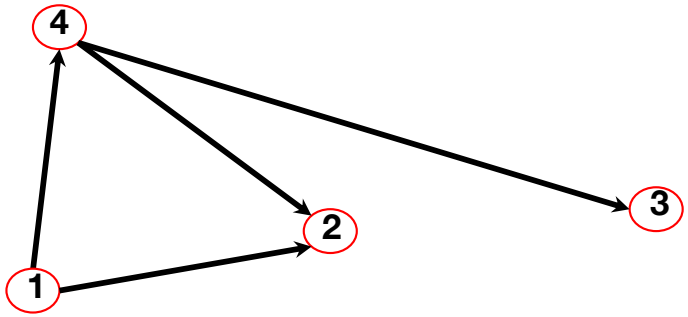
- **Undirected:** direction is meaningless

  (collaborations, social media friendships)



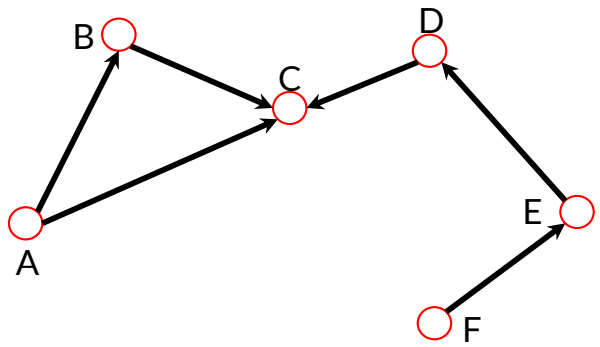
- **Directed:** direction is fundamental

  (follower/following relations, mobile communications)

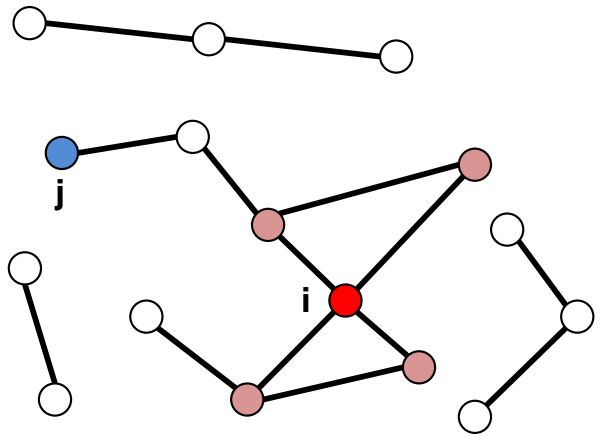


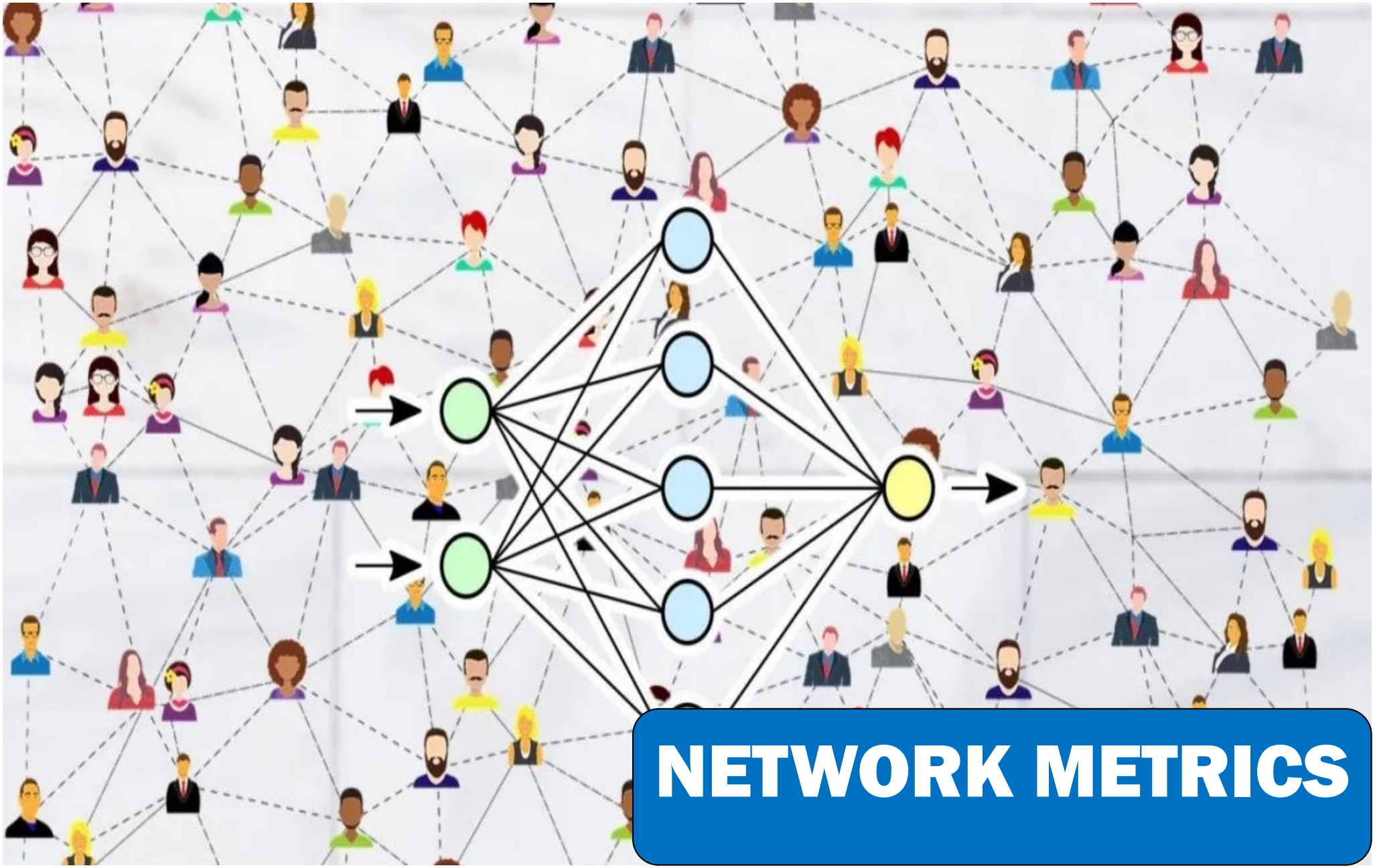
Graph Types

- **Connected:** all the nodes could be reached



- **Disconnected:** some nodes are isolated



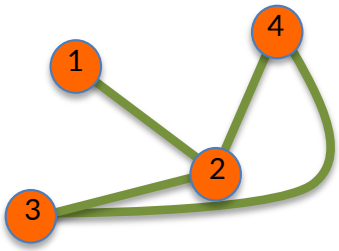


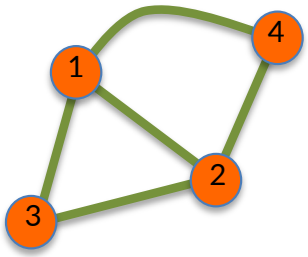
NETWORK METRICS

Connectivity

Node (edge) connectivity: minimum number of nodes (edges) to let a connected network become disconnected

Node and edge connectivity of these graphs?

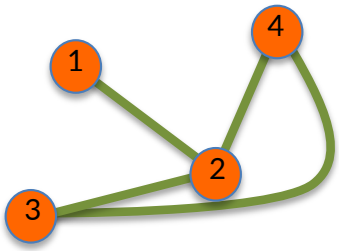




Connectivity Measures

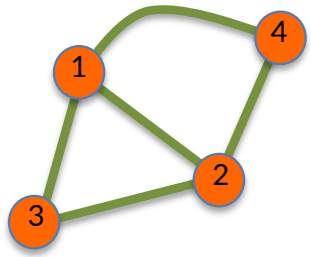
Node (edge) connectivity: minimum number of nodes (edges) to let a connected network become disconnected

Node and edge connectivity of these graphs?



Node connectivity: 1 (remove node 2)

Edge connectivity: 1 (remove edge 1-2)



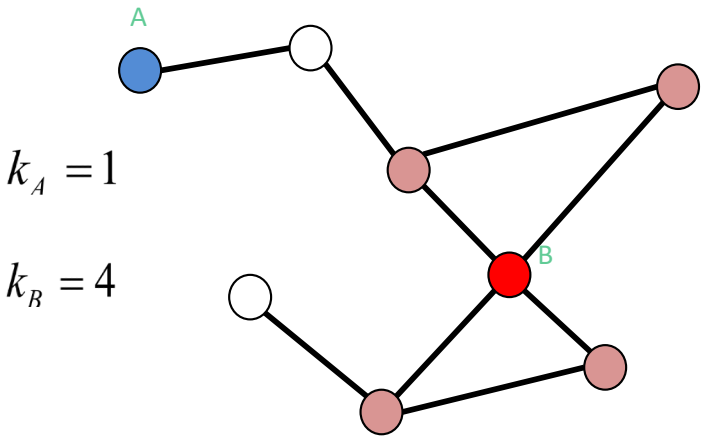
Node connectivity: 2 (remove nodes 1,2)

Edge connectivity: 2
(remove edges connected to 3 or 4)

Node Degree: K

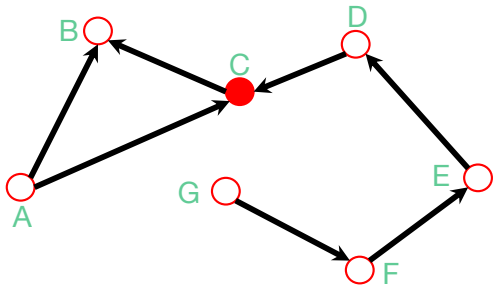
Undirected graphs

The number of links connected to the node



Directed graphs

- in-degree: incoming links
- out-degree: outgoing links
- total degree: sum of in- and out-degree



$k_C^{in} = 2$

$k_C^{out} = 1$

$k_C = 3$

Source: a node with $k^{in} = 0$ (es. A)

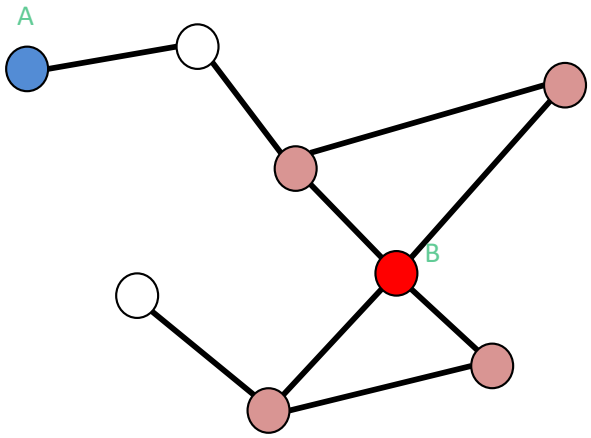
Sink: a node with $k^{out} = 0$ (es. B)

