ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Graph Theory & Social Network Analysis

Professor George M. Giaglis
Athens University of Economics & Business

Outline

Network properties

- *Adjacency matrices*
- *Paths, shortest paths*
- *Network diameter*

Node properties

- *Degree*
- *Centrality*
- *Clustering coefficient*

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

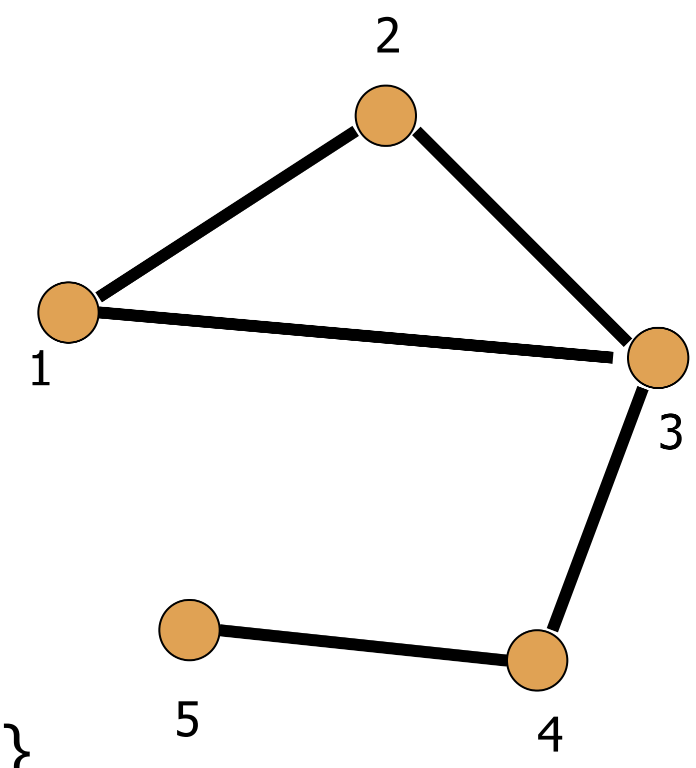
Network properties

Adjacency matrices
Paths, shortest paths
Network diameter

Networks as Graphs

Graph G = (V,E)

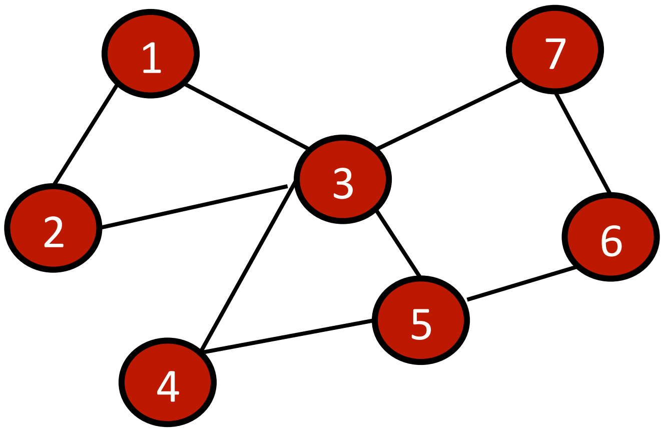
- V = set of nodes
- E = set of edges



V = {1, 2, 3, 4, 5}
E = {(1,2), (1,3), (2,3), (3,4), (4,5)}
G = (V,E)

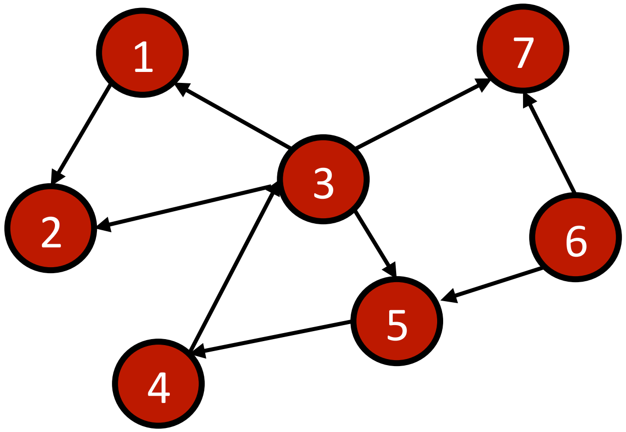
Networks as graphs

Undirected



An **undirected graph** is one in which edges have ***no orientation.*** The edge (a, b) is identical to the edge (b, a).

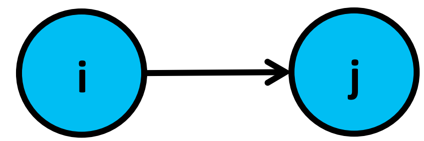
Directed



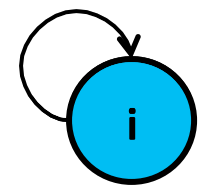
a directed graph is a graph, or set of nodes connected by edges, where the edges ***have a direction associated with them.***

Adjacency matrix

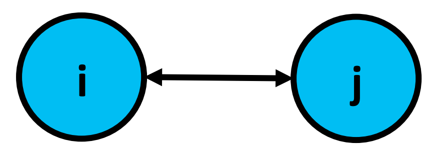
- **Representing edges (who is adjacent to whom) in a matrix**
  - $\mathbf{A_{ij}}$ **= 1** if node *i* has an edge to node *j*
    = 0 if node *i* does not have an edge to j



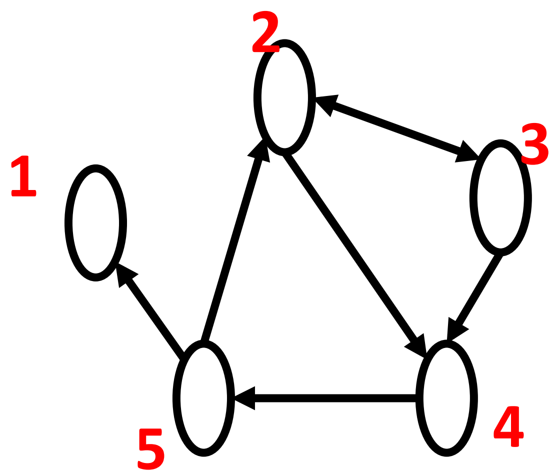
  - $\mathbf{A_{ii}}$ **= 0** unless the network has self-loops



