Lecture 3 Centrality Measures and Metrics

Measures and Metrics

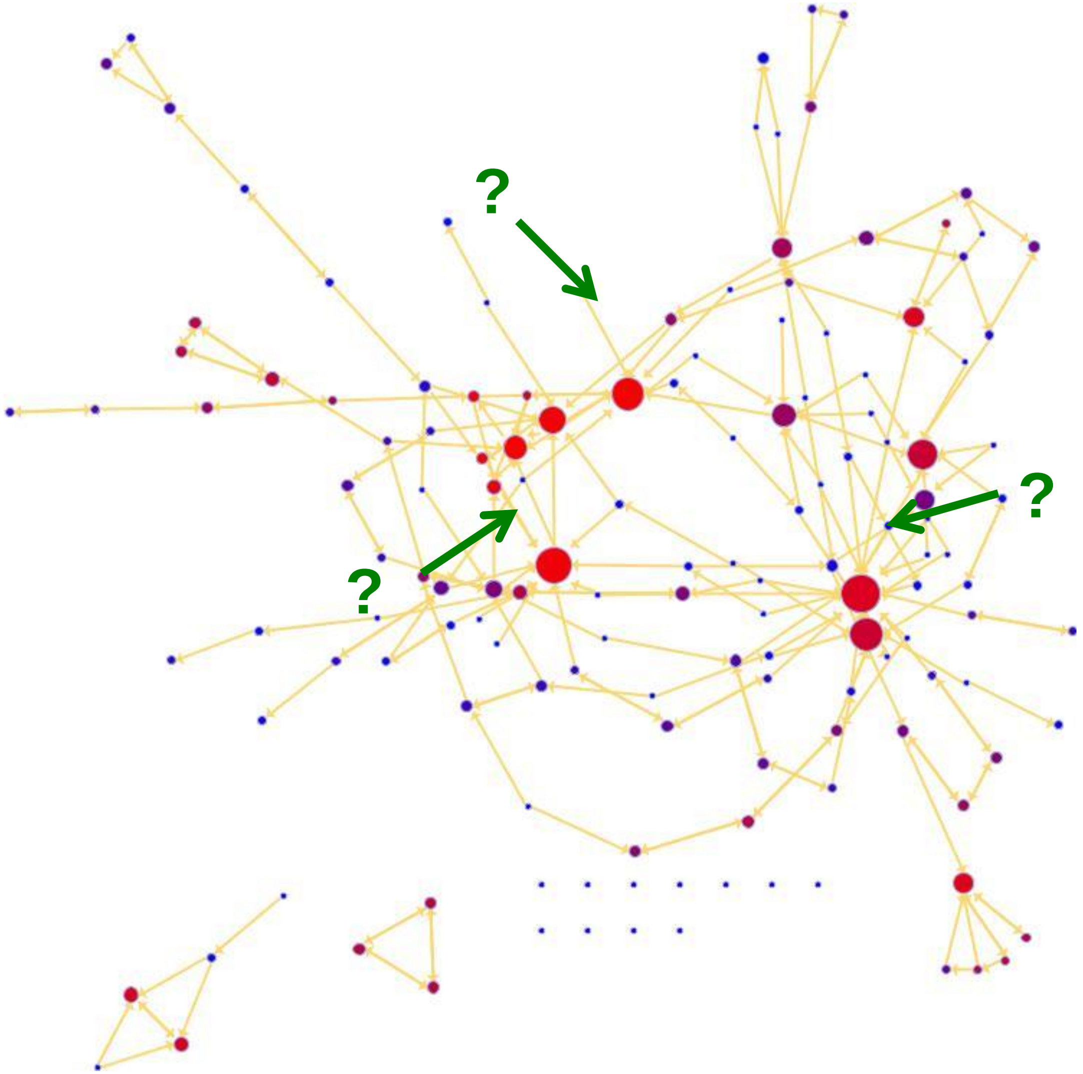
- Knowing the structure of a network, we can calculate various useful quantities or measures that capture particular features of the network topology.
  - basis of most of such measures are from *social network analysis*

- We have learned
  - Path length, Diameter, Degree, Density, Connectedness, etc

- We now learn Centrality
  - Degree, Eigenvector, Katz, PageRank, Hubs, Closeness, Betweenness, ….

- We will learn
  - Assortativity, Clustering coefficient, other graph metrics, etc

Characterizing networks:
Who is most central?



General Understanding

What vertices are ***most*** important?

- Important = prominent
- Important = admired
- Important = linchpin
- Important = listened to
- Important = in the know
- Important = gate keeper
- Important = involved

Translate into Network Language

Ordinary Description	Possible Network Interpretation
prominent	Vertex is “visible” to many other vertices
admired	Vertex is “chosen” by many other vertices
listened to	Vertex is “received” by many other vertices
in the know	Vertex is short distance from many sources of information
linchpin	Vertex irreplaceable part
gate keeper	Vertex stands between one part of graph and another
involved	Vertex connected to many parts of graph

Network Centrality

- Network Centrality measures address the question:
  "Who is the most important or central person in this network?"

- There are many answers to this question, depending on what we mean by importance.

- According to Scott Adams, the power a person holds in the organization is inversely proportional to the number of keys on his keyring.
  ➡ A janitor has keys to every office, and no power.
  ➡ The CEO does not need a key: people always open the door for him.

- There are a vast number of different centrality measures that have been proposed over the years.

- According to Freeman (1979):
  "There is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is little agreement on the proper procedure for its measurement."

Network Centrality

- Which nodes are most ‘central’?

- Definition of ‘central’ varies by context/purpose

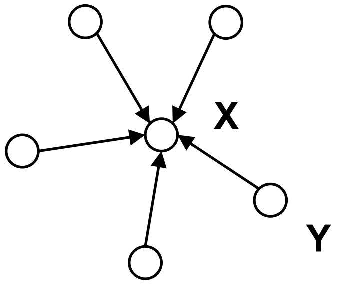
- Local measure:
  - degree

- Relative to rest of network:
  - closeness, betweenness, eigenvector, Katz, PageRank, …

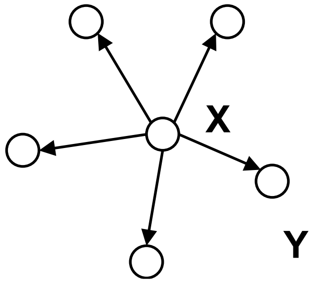
- How evenly is centrality distributed among nodes?
  - Centralization, hubs and authorities, …

Centrality: Who’s important based on their network position

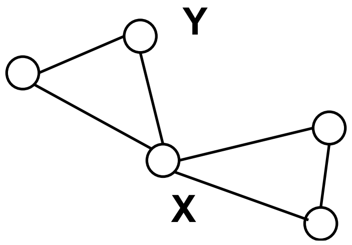
In each of the following networks, X has higher centrality than Y according to a particular measure