Average Node Degree: <K>

N: # nodes **L**: # links

Undirected graphs

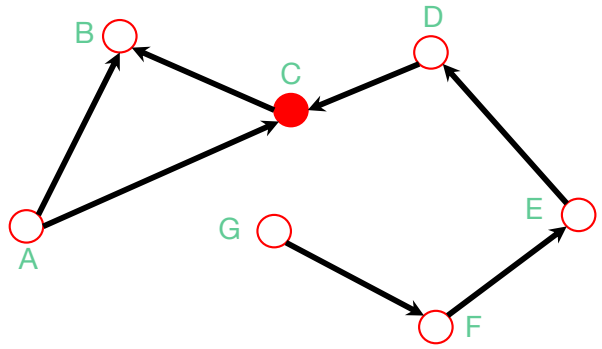
$$\langle \boldsymbol{k} \rangle \equiv \frac{\mathbf{1}}{N} \sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$



Directed graphs

$$\langle \boldsymbol{k}^{in} \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} k_i^{in} \qquad \langle \boldsymbol{k}^{out} \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{k}_i^{out}$$

$$\langle \boldsymbol{k}^{in} \rangle = \langle \boldsymbol{k}^{out} \rangle \quad \Rightarrow \quad \langle k \rangle = \frac{L}{N}$$

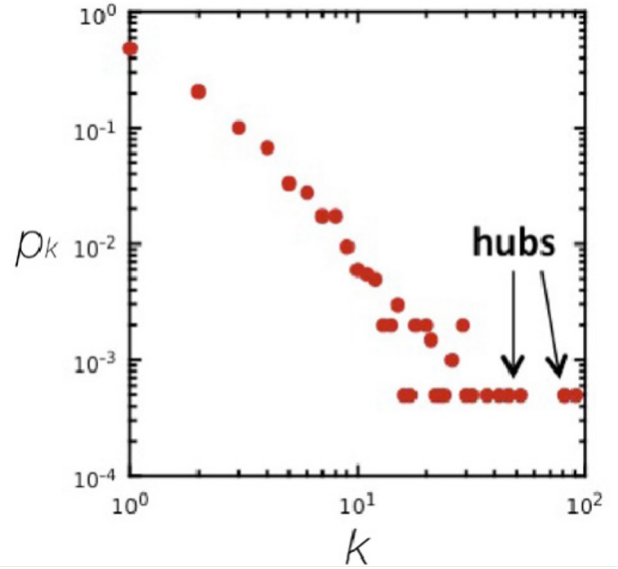


Degree Distribution

P(k): probability that a randomly chosen node has degree *k*

$\mathbf{N_k}$ = # nodes with degree k

$$\mathbf{P(k) = N_k / N}$$



Hubs: nodes with many edges or with edges that place them in central positions for facilitating traffic over a network

↓

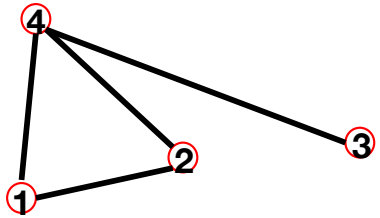
Nodes with high degree

Adjacency Matrix: A

Undirected graphs

1 if nodes i and j are adjacent

0 otherwise



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$k_i = \sum_{j=1}^{N} A_{ij} \qquad k_j = \sum_{i=1}^{N} A_{ij}$$

$$A_{ij} = A_{ji}$$

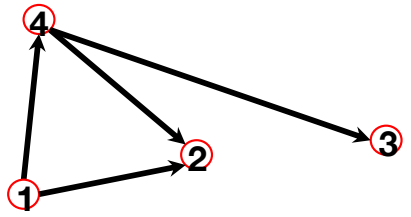
$$A_{ii} = 0$$

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{ij} A_{ij}$$

Directed graphs

1 if node i has an incoming (outgoing) edge from (to) node j

0 otherwise



$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$k_i^{in} = \sum_{j=1}^{N} A_{ij} \qquad k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

$$A_{ij} \neq A_{ji}$$

$$A_{ii} = 0$$

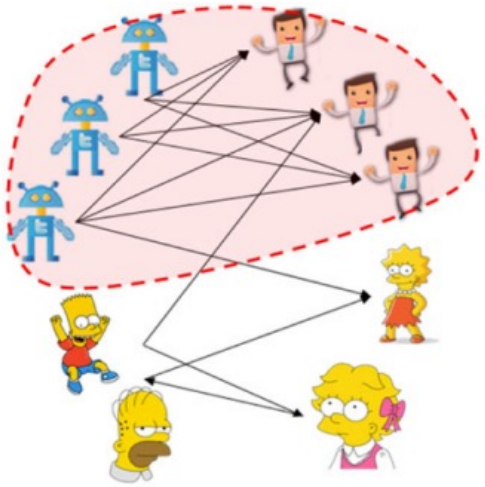
$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j} A_{ij}$$

A step back...

How can we recognize bots?

- Possible networks:
  - retweeters of a set of tweets;
  - followers of a set of accounts;
  - likers of a set of pages on Facebook;
  - users that shared a set of URLs;
  - …

Twitter-Style Network



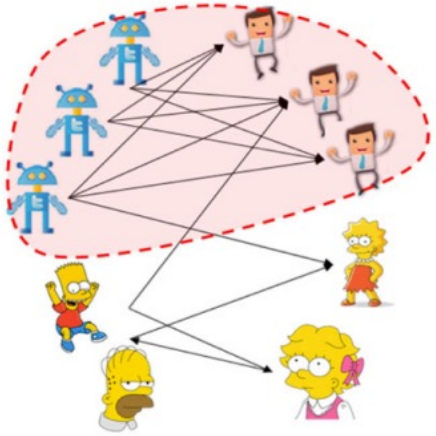
Undirected/directed Graph

	0	0	0	1	1	1	0	0	0	0
	0	0	0	1	1	1	0	0	0	0
	0	0	0	1	1	1	0	0	0	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	0	1	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0

A step back...

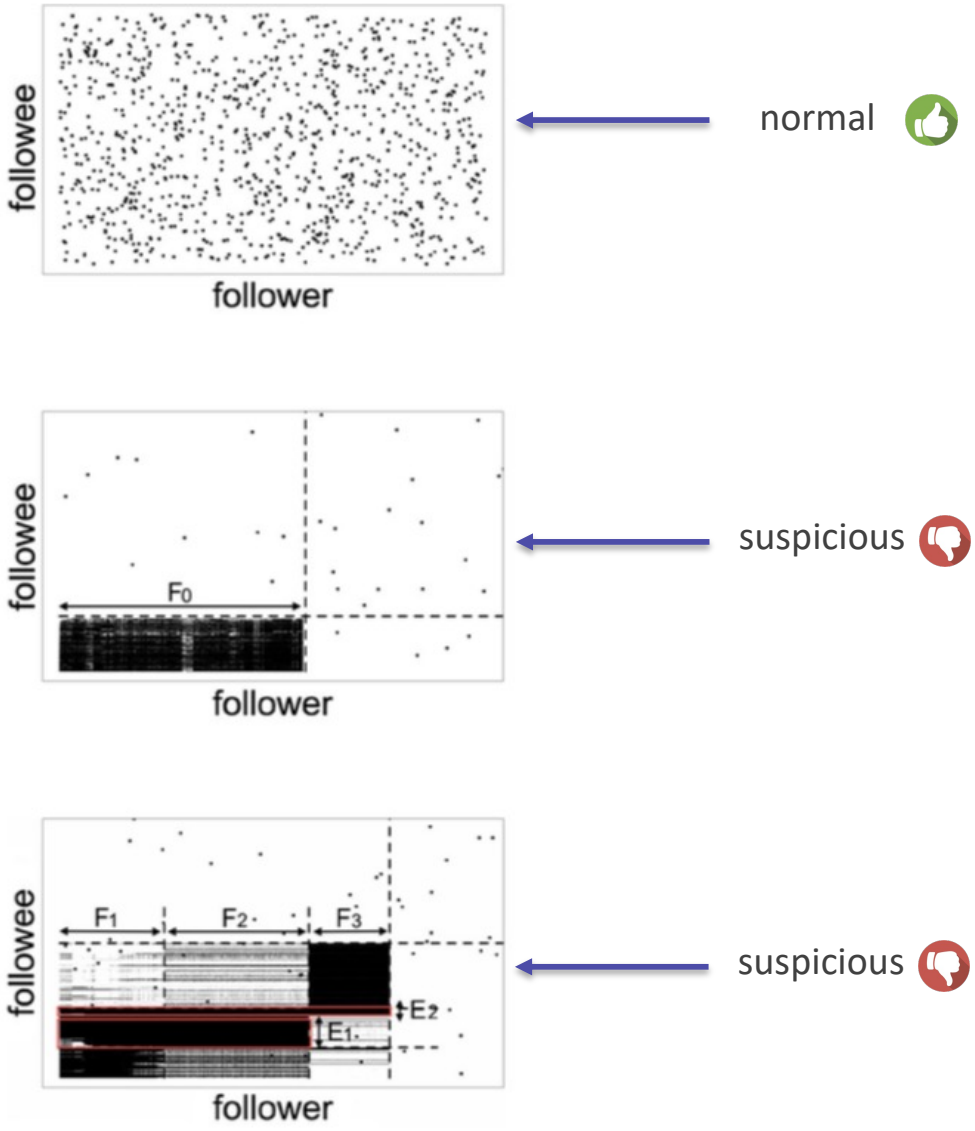
Intuition: automated accounts have patterns in the adjacency matrix

Twitter-Style Network



Undirected/directed Graph

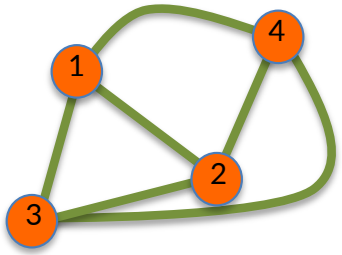
	0	0	0	1	1	1	0	0	0	0
	0	0	0	1	1	1	0	0	0	0
	0	0	0	1	1	1	0	0	0	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	0	1	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0