  - $\mathbf{A_{ij}}$ **=** $\mathbf{A_{ji}}$ if the network is undirected, or if *i* and *j* share a reciprocated edge



Adjacency matrix example



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

Walks, Paths, Cycles, and Geodesics

- **Walk from i_1 to i_K:** a sequence of nodes $(i_1, i_2, \ldots i_K)$ and a sequence of links $(i_1 i_2, i_2 i_3, \ldots, i_{K-1} i_K)$ such that $i_{k-1} i_k$ in E for each k

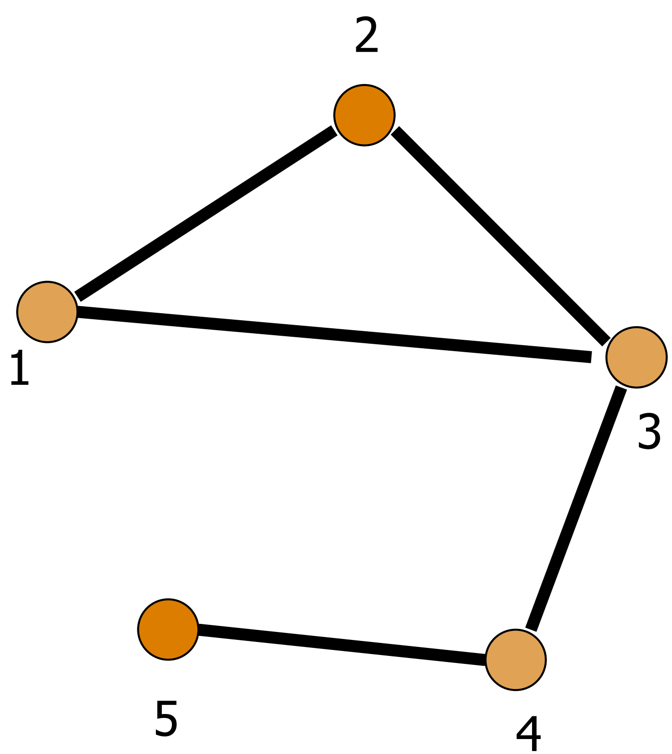
- **Path:** a walk $(i_1, i_2, \ldots i_K)$ with each node i_k is distinct

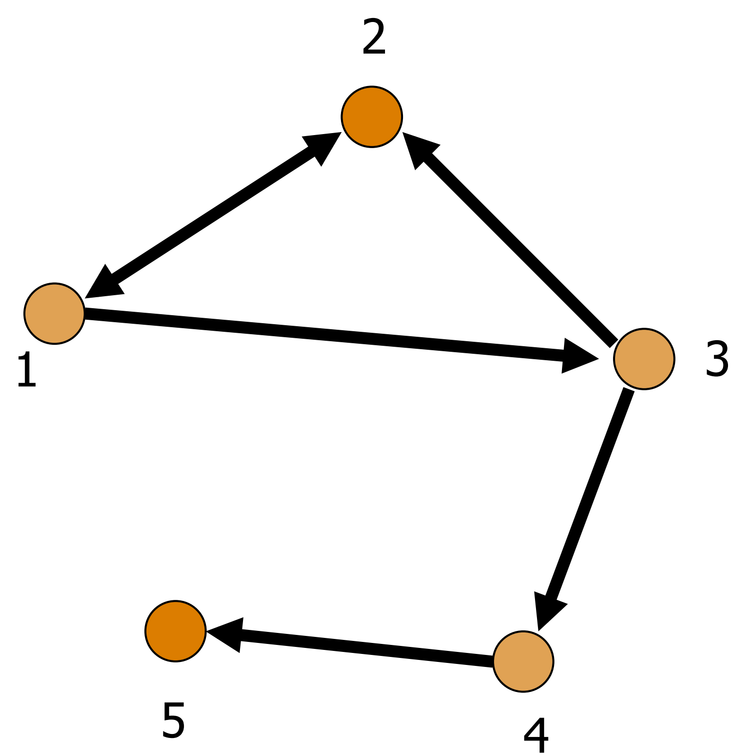
- **Cycle:** a walk where $i_1 = i_k$

- **Geodesic:** a shortest path between two nodes

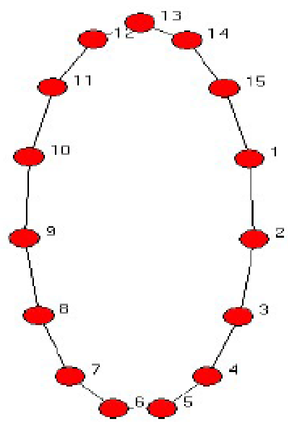
Network Diameter

- **Diameter = the *longest shortest path* in the network**
  - Represents a worst-case scenario in network size
  - Left example (undirected network): diameter=?
  - Right example (directed network): diameter=?