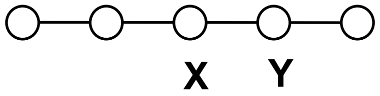
indegree



outdegree



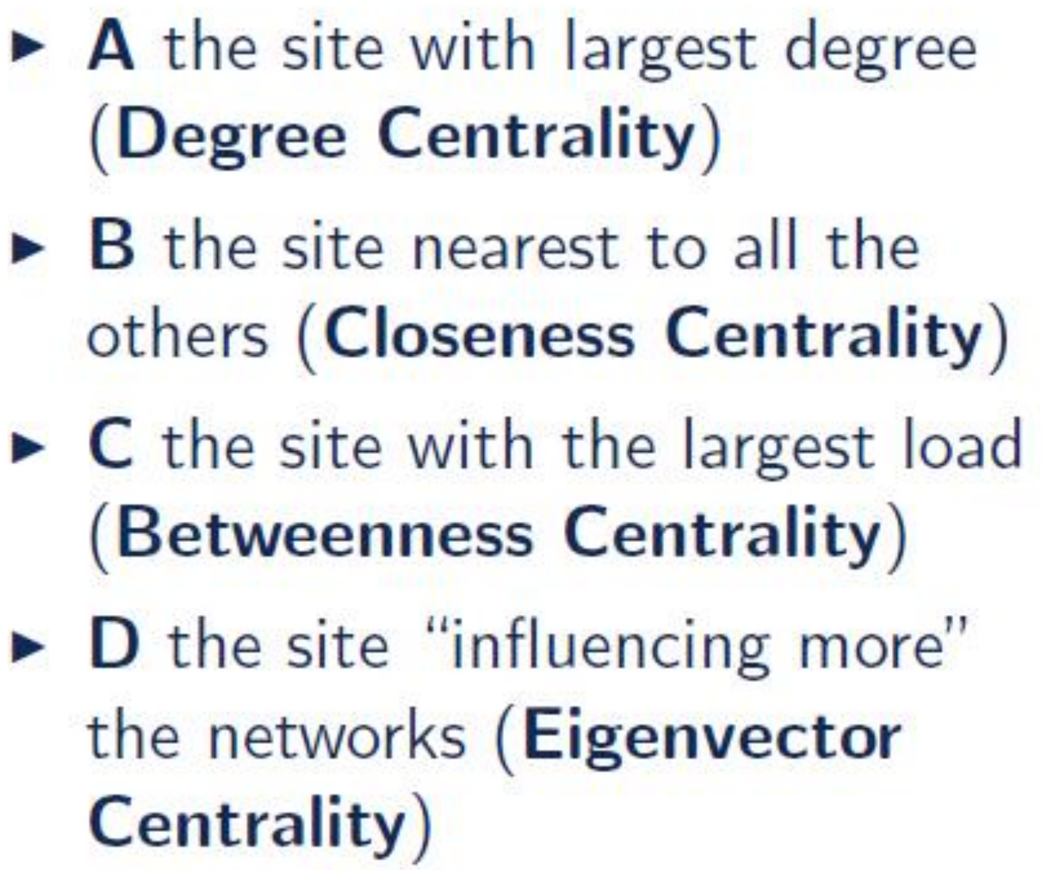
betweenness



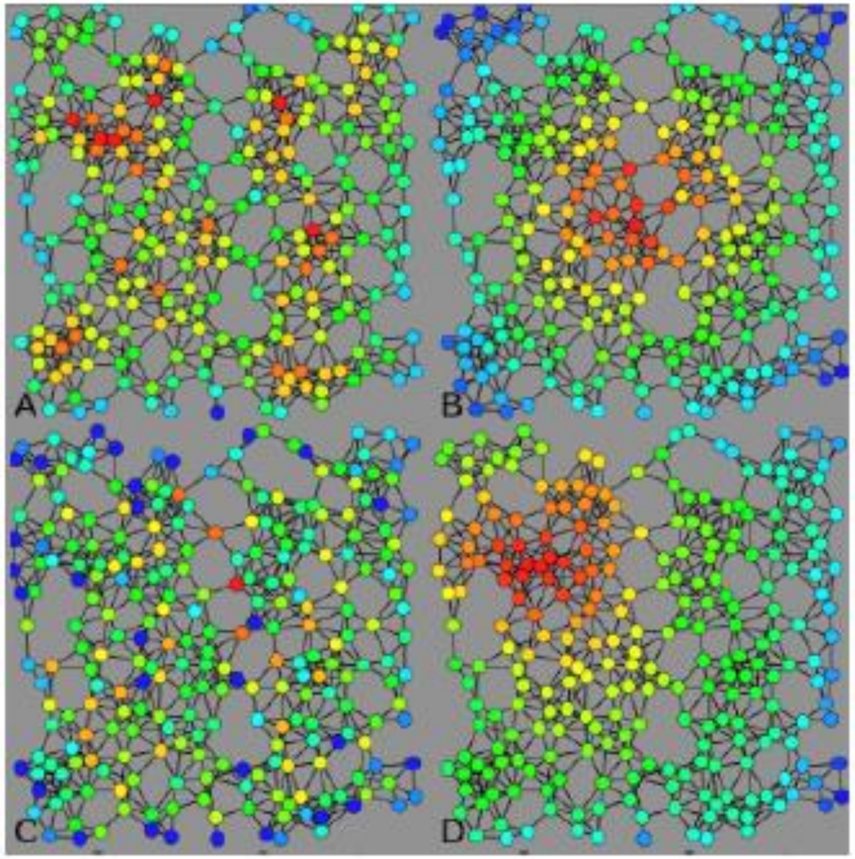
closeness

Common Centrality Measures

“Central” nodes connect the network. By removing these nodes, one can split the network in different parts.

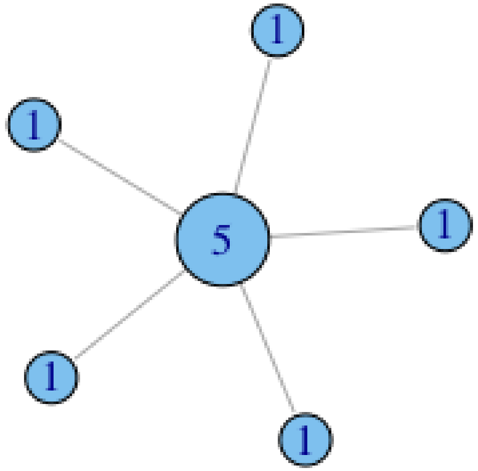


- **A** the site with largest degree (**Degree Centrality**)
- **B** the site nearest to all the others (**Closeness Centrality**)
- **C** the site with the largest load (**Betweenness Centrality**)
- **D** the site “influencing more” the networks (**Eigenvector Centrality**)



Degree Centrality (undirected)

He who has many friends is most important.

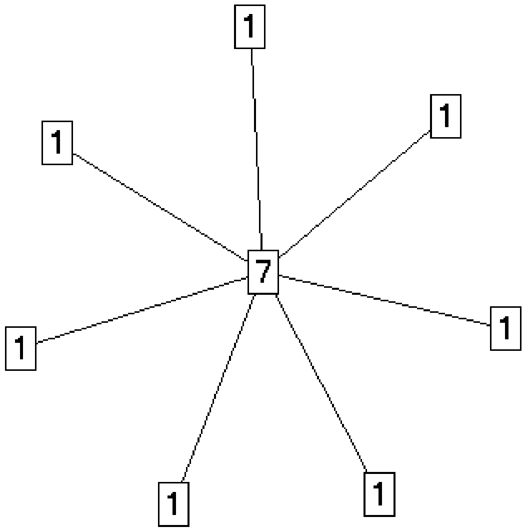


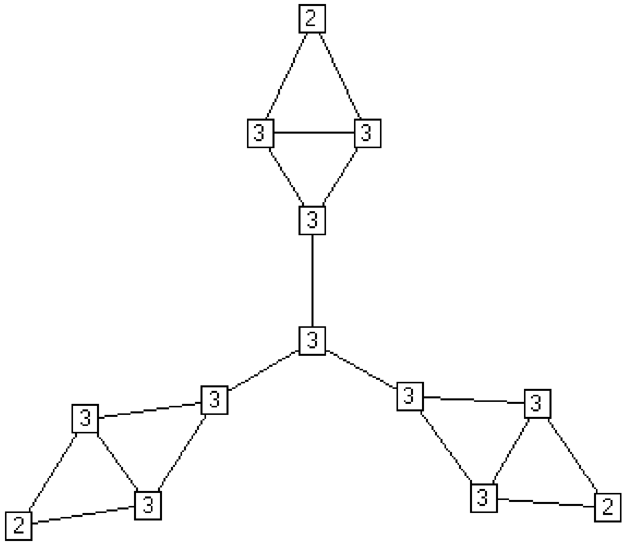
When is the number of connections the best centrality measure?