Complete Graph

If the graph **contains all the possible links**, is called **COMPLETE**

$$L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$



$L=L_{max}$

<k>=N-1

$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

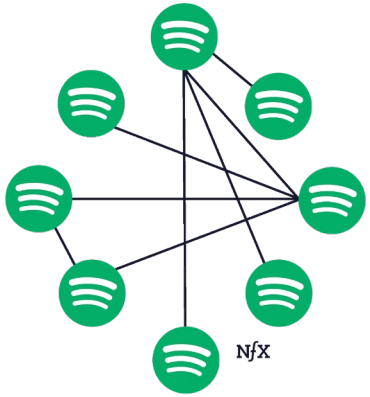
Density: D

Density: ratio of existing edges over possible ones

A graph with low density is called **SPARSE**

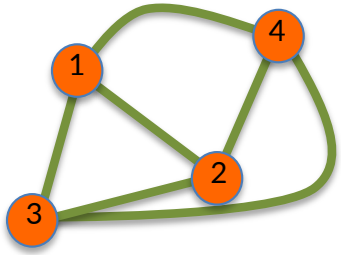
$$L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$

$$d(G) = \frac{L}{L_{max}}$$

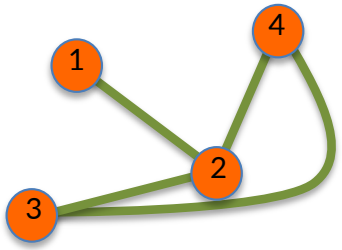


Low Density

High Density



Complete graph
d(G) = 1

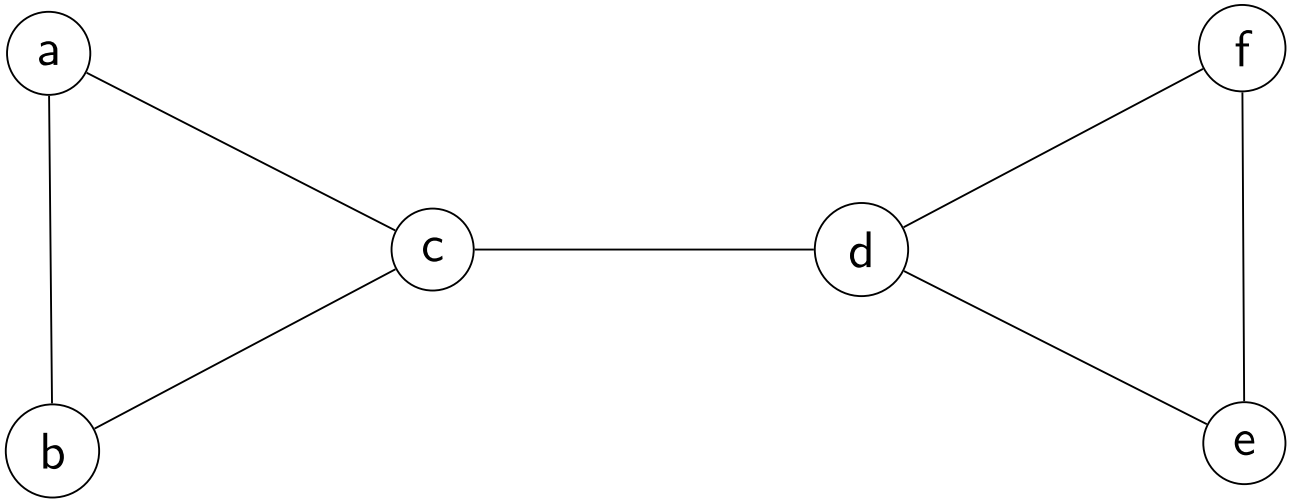


Sparse graph
d(G) = 2/3

Global Clustering Coefficient: C

- The global clustering coefficient indicates how "clustered" the network is, accounting for **triangles and triplets**
- **Triangles:** set of 3 nodes connected by 3 edges
- **Triplets:** set of 3 nodes connected by 2 edges (include triangles!)
- *What is the global clustering coefficient of this graph?*

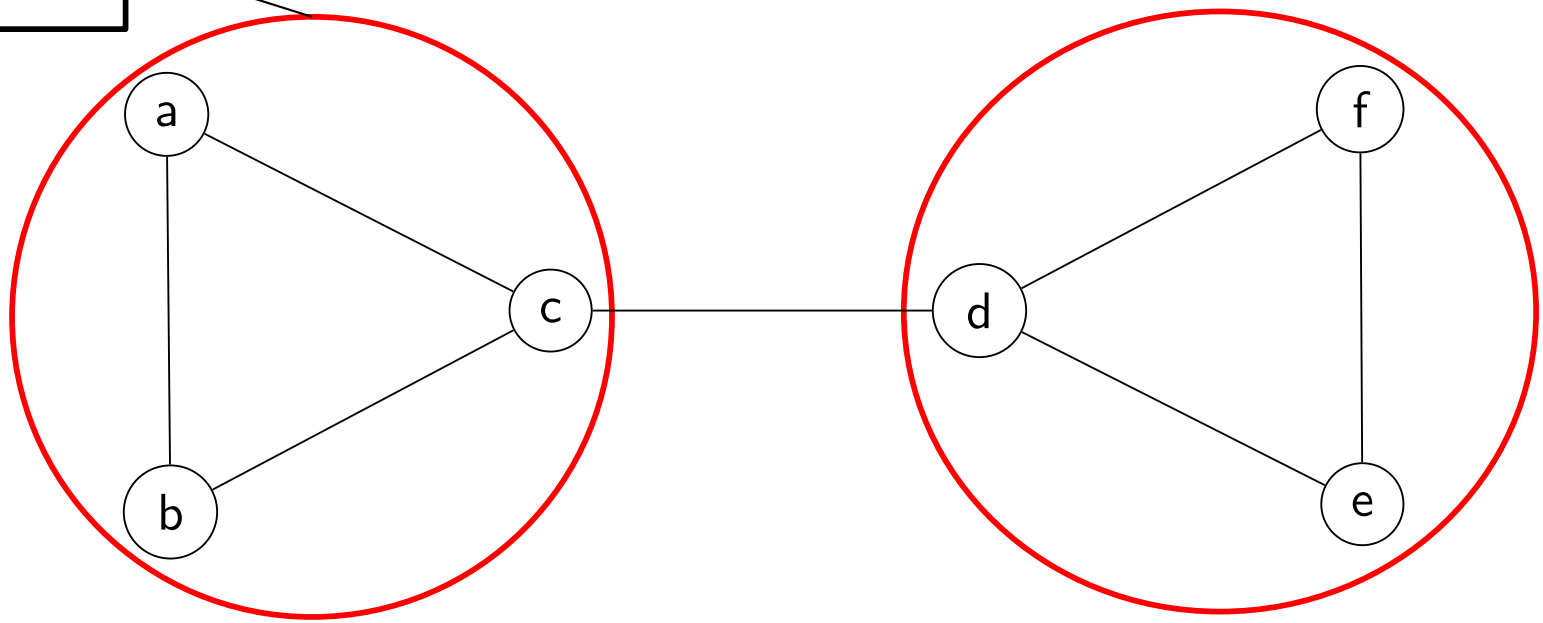
$$C = \frac{3 \times \text{number of triangles}}{\text{number of all triplets}} \longrightarrow \text{C in [0,1]}$$



Global Clustering Coefficient: C

1 triangle
3 combinations:
A-C-B
C-B-A
B-A-C

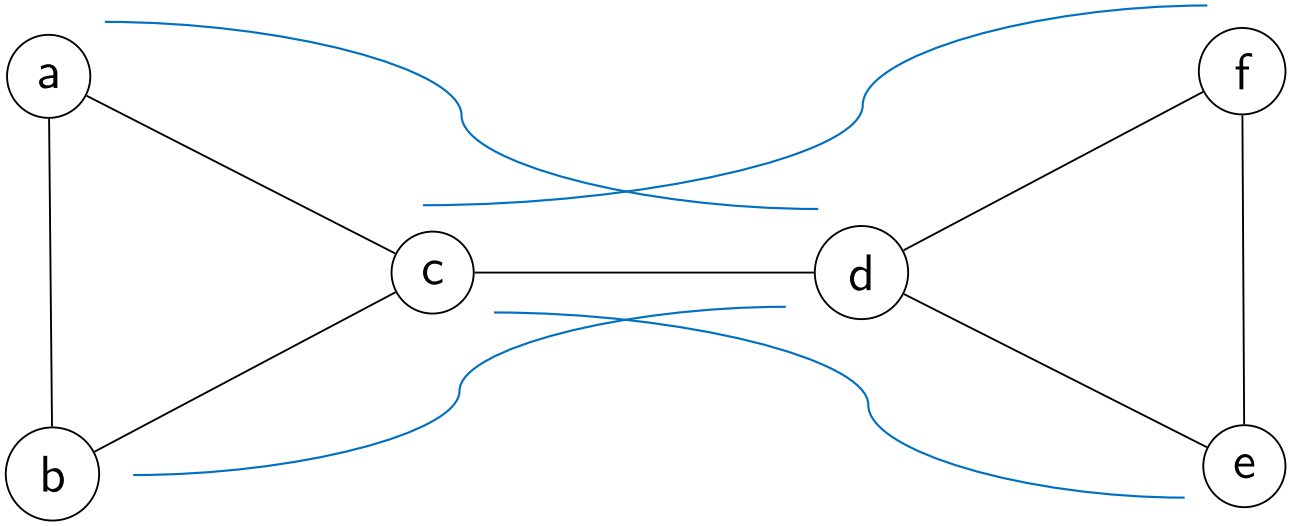
$$C = \frac{3 \times \text{number of triangles}}{\text{number of all triplets}}$$



2 triangles
3 triplets for each triangle

Global Clustering Coefficient: C

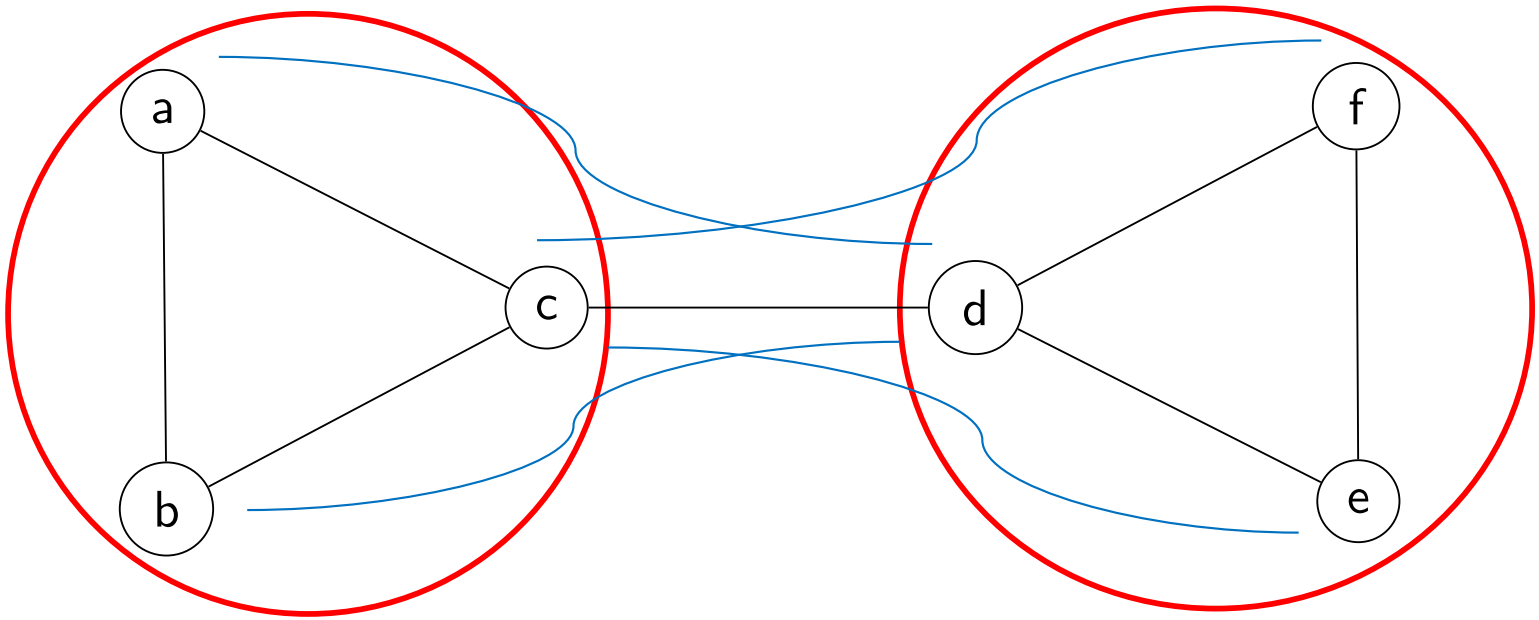
$$C = \frac{3 \times \text{ number of triangles}}{\text{number of all triplets}}$$