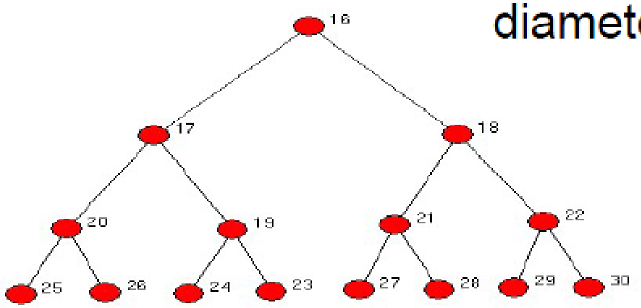




Diameter scenarios



diameter is either
n/2 or (n-1)/2



K levels has $n = 2^{K+1}-1$ nodes
so, $K = \log_2(n+1) - 1$
diameter is 2K

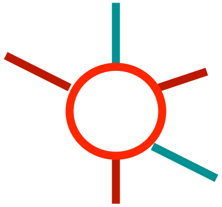
diameter is on order of
$2 \log_2(n+1)$

Node properties

Degree
Centrality
Clustering Coefficient

Degree of Nodes

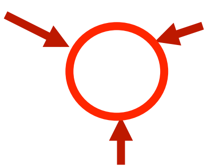
Degree: number of edges incident on a node



degree=5

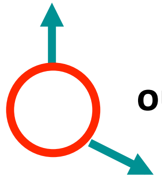
Two different degree types in directed networks

1. **Indegree :** how many directed edges (arcs) terminate at a node



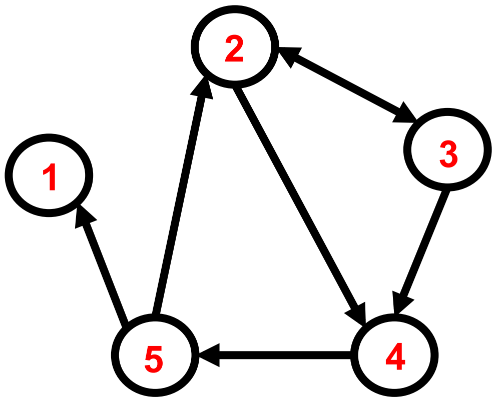
indegree=3

2. **Outdegree:** how many directed edges (arcs) originate at a node



outdegree=2

Node degree from matrix values



Outdegree

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ \boxed{0 \quad 1 \quad 0 \quad 1 \quad 0} \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Indegree

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

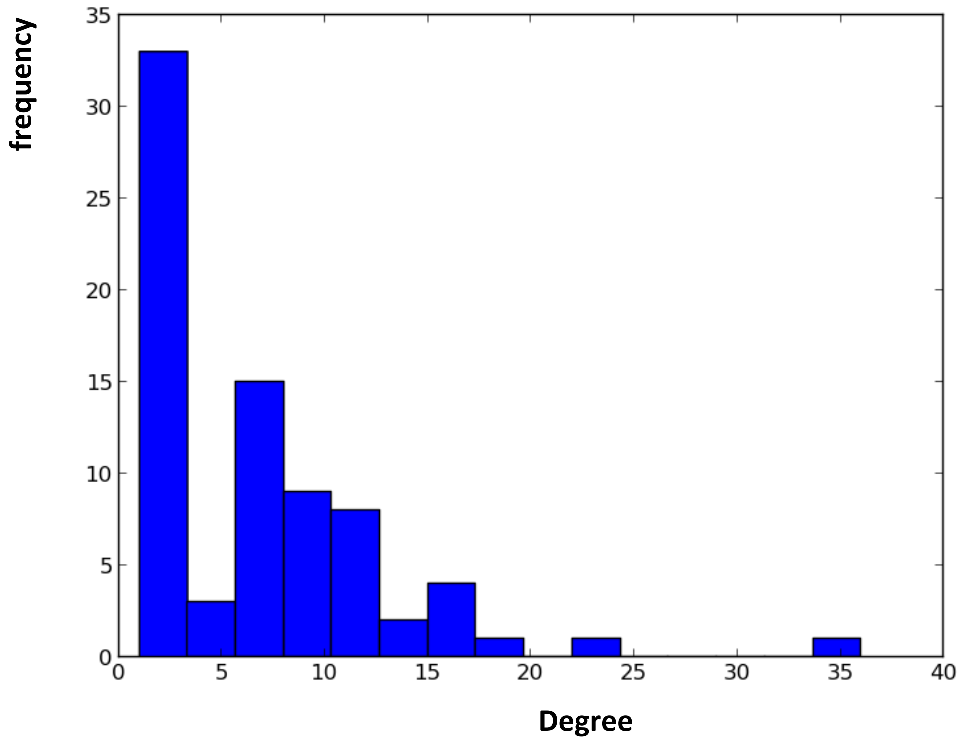
example: outdegree for node 3 is **2**, which we obtain by summing the number of non-zero entries in the 3rd *row*

example: the indegree for node 3 is **1**, which we obtain by summing the number of non-zero entries in the 3rd column

Degree distribution

- **Degree distribution**: A frequency count of the occurrence of each degree in a network
- Degree distributions are far from normal in most real-world networks (**hubs**)

Example: In the figure we witness many nodes with very small degree and few nodes with high degree, implying the presence of a ***hub*** or ***fat tail.***



Centrality Measures

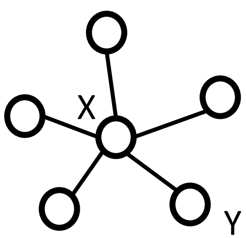
Centrality: Captures the idea of how central a node is in the network

Can be categorized into four main types

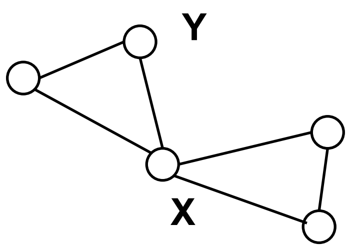
1. **Degree Centrality:** Shows how connected a node is
2. **Betweenness Centrality:** Shows how important a node is in terms of connecting other nodes
3. **Closeness Centrality:** Shows how easily a node can reach other nodes (i.e. how close the node is to the center of the network)
4. **Eigenvector / Bonacich Centrality:** Show how much a node is connected to other important nodes in the network.