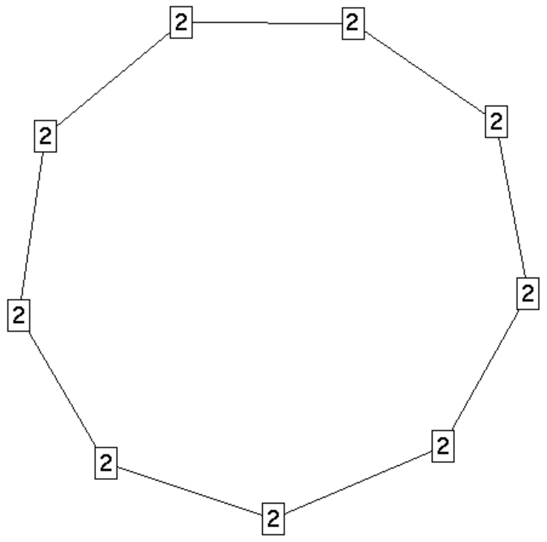
- people who will do favors for you
- people you can talk to (influence set, information access, …)
- influence of an article in terms of citations (using in-degree)

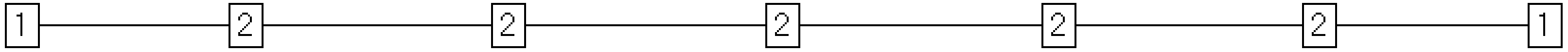
Degree Centrality in Social Networks

The most intuitive notion of centrality focuses on **degree**:
The actor with the most ties is the most important





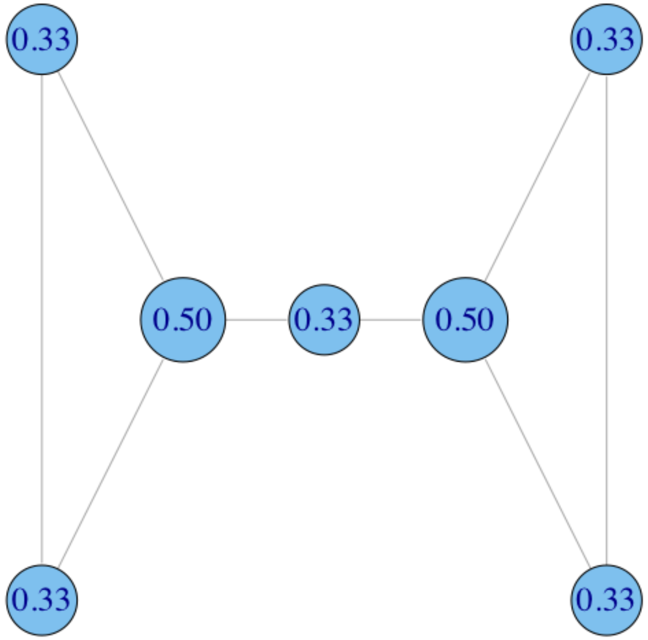


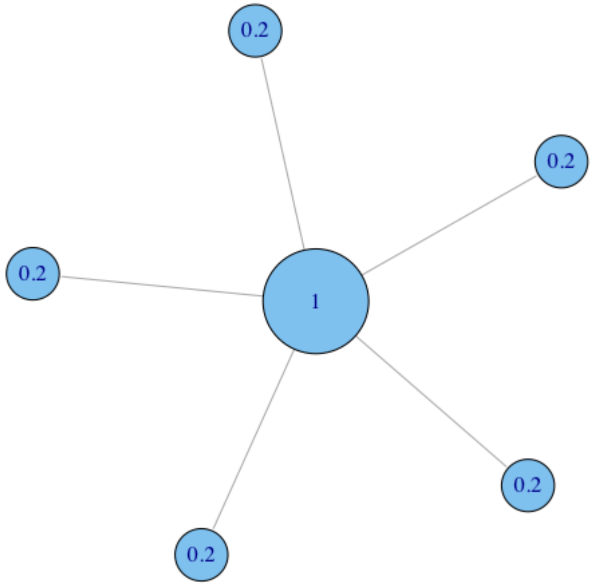


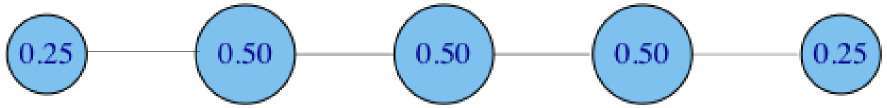
$$C_D(i) = \deg(i) = k_i = \sum_{j=1}^{N} A_{ij}$$

Degree: Normalized Degree Centrality

Divide by the maximum possible, i.e. (*N*-1)

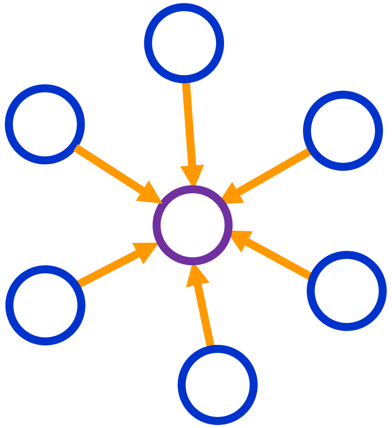




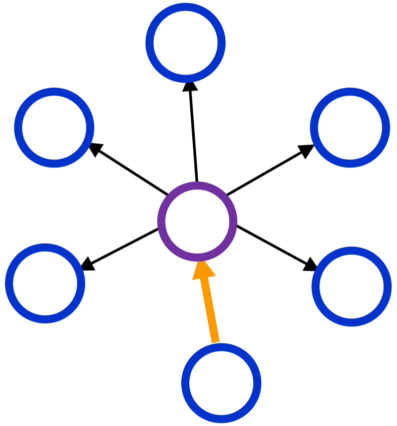


Extensions of undirected degree centrality

- Degree Centrality
  - Indegree centrality
    - A paper that is cited by many others
    - A person nominated by many others for a reward
    - Nodes with higher indegree is more ***prestigious*** (choices received)
  - Outdegree centrality
    - Nodes with higher outdegree is more ***central*** (choices made)



In-degree Centrality



Out-degree Centrality

Degree centralization: How equal are the nodes?

How much variation is there in the centrality scores among the nodes?