4 missing triplets

Global Clustering Coefficient: C

$$C = \frac{3 \times \text{ number of triangles}}{\text{number of all triplets}}$$



C = 6/(6+4) = 6/10 = 3/5 = 0,6

Local Clustering Coefficient: C_i

- The local clustering coefficient is computed for each node
- *Clustering coefficients of nodes A and C?*
- *Average local clustering coefficient?*

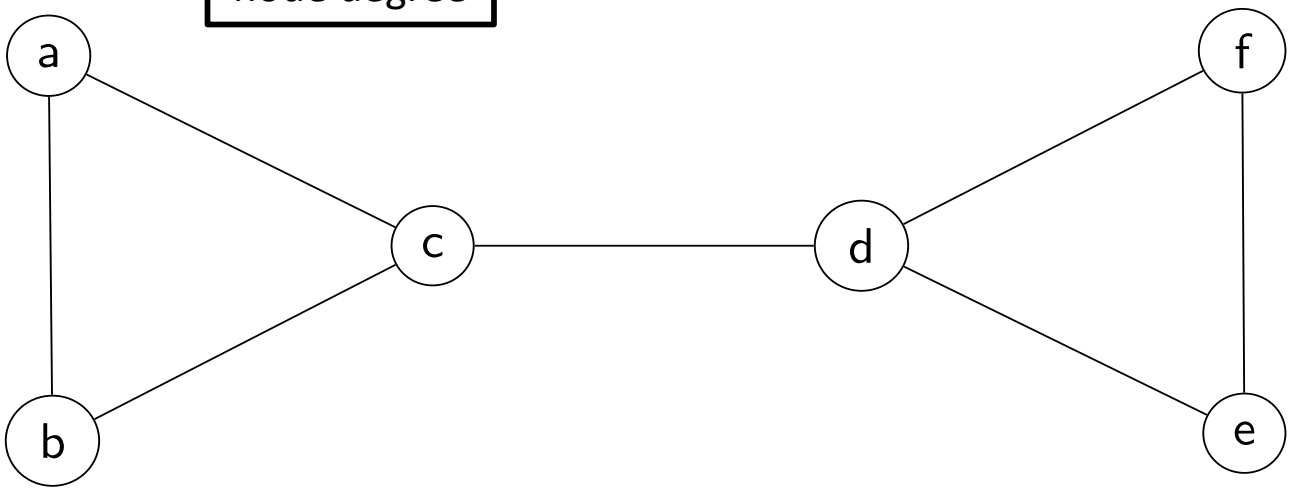
$$C_i = \frac{2e_i}{k_i(k_i-1)}$$

triangles involving the node i

C_i in [0,1]

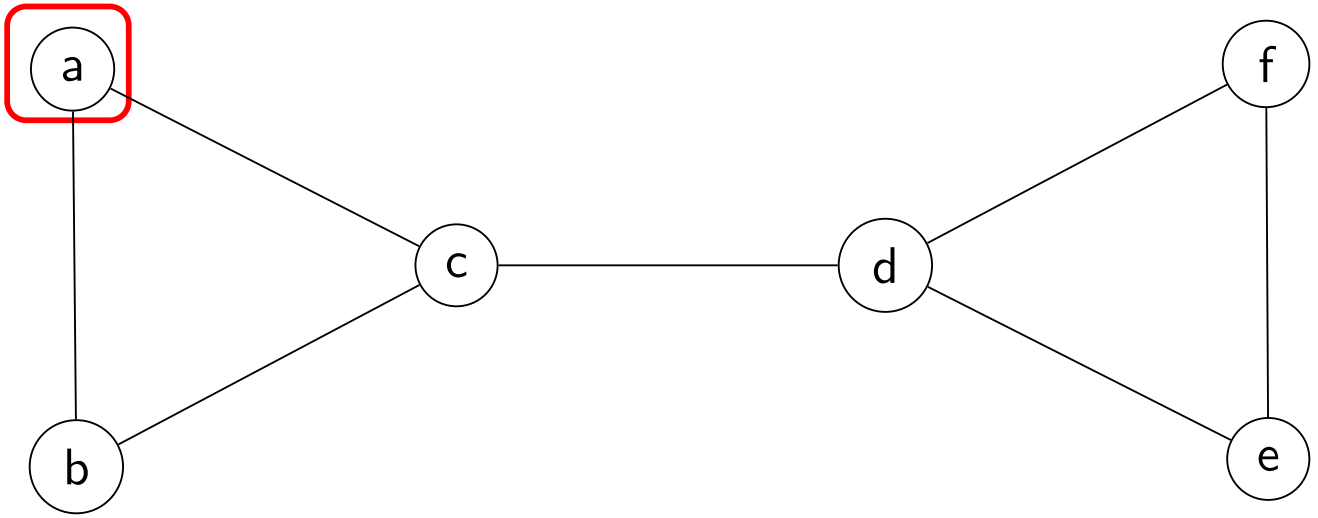
triplets involving the node i

node degree



Local Clustering Coefficient: C_i

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



Node A

E_i = 1

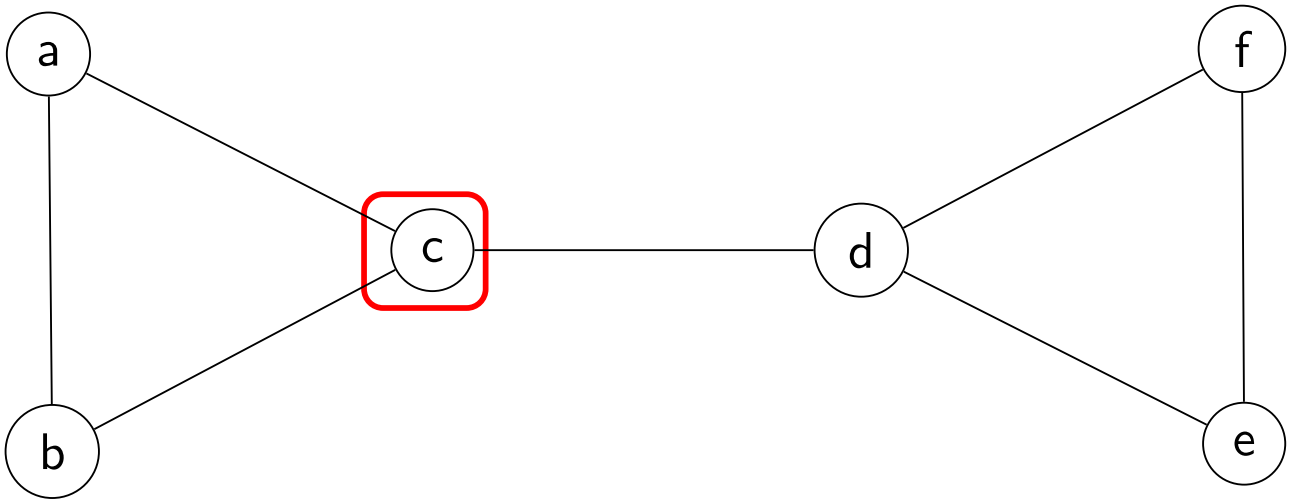
K_i = 2

Triplets = 2

C_i = 1

Local Clustering Coefficient: C_i

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



Node C

E_i = 1

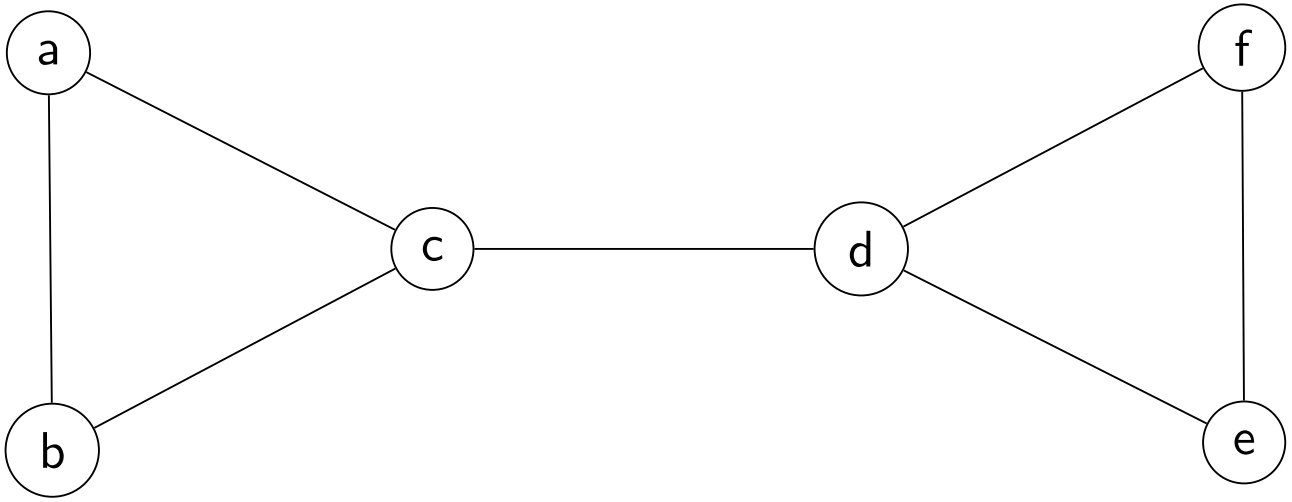
K_i = 3

Triplets = 6

C_i = 1/3

Local Clustering Coefficient: C_i

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



Average clustering coefficient:

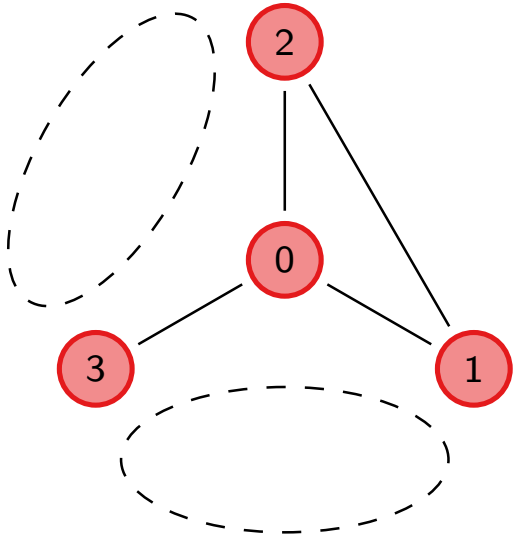
4 nodes (A-B-E-F) with $C_i = 1$

2 nodes (C-D) with $C_i = 1/3$

<C> = 7/9 = 0,77

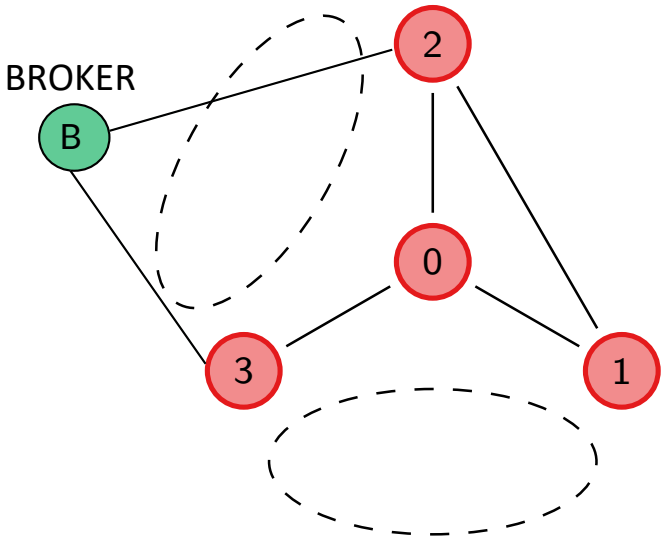
Brokers

- "Missing links" in a graph are called **STRUCTURAL HOLES**
- A node that fills a structural hole is called local bridge or **BROKER**
- Brokers are common in low-clustered graphs