Centrality Measures: An example

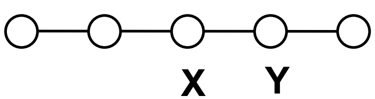
In each of the following networks, X has higher centrality than Y according to a particular measure



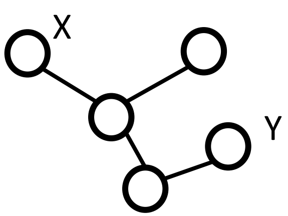
Degree



Betweenness



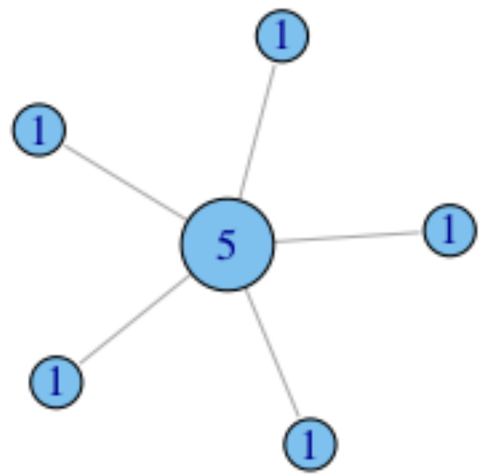
Closeness



Neighbor based

Degree Centrality

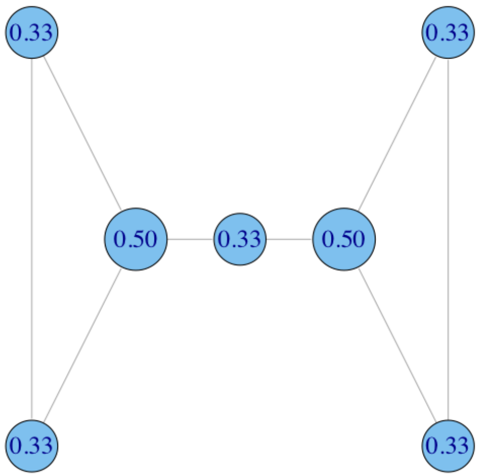
Nodes with more friends are more important

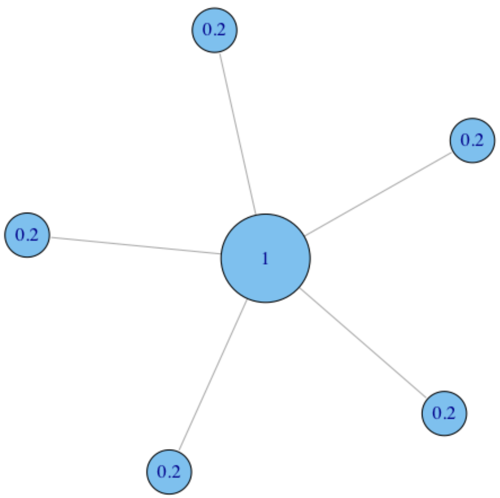


Assumption: the connections that your friends have don't matter, it is what they can do directly that does

Normalization of Degree Centrality

- Divide degree by the maximum possible (**N-1**)
- Normalized Degree Centrality ranges **from 0 to 1**
- **Allows comparisons between networks of different sizes**





Centralization: skew in distribution

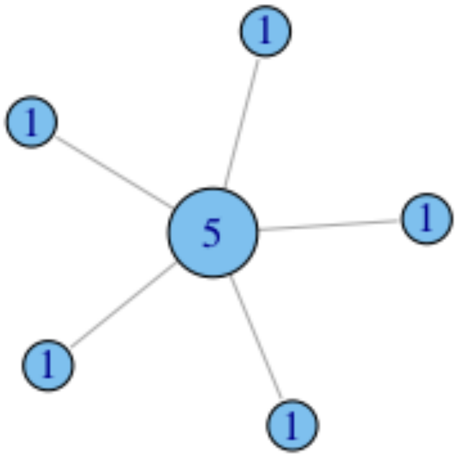
- **Centrality refers to an individual node but there is a need to capture the inequality in the distribution of centralities characterizing the entire network.**

- Using Freeman's general formula for **centralization** we can capture the inequality of degree between the nodes of the network:

maximum degree value in the network

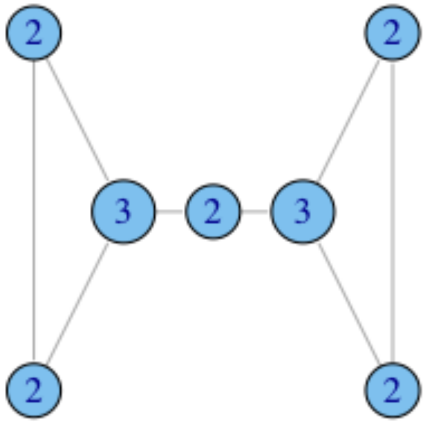
$$C_D = \frac{\sum_{i=1}^{n} \left[\boxed{C_D(n^*)} - C_D(i) \right]}{[(N-1)(N-2)]}$$

Degree centralization examples



$$C_D = \frac{(5-1)+(5-1)+(5-1)+(5-1)+(5-1)+(5-5)}{5 \cdot 4} = 1.0$$

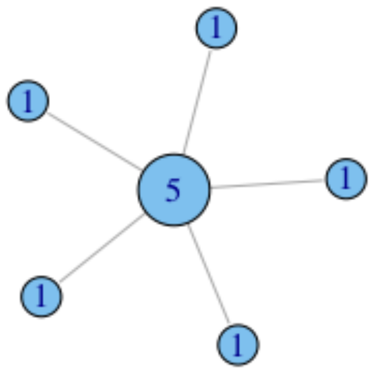
$C_D = 0.167$

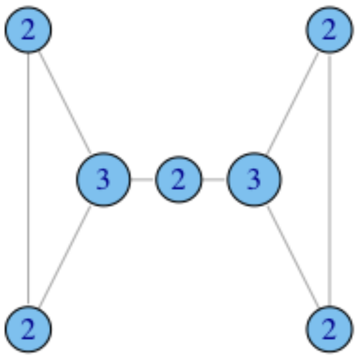


$C_D = 0.167$

What does degree not capture?

In what ways does degree fail to capture centrality in the following graphs?