Freeman's general formula for centralization:

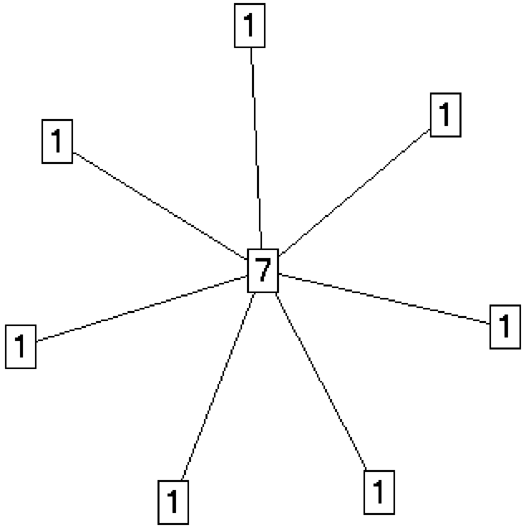
(One can use other metrics, e.g. Gini coefficient or standard deviation)

maximum value in the network

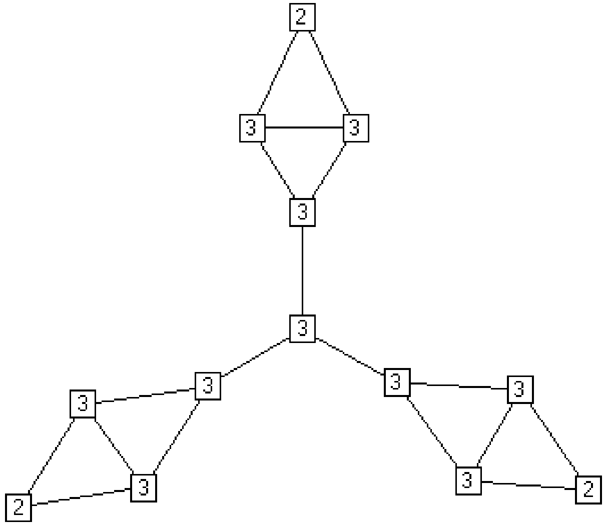
$$C_D = \frac{\sum_{i=1}^{N}[\boxed{C_D(n^*)} - C_D(i)]}{(N-1)(N-2)}$$