Brokers

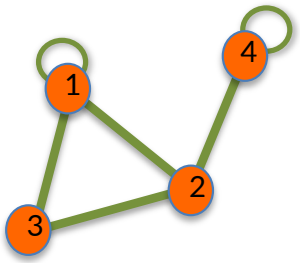
- "Missing links" in a graph are called **STRUCTURAL HOLES**
- A node that fills a structural hole is called local bridge or **BROKER**
- Brokers are common in low-clustered graphs




NETWORK FEATURES

Loops

A node could be connected to itself
(ex. protein-protein interaction network)



$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

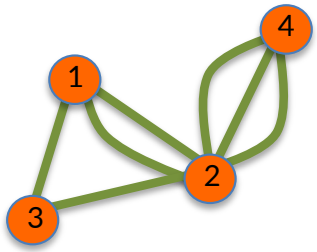
$$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii}$$

Directed graphs could further present loops

Multigraph

A node could have multiple connections to another node (ex. street path networks)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

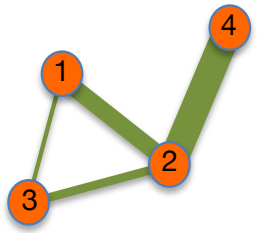
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad < k >= \frac{2L}{N}$$

We can also have directed multigraphs

Weighted Graphs

The paths between nodes could present different weights (ex. highway networks)



$$A_y = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

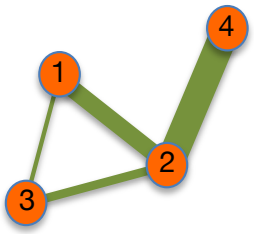
$$L = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad < k >= \frac{2L}{N}$$

Directed graphs could further present weighted links

Weighted Graphs

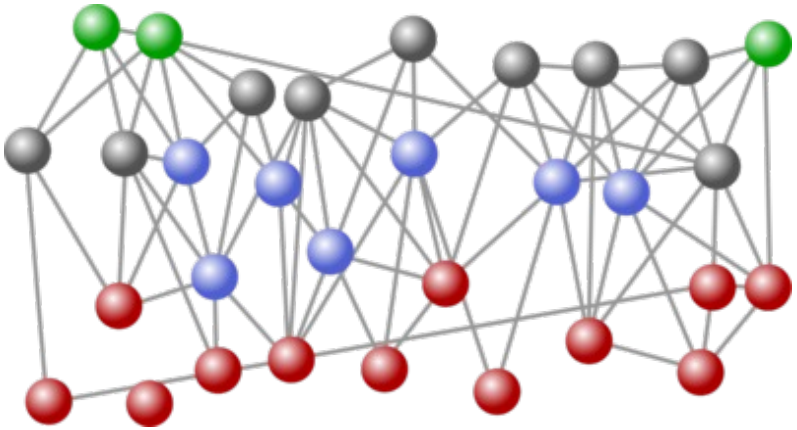
In weighted graphs, we could define the node **STRENGTH** (S)

- Degree K: considers 1 for each connected node
- Strength S: considers Aij for each connected node



$$A_y = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

Real Networks: Examples

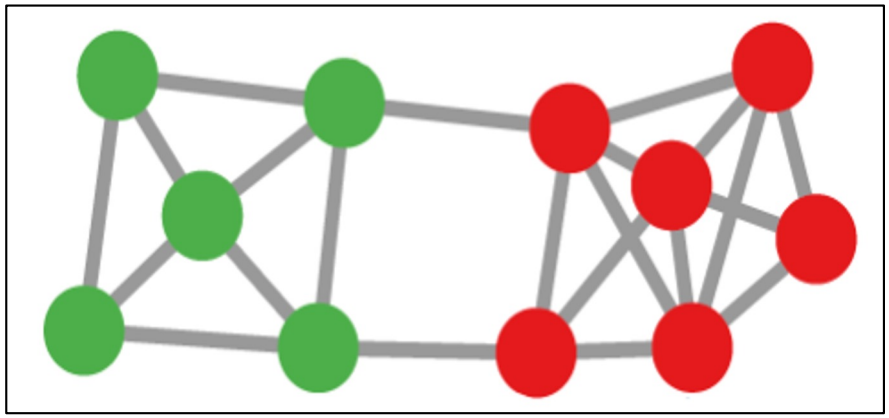


Most networks observed in real systems are **sparse**

Network	Directed	Weighted	Multigraph	Self-loops
WWW	yes	no	yes	yes
Protein interactions	no	no	no	yes
Collaboration network	no	yes	yes	no
Mobile phone calls	yes	yes	no	no
Facebook Friendship	no	no	no	no

Categorical Networks

- The nodes of a network could belong to different categories
- Connections between nodes of different categories are allowed
  - Ex. mobile phone network accounting for communication providers



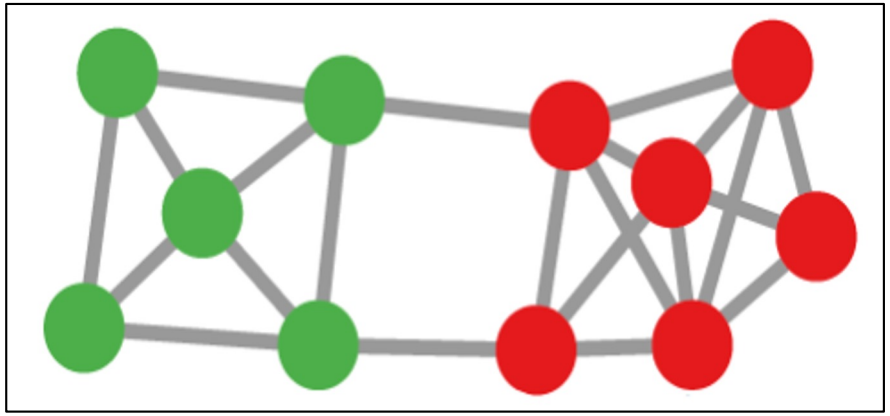
Assortativity: R

- To what extent are nodes belonging to the same category connected to each other?
- Assortativity quantifies link homophily

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

- e_{ij} fraction of links connecting nodes of type i and j
- a_i fraction of out-links from nodes of type a
- b_i fraction of in-links for type b nodes

- R = 0: no assortative mixing
- R = 1: perfectly assortative
- -1 < R < 0: disassortative mixing



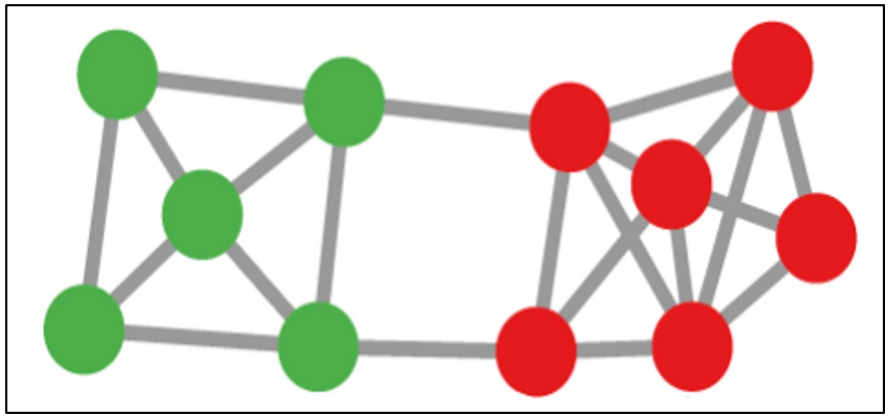
Degree Assortativity

- In the degree assortativity, there are no distinct categories: the category of nodes reflects their degree level (high/not high)
- N.B. in undirected graphs $a_i = b_i$
- *What is the degree assortativity of the graph below?*

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

- e_{ij} fraction of links connecting nodes of type i and j
- a_i fraction of out-links from nodes of type a
- b_i fraction of in-links for type b nodes

- R = 0: no assortative mixing
- R = 1: perfectly assortative
- -1 < R < 0: disassortative mixing

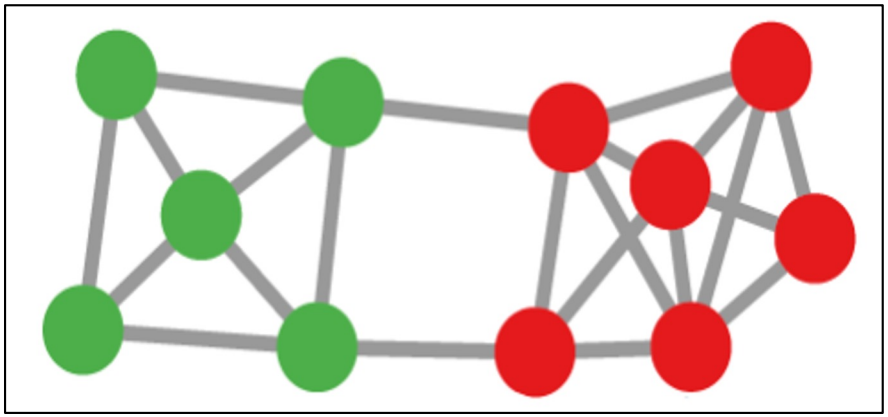


Degree Assortativity

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

- e_{ij} fraction of links connecting nodes of type i and j
- a_i fraction of out-links from nodes of type a
- b_i fraction of in-links for type b nodes

- R = 0: no assortative mixing
- R = 1: perfectly assortative
- -1 < R < 0: disassortative mixing



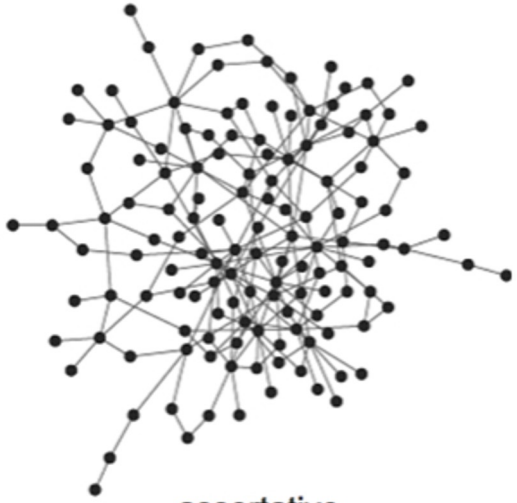
Undirected graph
$A_i = B_i$

$$r = \frac{((8/22)+(12/22))-((10/22)^2+(14/22)^2)}{1-((10/22)^2+(14/22)^2)} = \text{~0.766}$$