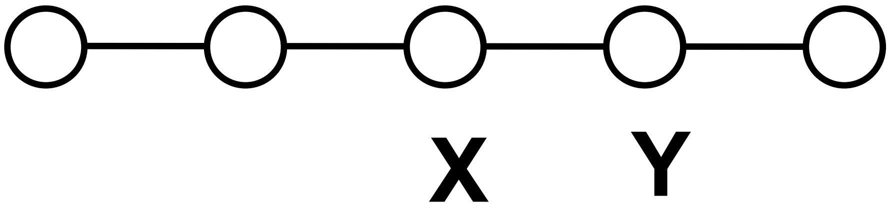




Brokerage not captured by degree !

Betweenness Centrality

How many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?



Betweenness Centrality

$$C_B(i) = \sum_{j<k} g_{jk}(i) / g_{jk}$$

- g_{jk} = the number of shortest paths connecting *jk*
- $g_{jk}(i)$ = the number that actor *i* is on.

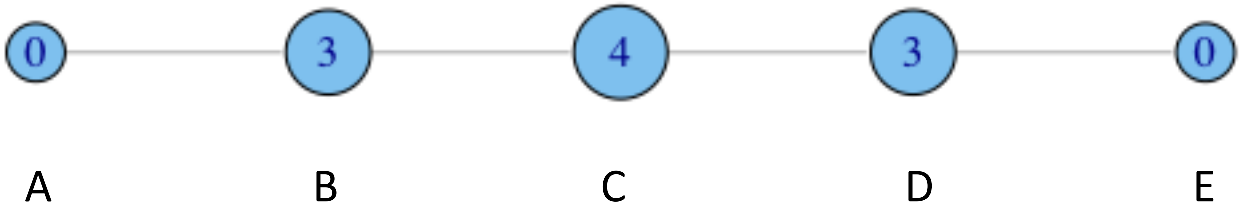
Usually normalized by:

$$C'_B(i) = C_B(i\ ) / \boxed{[(n-1)(n-2)/2]}$$

number of pairs of vertices excluding the vertex itself

Betweenness on toy networks

non-normalized version

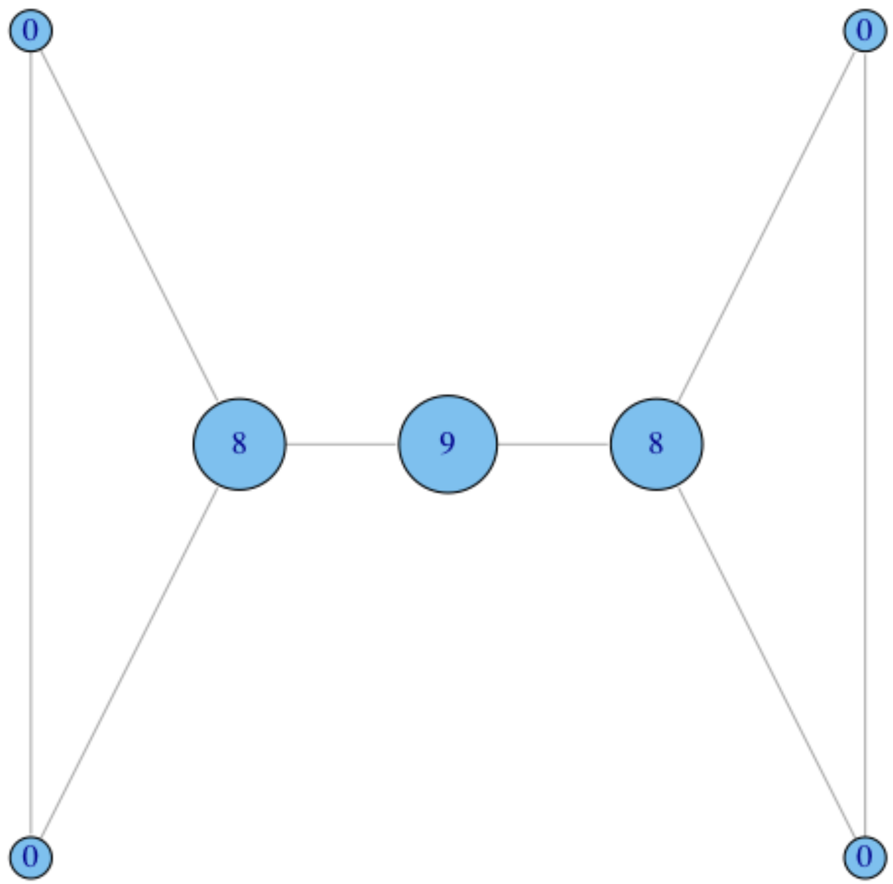


- A lies between no two other vertices
- B lies between A and 3 other vertices: C, D, and E
- C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)

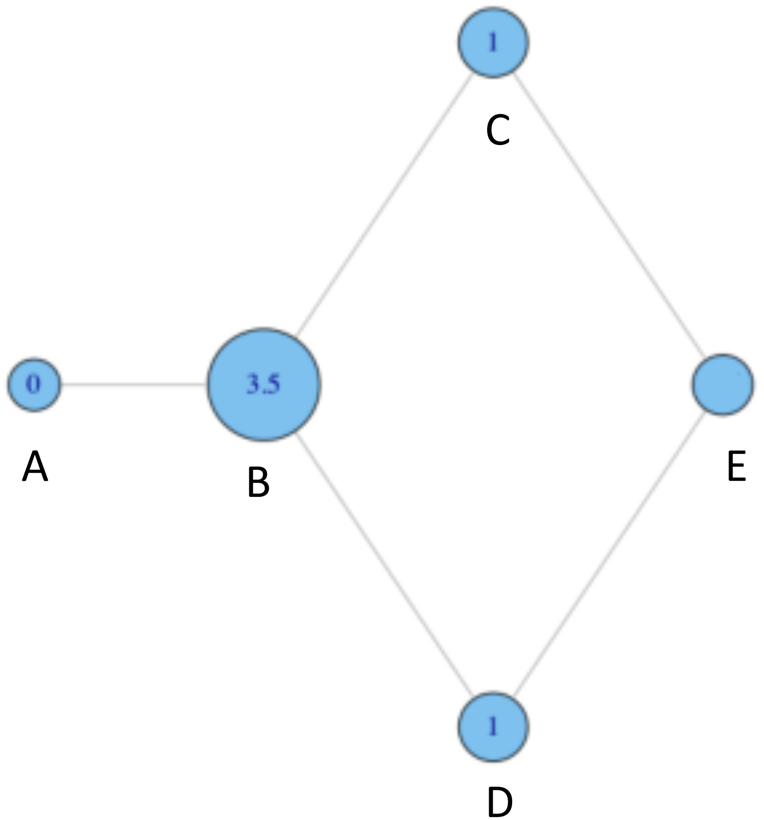
- note that there are no alternate paths for these pairs to take, so C gets full credit