Degree Centrality in Social Networks

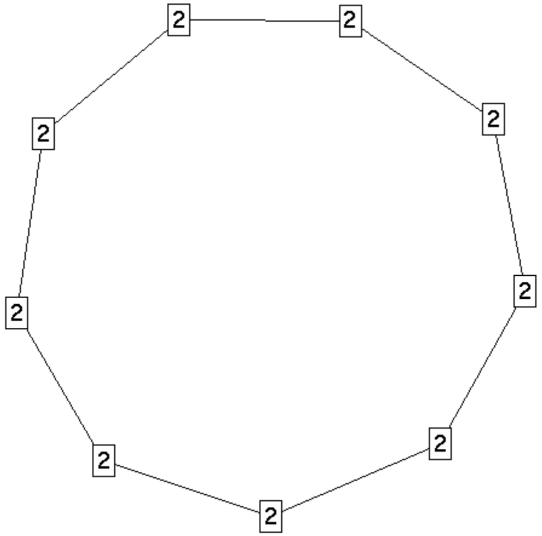
Degree <u>Centralization</u> Scores



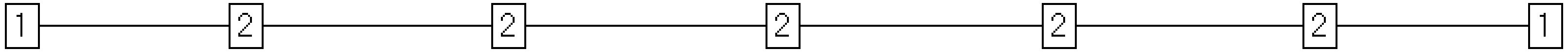
Freeman: 1.0
Variance: 3.9



Freeman: .02
Variance: .17

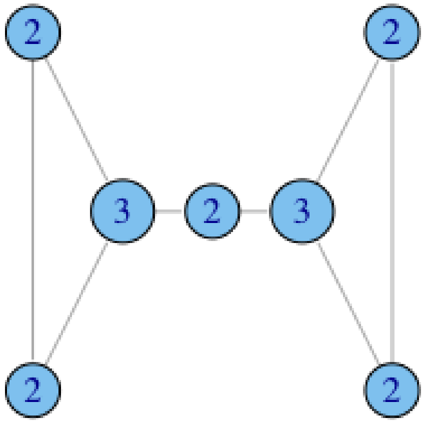


Freeman: 0.0
Variance: 0.0

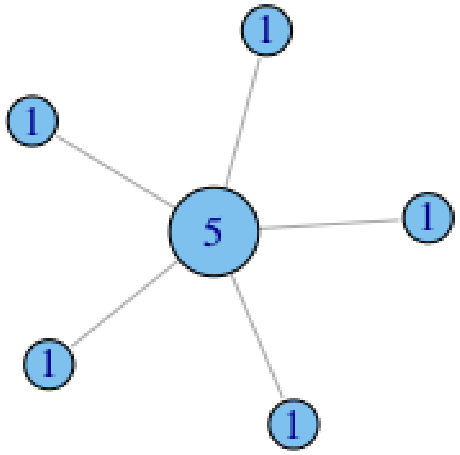


Freeman: .07
Variance: .20

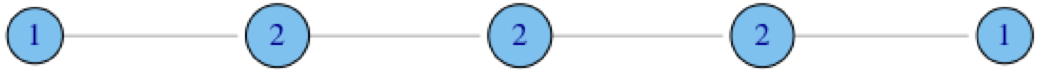
Degree Centralization examples



$C_D = 0.167$



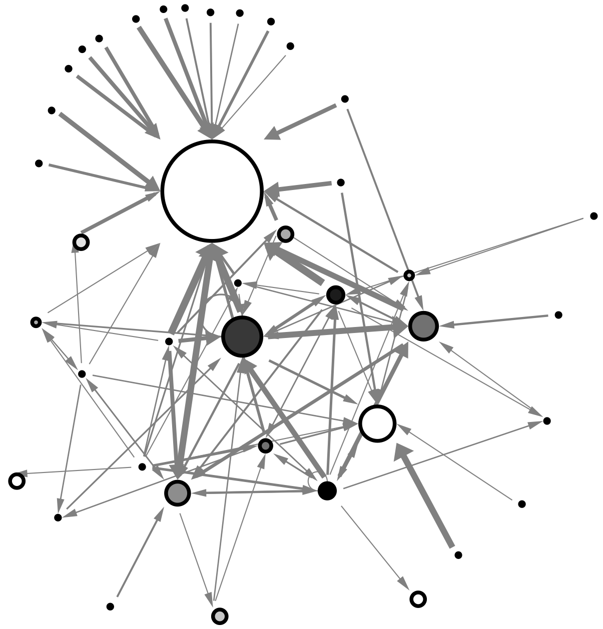
$C_D = 1.0$

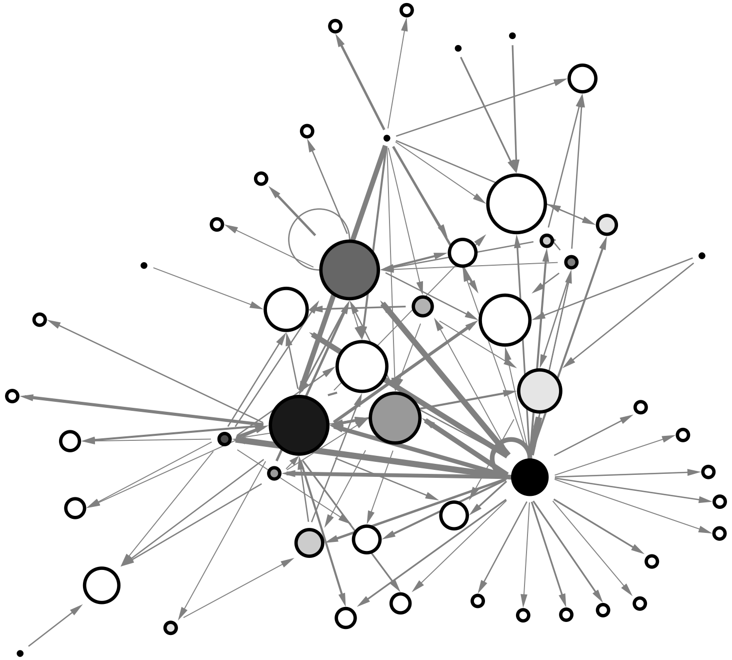


$C_D = 0.167$

Degree Centralization examples

Example: Financial trading networks

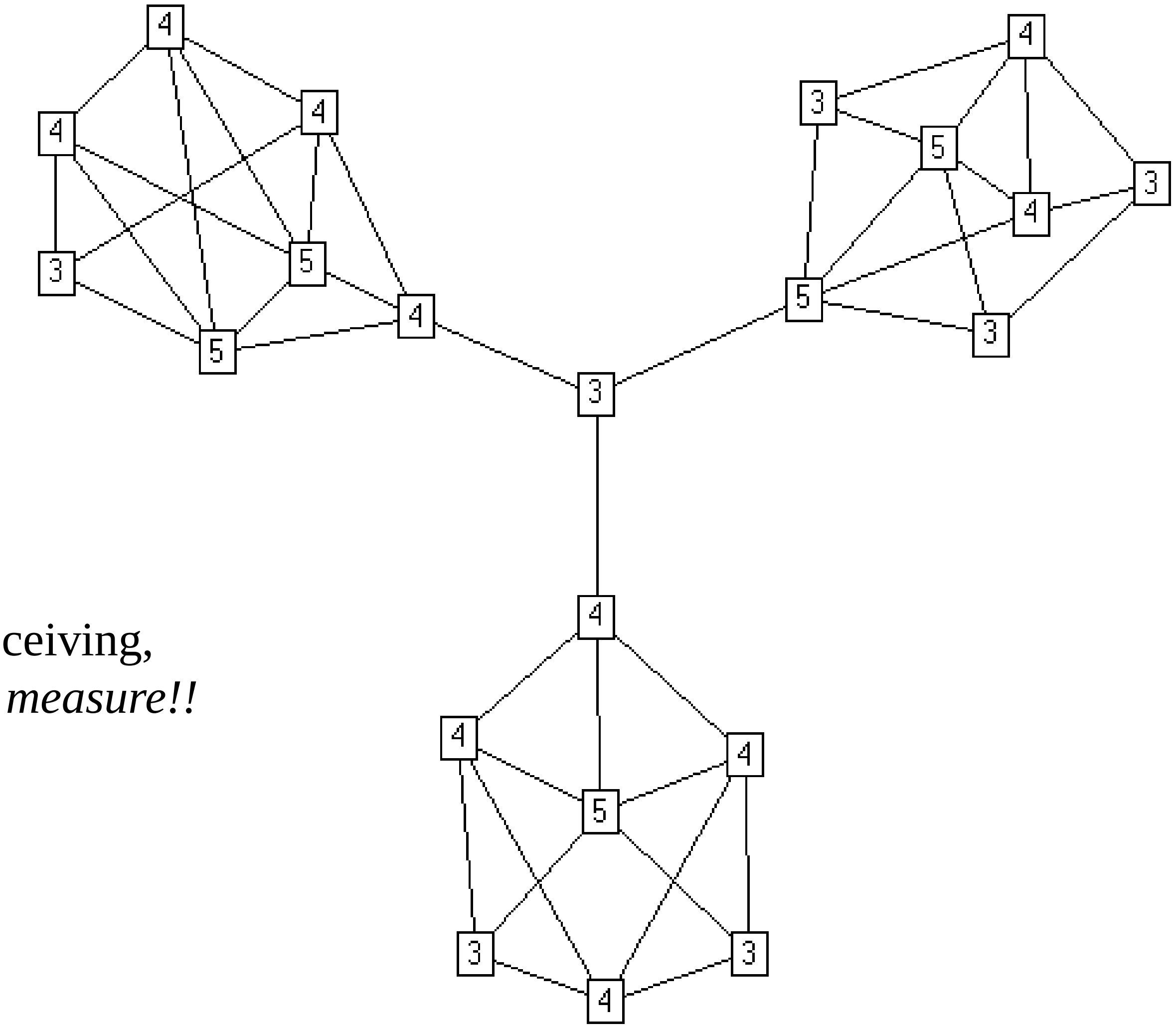




High centralization: one node trading with many others

Low centralization: trades are more evenly distributed

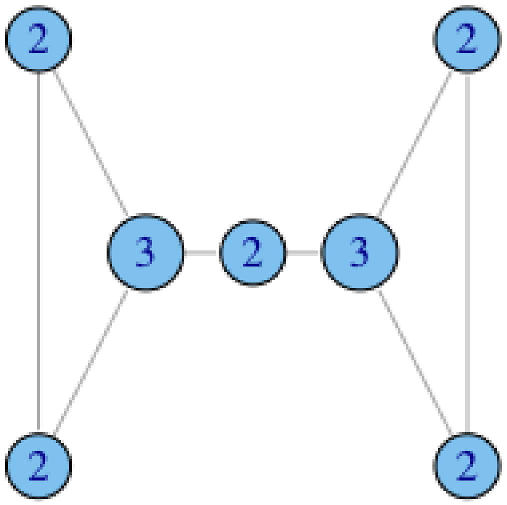
Degree Centrality in Social Networks

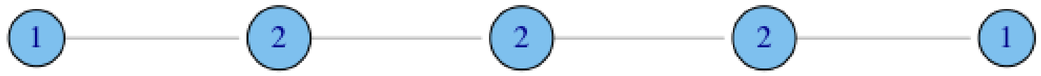


Degree centrality can be deceiving,
*because it is a **purely** local measure!!*

When degree isn't everything……

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





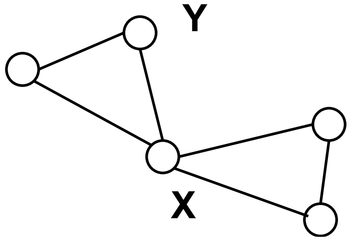
- ability to mediate between groups
- likelihood that information originating anywhere in the network reaches you…

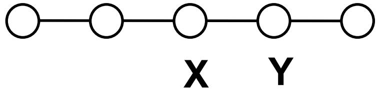
Betweenness Centrality

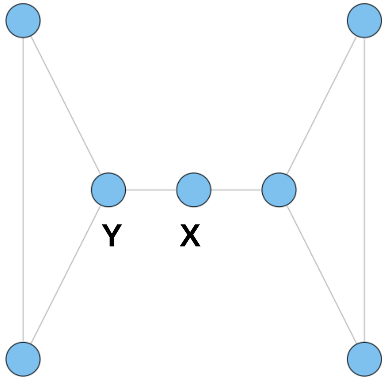
Betweenness: another centrality measure that can be obtained by considering the number of times that we cross one vertex *i* when going from vertex *j* to *k* following the shortest path.

- *Intuition:* how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?

- Who has higher betweenness, X or Y?