Degree Assortativity

Assortative Mixing

Nodes tend to connect homogeneously w.r.t. their degree (e.g., hubs with hubs)



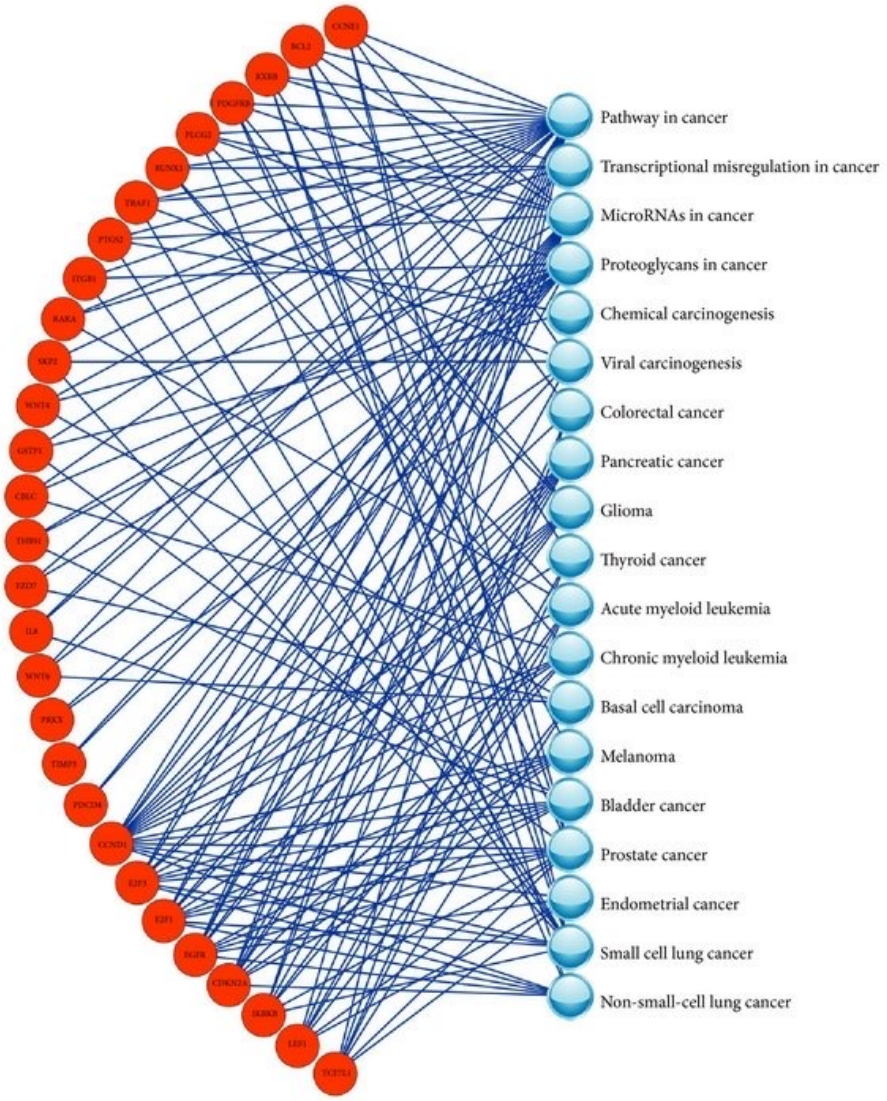
assortative

Disassortative Mixing

Nodes tend to connect in a star-like topology

disassortative

Bipartite Networks



«The gene-pathway bipartite network constructed with 29 gene signatures that were used for predicting the reoperative treatment response of breast cancer»

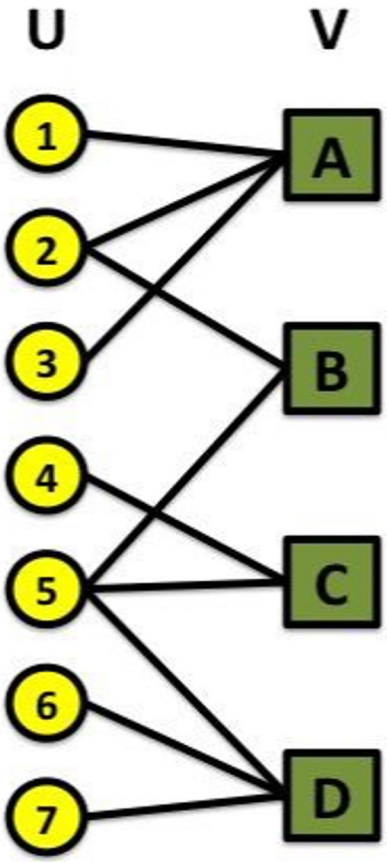
- Categorical network with two **disjoint** sets of nodes: U and V
- Nodes in U could connect **only** to nodes in V and vice-versa (differently from classic categorical networks)
- Used to model different object categories

Bipartite Networks

- **Projection of a bipartite network through the category U (V):** connects two nodes of U (V) which are connected to the same object in V (U)

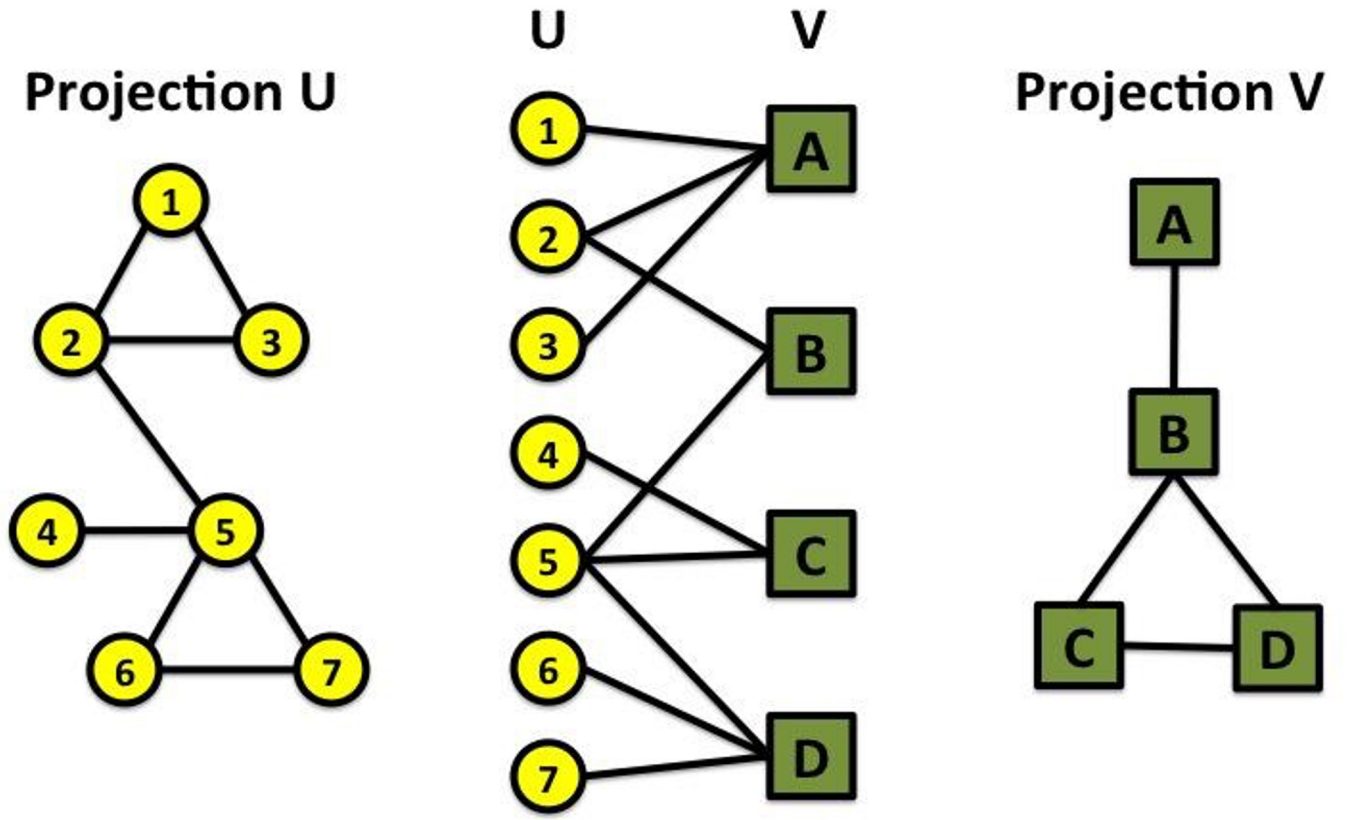
- **If weighted**: edge weight is the number of shared connections

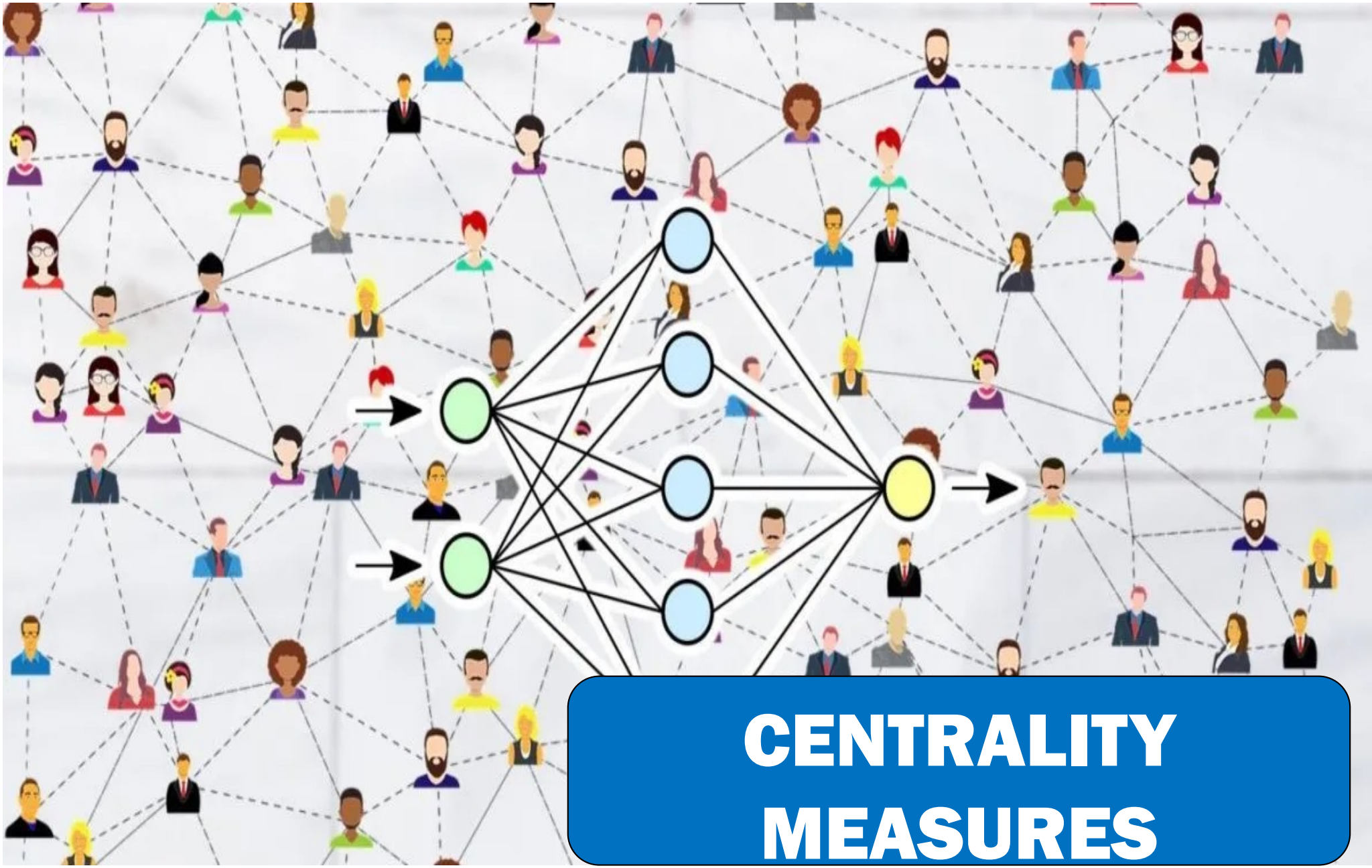
- *What is the projection of this network through the categories U and V?*