Betweenness on toy networks

non-normalized version



non-normalized version



- why do C and D each have betweenness 1?
- They are both on shortest paths for pairs (A,E), and (B,E), and so must share credit:
  - ½+½ = 1

Betweenness centrality in directed networks

- We now consider the fraction of all directed paths between any two vertices that pass through a node

Betweenness of vertex i

paths between j and k that pass through i

$$C_B(i) = \sum_{j,k} g_{jk}(i) / g_{jk}$$

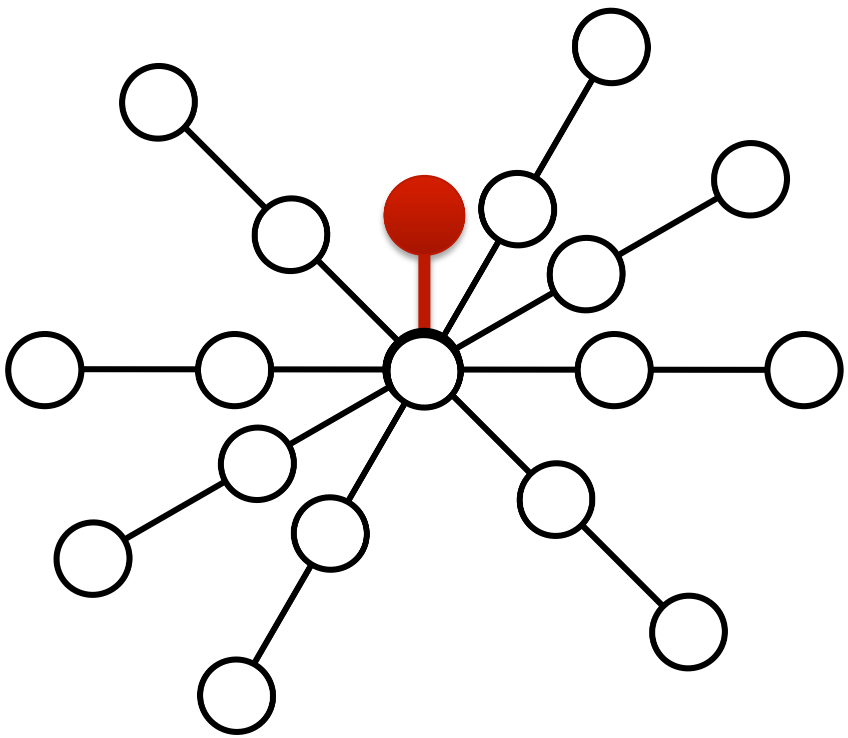
all paths between j and k

Only modification: when normalizing, we have (N-1)*(N-2) instead of (N-1)*(N-2)/2, because we have twice as many ordered pairs as unordered pairs

$$C'_B(i) = C_B(i) / [(N-1)(N-2)]$$

Closeness Centrality

- What if it's not so important to have many direct friends?
- Or be "between" others
- But one still wants to be in the "middle" of things, not too far from the center



Closeness Centrality

Closeness is based on the length of the average shortest path between a node and all other nodes in the network

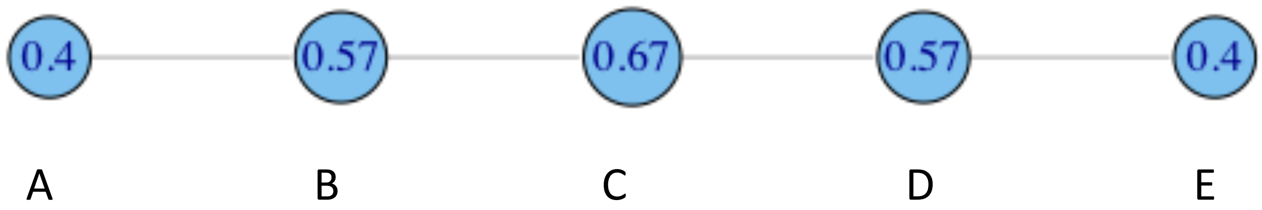
Closeness Centrality:

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

Normalized Closeness Centrality

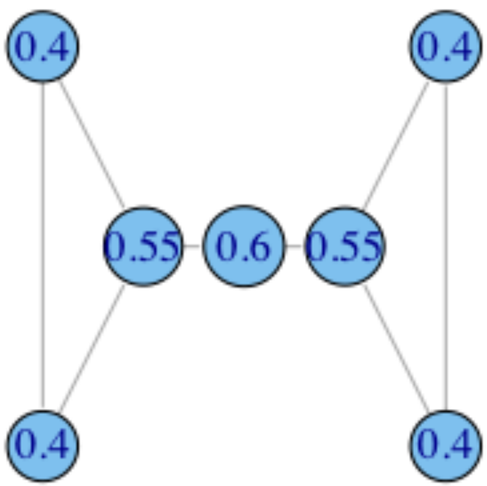
$$C_C^{'}(i) = (C_C(i))/(N-1)$$

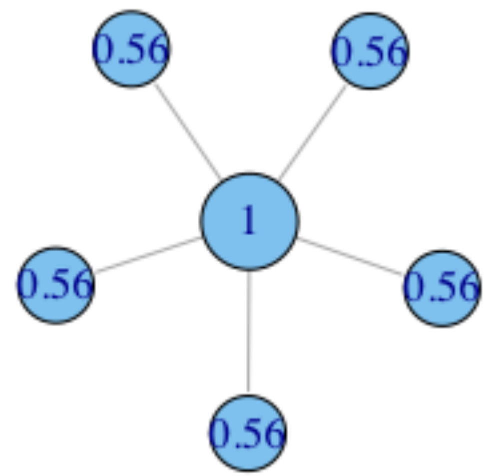
Closeness Centrality Toy Example



$$C_c'(A) = \left[\frac{\sum_{j=1}^{N} d(A,j)}{N-1} \right]^{-1} = \left[\frac{1+2+3+4}{4} \right]^{-1} = \left[\frac{10}{4} \right]^{-1} = 0.4$$

Closeness Centrality Other Examples

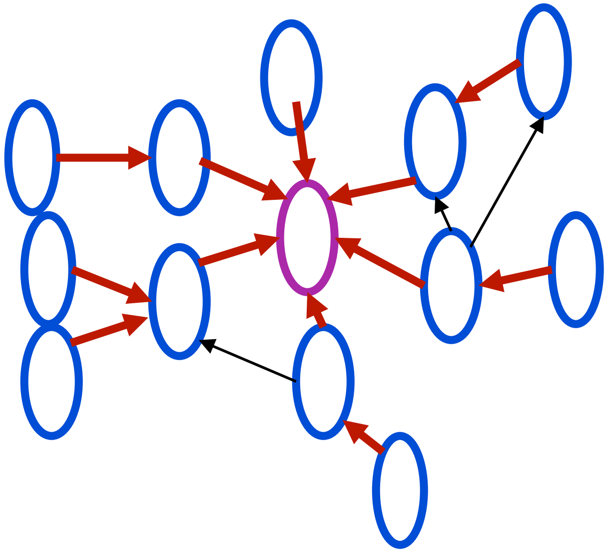




Directed closeness centrality

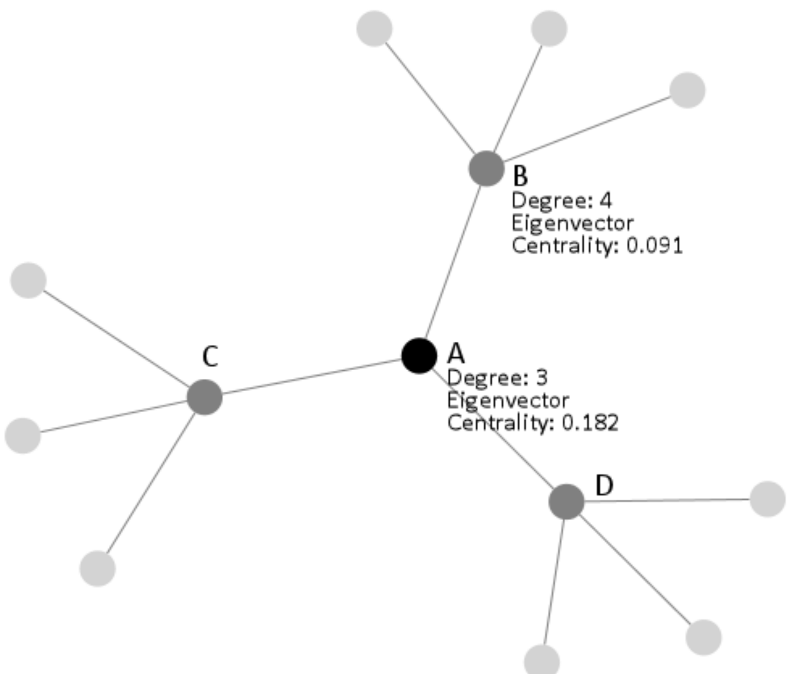
- **Choose a direction**
  - **in-closeness** (e.g. prestige in citation networks)
  - **out-closeness**

- Usually consider only vertices from which the node in question can be reached



Eigenvector centrality

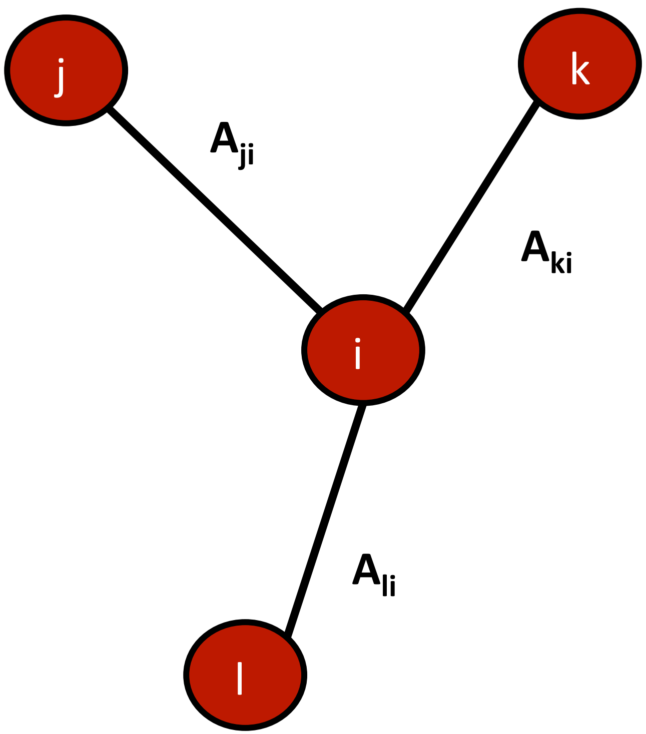
How central you are depends on how central your neighbors are



While the **degree** for node A in a social network measures how many ties A has, the **eigenvector centrality** of node A is measured based on ***how many ties A's connections have.***

Eigenvector centrality

How central you are depends on how central your neighbors are



The centrality score ***c(i)*** of each node i is proportional to its neighbors' scores

$$C(i) = A_{ji}C(j) + A_{ki}C(k) + A_{li}C(l)$$

Bonacich eigenvector centrality

- The **Bonacich Centrality** measure is also based on the premise that a node's importance is determined by *how important its neighbors* are.

$$c_i(\beta) = \sum_j (\alpha + \beta c_j) A_{ji}$$

- α is a normalization constant
- β determines how important the centrality of your neighbors is
- $\mathbf{A}$ is the adjacency matrix (can be weighted)

- This notion is central to citation rankings and things like Google page rankings.