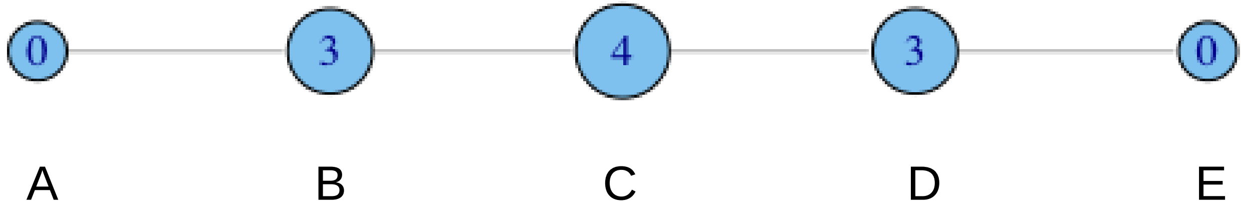




Betweenness Centrality

Toy networks

- Non-normalized version:



- A and E lie 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 Centrality

Definition:

betweenness of vertex i

paths between j and k that pass through i

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

all paths between j and k

Where g_{jk} = the number of geodesics connecting *j*-*k*, and
g_{jk} (*i*) = the number that actor *i* is on.

Usually normalized by:

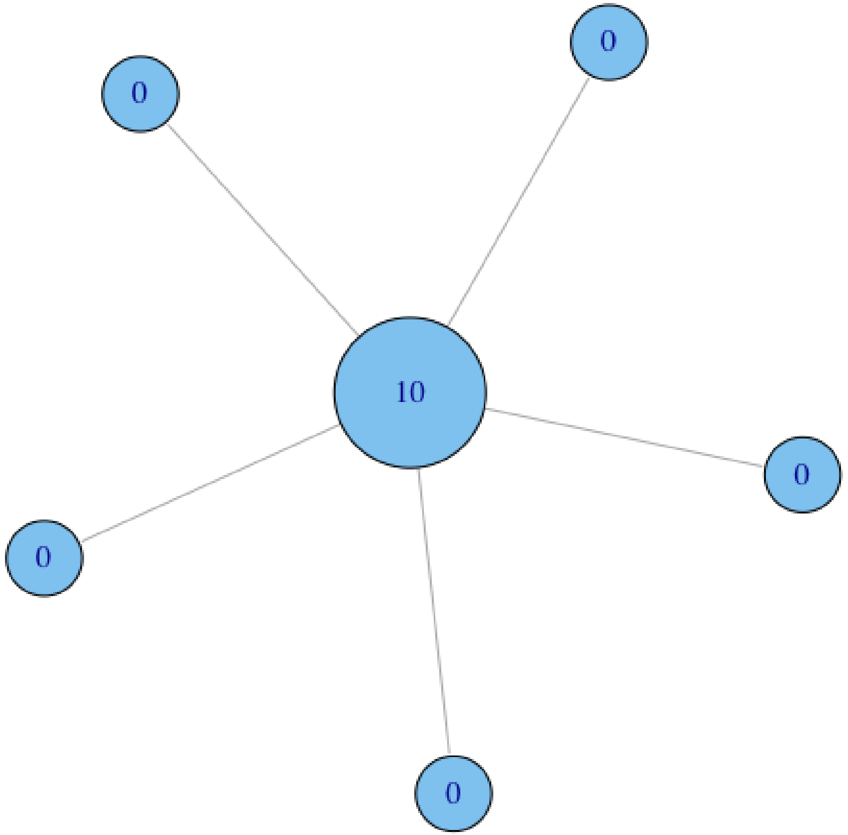
$$C'_B(i) = \frac{C_B(i)}{\boxed{(n-1)(n-2)/2}}$$

number of pairs of vertices
excluding the vertex itself

Betweenness Centrality

More toy networks

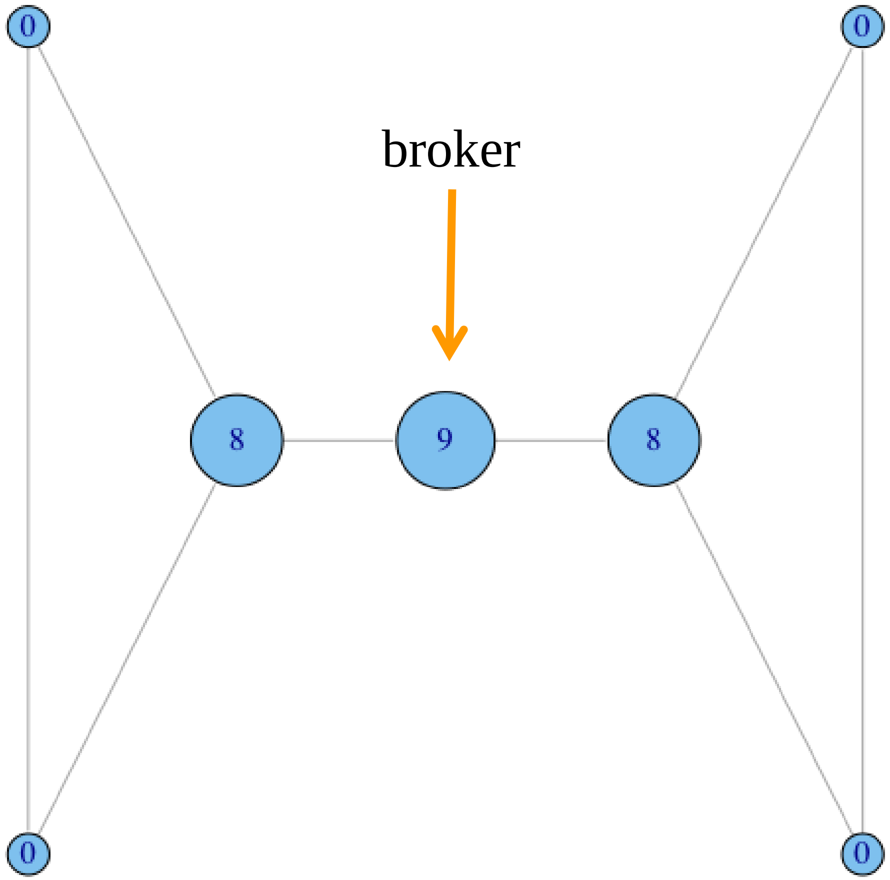
Non-normalized version:



Betweenness Centrality

More toy networks

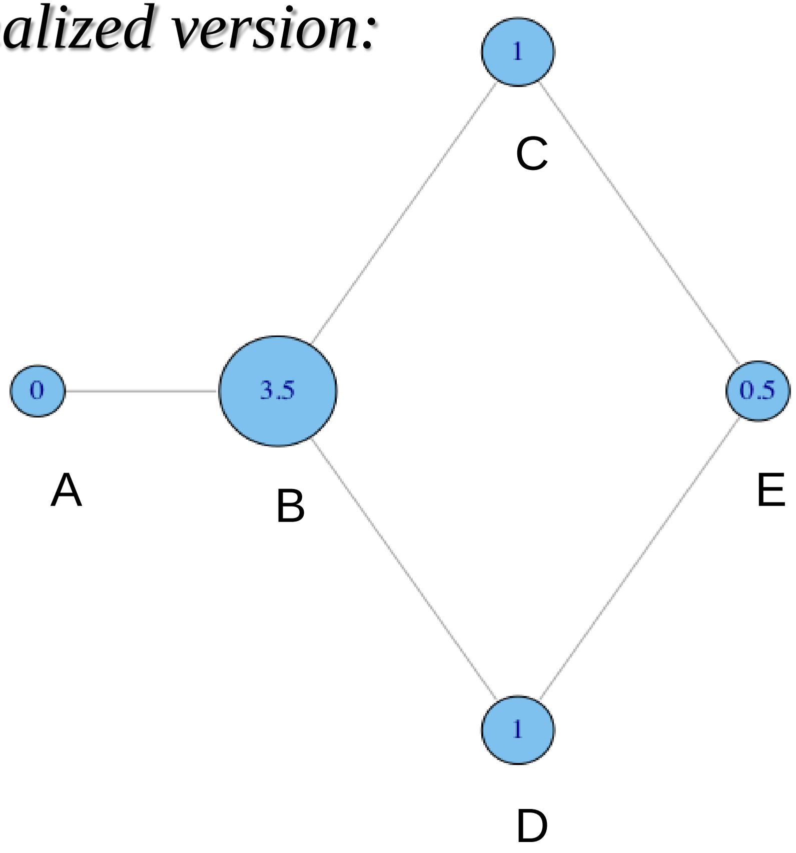
Non-normalized version:



Betweenness Centrality

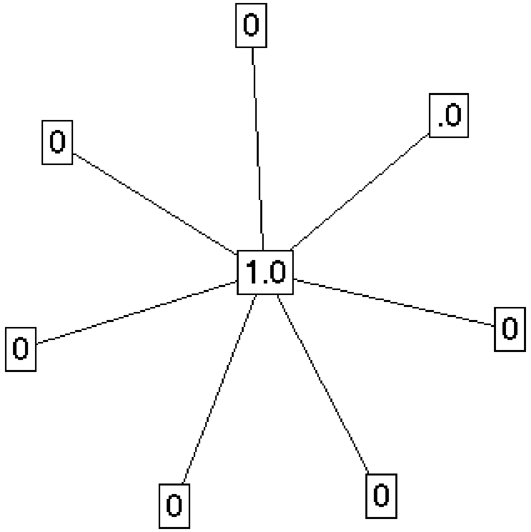
More toy networks

Non-normalized version:

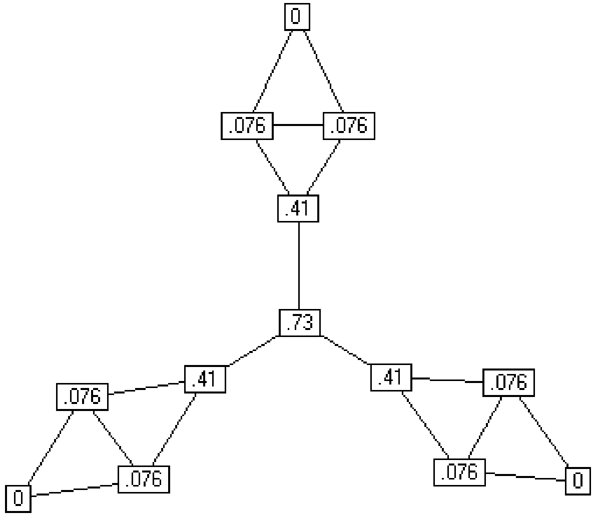


- Why do C and D each have betweenness 1?
- They are both on shortest paths for pairs (A,E), and (B,E), and so must share credit:
  - ½+½ = 1
- Can you figure out why B has betweenness 3.5 while E has betweenness 0.5?

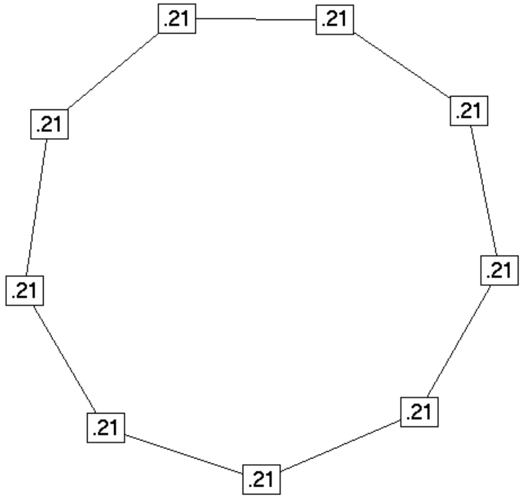
Betweenness Centrality in Social Networks