Bipartite Networks

- **Projection of a bipartite network through the category U (V):** connect two nodes of U (V) which are connected to the same object in V (U)



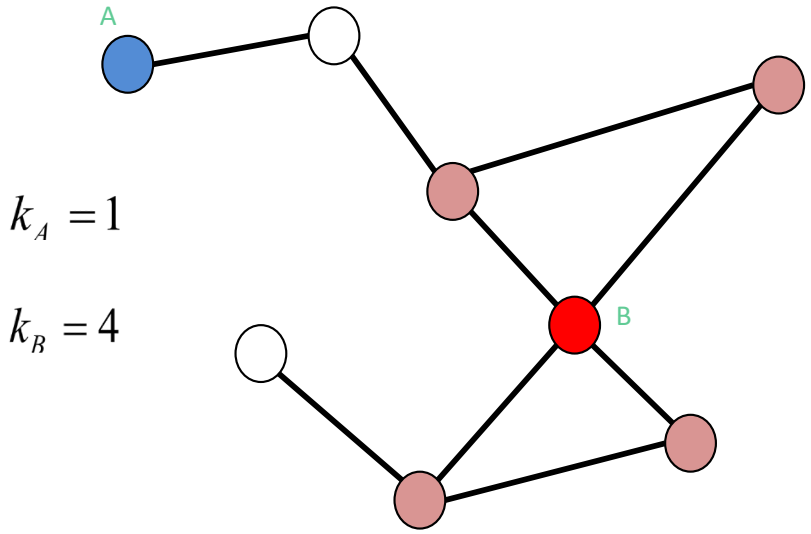

CENTRALITY
MEASURES

Centrality Measures

- We can measure nodes' importance using the so-called centrality
- Nothing to do with being central in general
- Widely used for machine-learning tasks

Degree Centrality

- Important nodes have a high degree
  - main characters of a tv series talk with more people
- Not always…
  - Twitter users with the most contacts are often spam
  - Webpages/wikipedia pages with most links are often lists of references



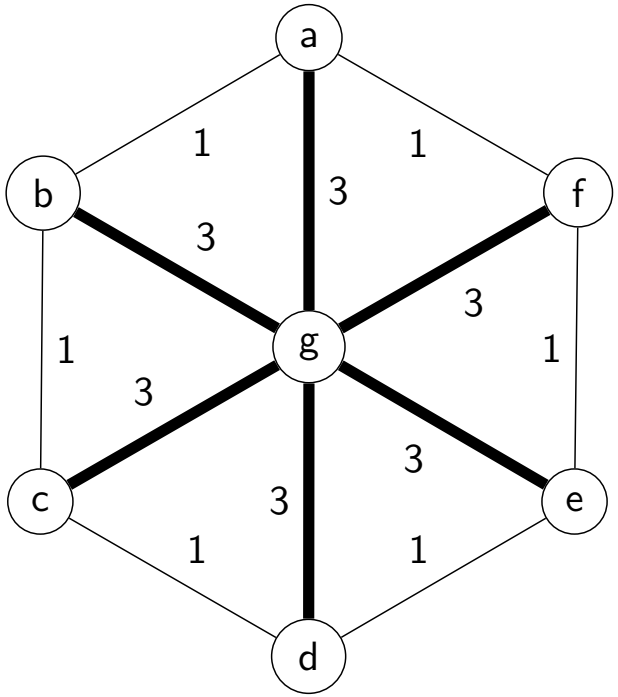
Eigenvector Centrality: X_v

- Nodes connected to important nodes are central
- Specific for weighted networks

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

$$\mathbf{Ax} = \lambda \mathbf{x}$$

- M(v) is the set of v's neighbours
- λ is the largest eigenvalue of Ax

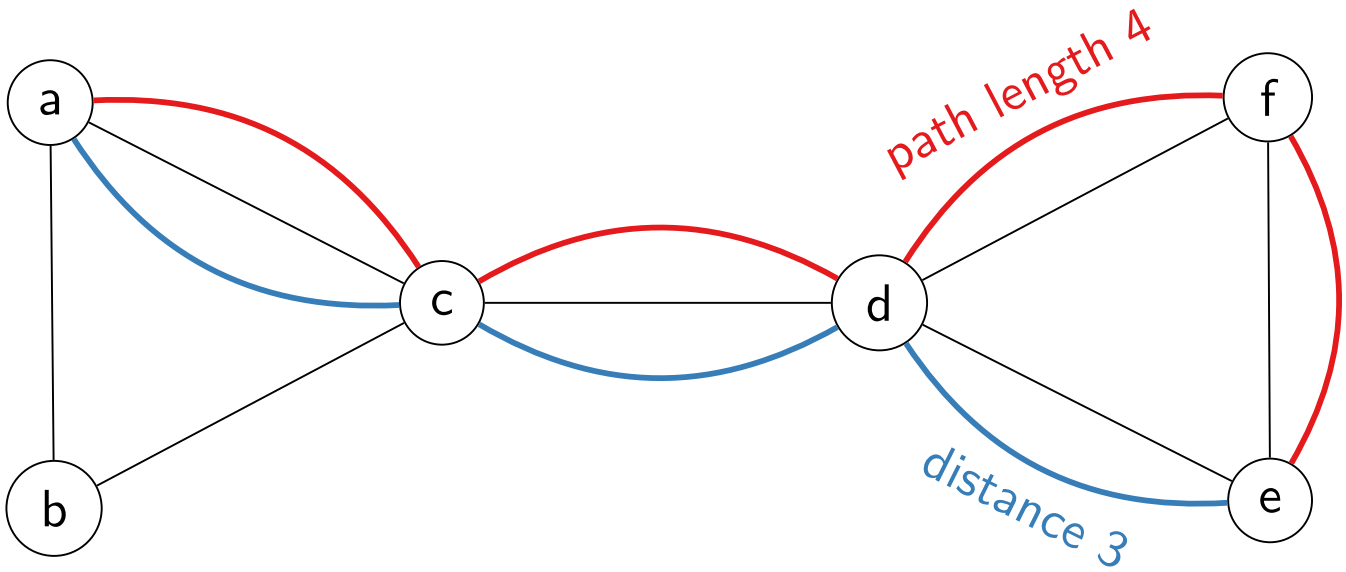


$X_V(a) = 0,31$

$X_V(g) = 0,65$

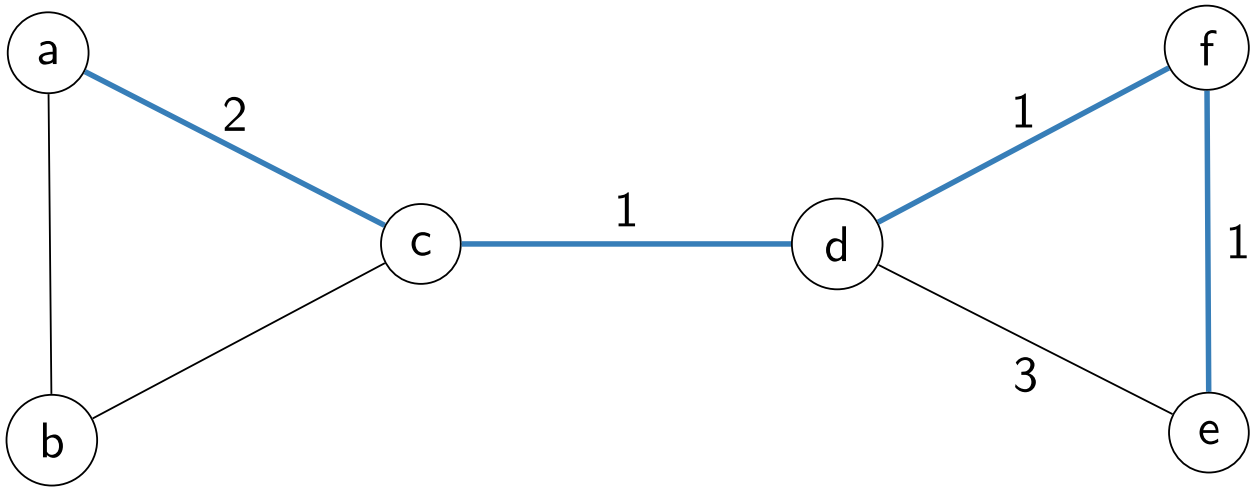
Path and Distance

- Path and distance between nodes A and E
- In unweighted networks, the path with the lowest geodesic distance is called the **SHORTEST PATH**
- The longest shortest path is called network **DIAMETER**



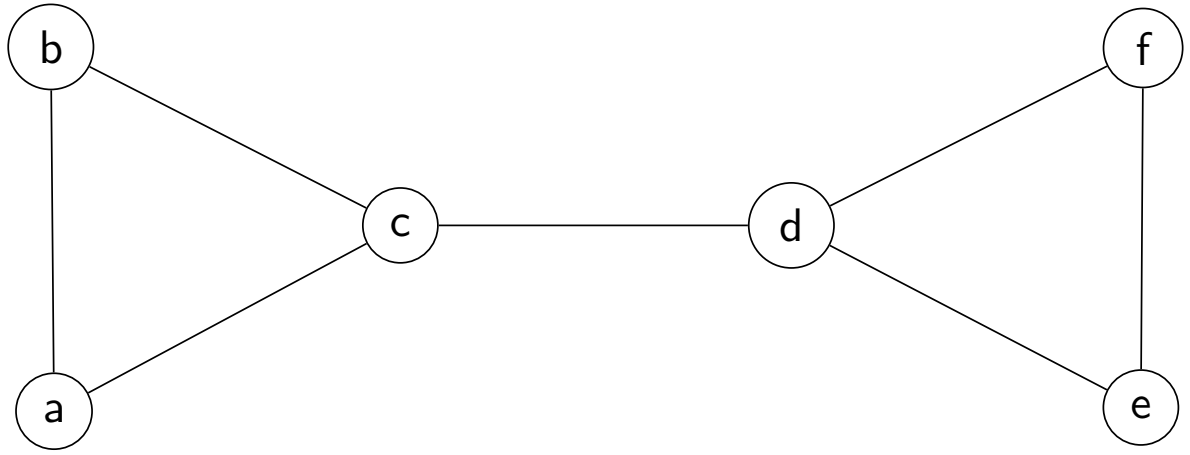
Path and Distance

- In weighted networks, the distance should account for the weights
- Also the shortest path should account for the weights
- <u>The diameter doesn't account for the weights!</u>