Bonacich Power Centrality: attenuation factor b

- **small β ➔ high attenuation**
  - *only your immediate friends matter, and their importance is factored in only a bit*

- **high β ➔ low attenuation**
  - *global network structure matters (your friends, your friends' of friends etc.)*

- **β = 0** yields simple degree centrality

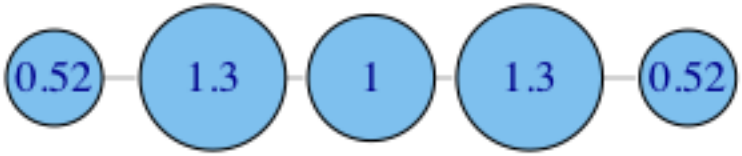
- **If β > 0,** nodes have higher centrality when they have edges to other central nodes.

- **If β < 0,** nodes have higher centrality when they have edges to less central nodes.

Bonacich Power Centrality: examples

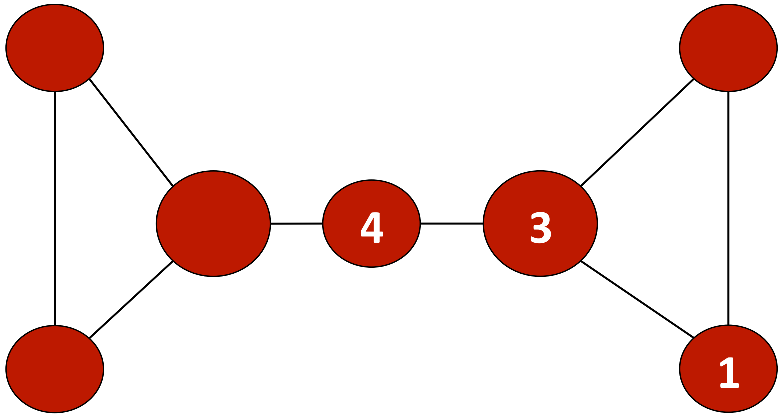
β = .25

β = -.25



Why does the middle node have lower centrality than its neighbors when β is negative?

Example Centrality Measures



	Node 1	Node 3	Node 4
Degree	.33	.50	.33
Betweeness	.00	.53	.60
Closeness	.40	.55	.60
Eigenvector	.47	.63	.54
Bonacich β=1/3, a=1	9.4	13	11
Bonacich β=1/4, a=1	4.9	6.8	5.4

Clustering Coefficient

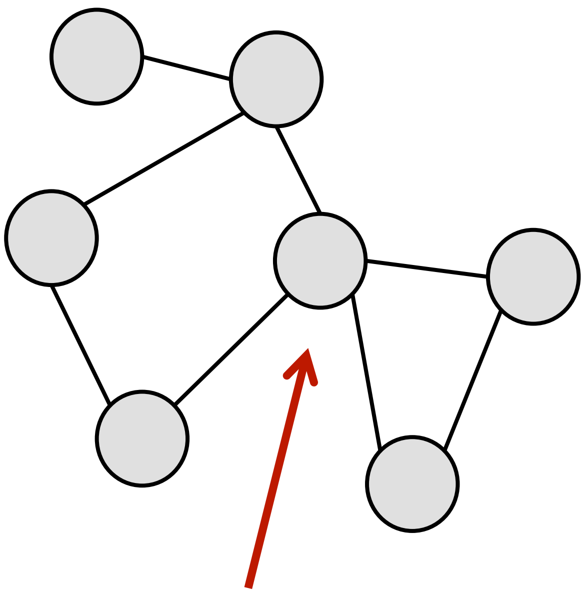
(Local) clustering coefficient for a node is the probability that two randomly selected friends of a node are friends with each other

$$C_i = \frac{2\,|\{e_{jk}\}|}{k_i(k_i - 1)}$$

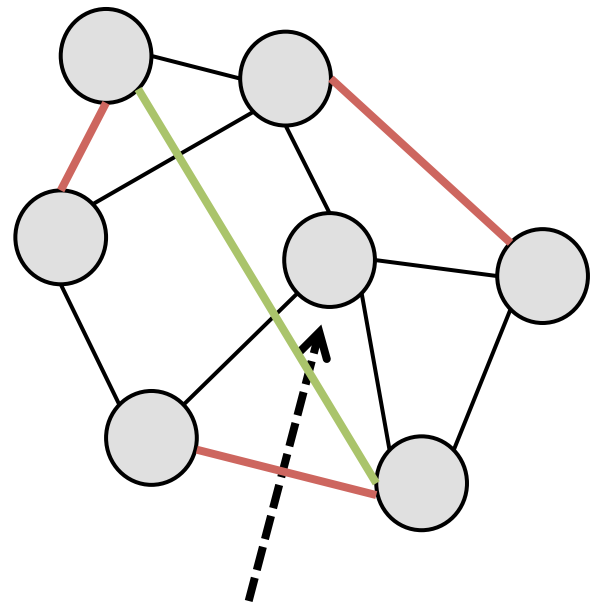
$e_{jk} \in E, u_i, u_j \in N_i$, k size of N_i, N_i neigborhood of u_i

Fraction of the friends of a node that are friends with each other (i.e., connected)

Clustering Coefficient



1/6



1/2

* Ranges from 0 to 1