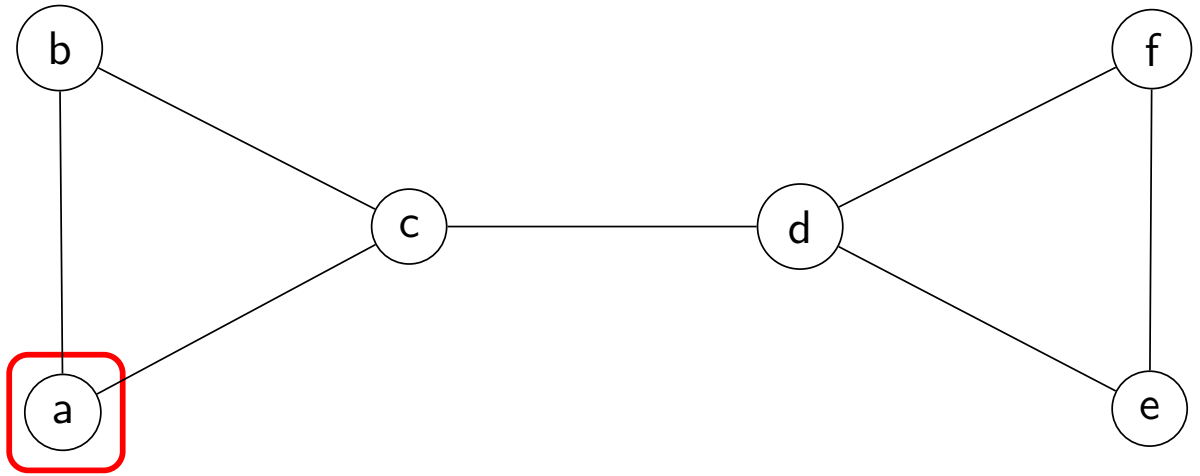
Closeness Centrality: X_c

- The closeness centrality is the inverse of the average distance from a node and all the network
- A node is central if is near to the other nodes
- *Closeness centrality of nodes A and C?*



Closeness Centrality: X_c

- The closeness centrality is the inverse of the average distance from a node and all the network
- A node is central if is near to the other nodes



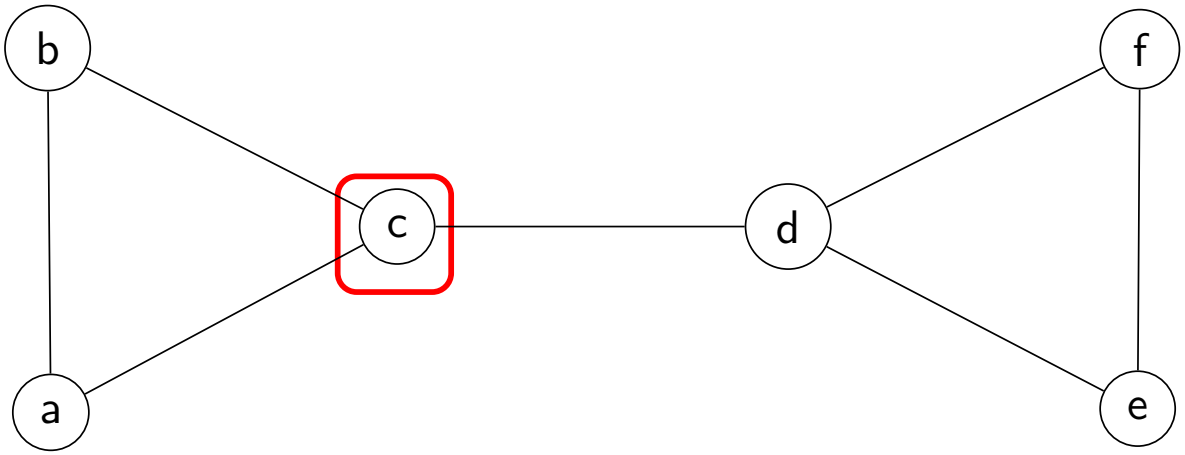
Closeness centrality of A

A-B = 1
A-C = 1
A-D = 2
A-E = 3
A-F = 3

Sum of distances: 10
Average distance: 10/5 = 2
X_C(a): 1/2 = 0,5

Closeness Centrality

- The closeness centrality is the inverse of the average distance from a node and all the network
- A node is central if is near to the other nodes



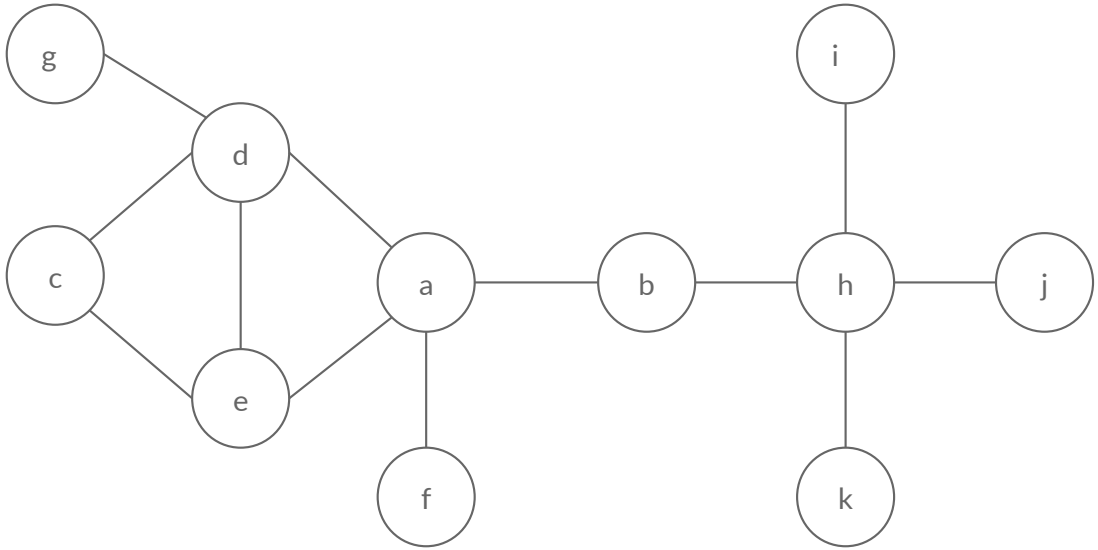
Closeness centrality of C

C-A = 1
C-B = 1
C-D = 1
C-E = 2
C-F = 2

Sum of distances: 7
Average distance: 7/5 = 1,4
$X_C(c)$: 1/(7/5) = 0,71

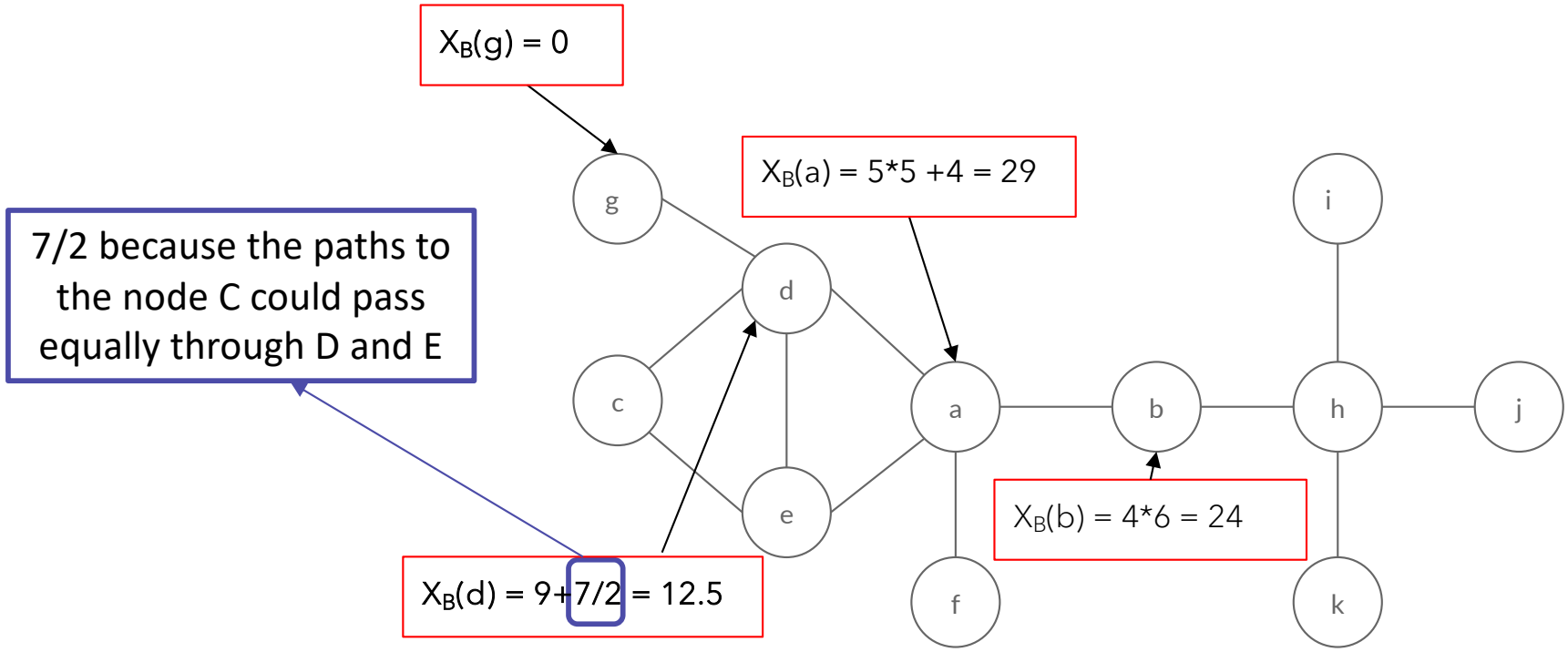
Betweenness Centrality: X_B

- Number of shortest paths that go through a node
- Important vertices are **bridges** over which information flows
- If the shortest path could equally pass through M different nodes, the contribution is 1/M
- *Betweenness centrality of nodes A, B, D and G?*



Betweenness Centrality: X_B

- Number of shortest paths that go through a node
- Important vertices are **bridges** over which information flows
- If the shortest path could equally pass through M different nodes, the contribution is 1/M



Centrality Measures

- Each centrality measure is a **proxy** of an underlying network process.
- If such a process is irrelevant to the network then the centrality measure makes no sense
  - Ex. If the information does not spread through the shortest paths, betweenness centrality is irrelevant
- Centrality measures should be used with caution:
  - (a) for exploratory purposes
  - (b) for characterization