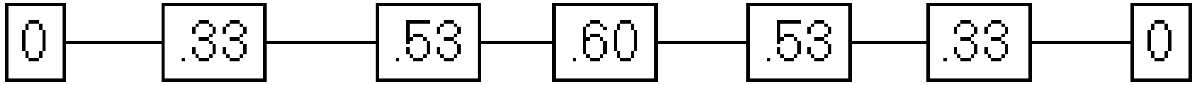
Centralization: 1.0



Centralization: .59

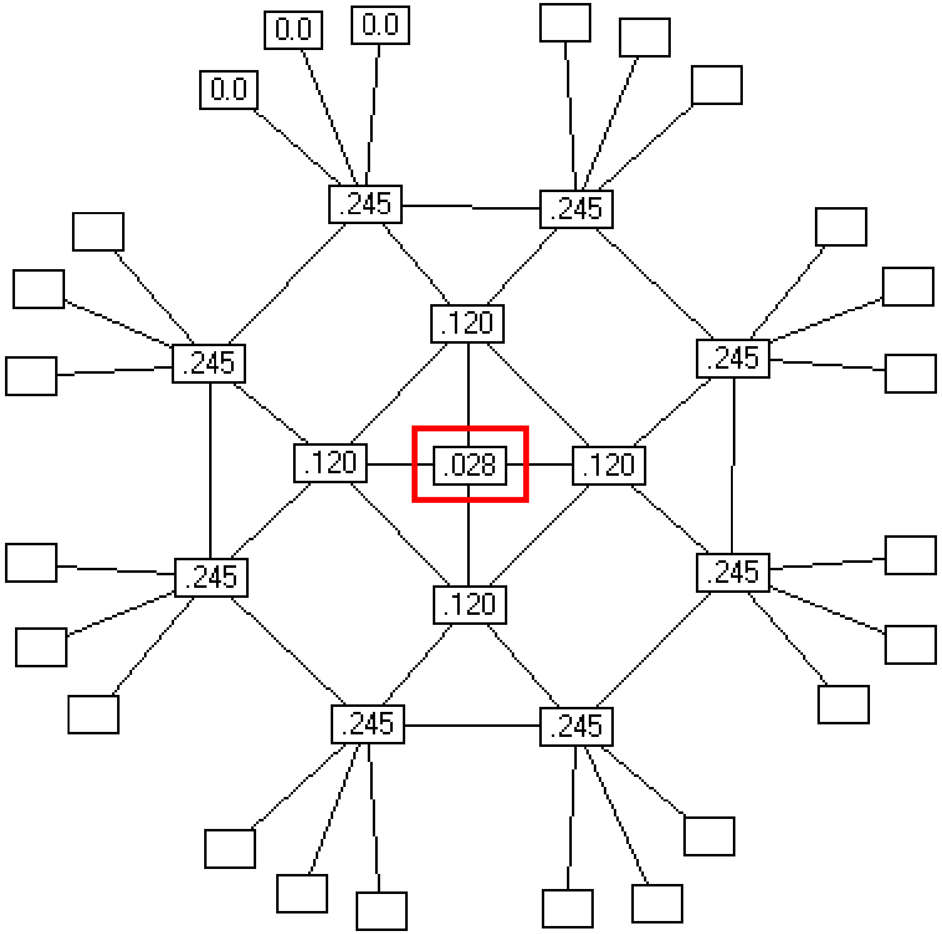


Centralization: 0



Centralization: .31

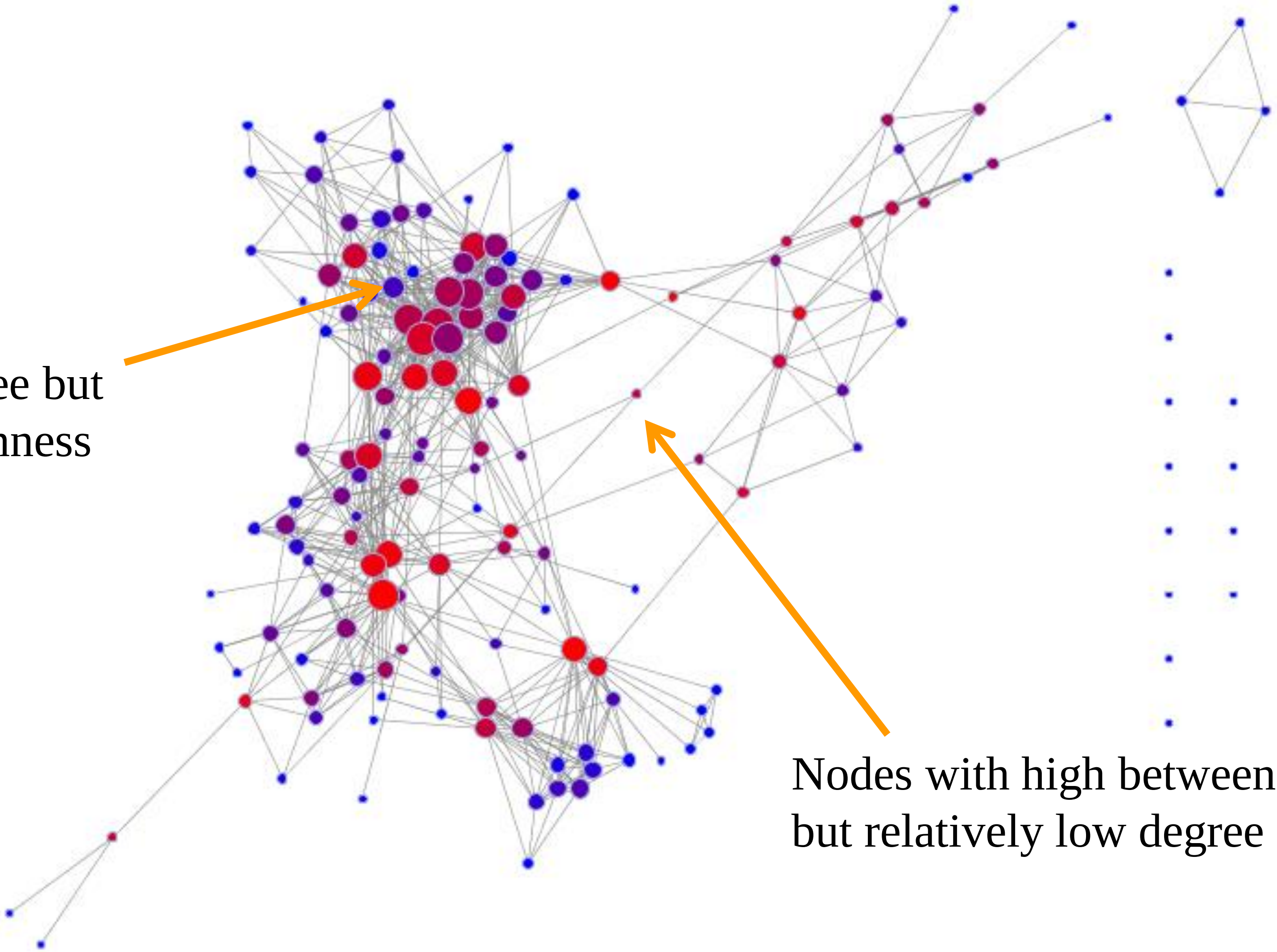
Betweenness Centrality in Social Networks



Centralization: .183

Betweenness Centrality

Example: Nodes are sized by degree, and colored by betweenness.



Nodes with high degree but
relatively low betweenness

Nodes with high betweenness
but relatively low degree

Betweenness Centrality

Extend to directed networks:

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*

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

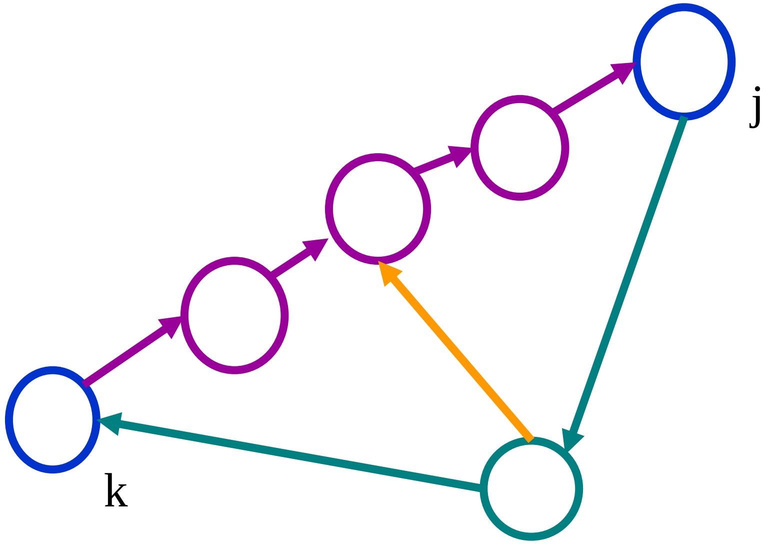
all paths between *j* and *k*

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

$$C'_B(i) = \frac{C_B(i)}{(n-1)(n-2)}$$

Directed geodesics

- A node does not necessarily lie on a geodesic from *j* to *k* if it lies on a geodesic from *k* to *j*.



Alternative betweenness computations

- Slight variations in geodesic path computations
  - inclusion of self in the computations

- Flow betweenness
  - Based on the idea of maximum flow
    - edge-independent path selection effects the results
    - may not include geodesic paths

- Random-walk betweenness
  - Based on the idea of random walks
    - yields ranking similar to geodesic betweenness

- Many other alternative definitions exist based on diffusion, transmission or flow along network edges

Closeness Centrality

Closeness: another centrality measure

- 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

- In connected graphs there is a natural distance metric between all pairs of nodes, defined by the length of their shortest paths. The ***farness*** of a node *i* is defined as the sum of its distances to all other nodes, and its ***closeness*** is defined as the inverse of the farness. Naturally, the ***closeness*** in an unconnected graph would be 0.

Closeness Centrality

***Definition*:** Closeness is based on the length of the average shortest path between a vertex *i* and all vertices *j* in the graph

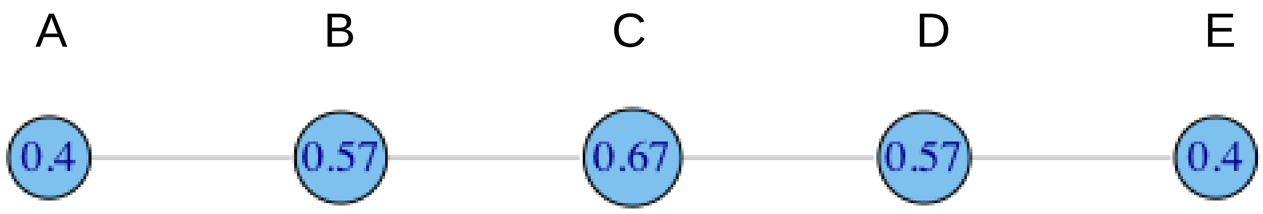
Closeness Centrality:

$$C_i = N\left[\sum_{j=1}^{N} d_{ij}\right]^{-1}$$

In other words, it depends on inverse distance to other vertices.
Note that some people use *N* - 1 instead of *N* in the definition.

Closeness Centrality

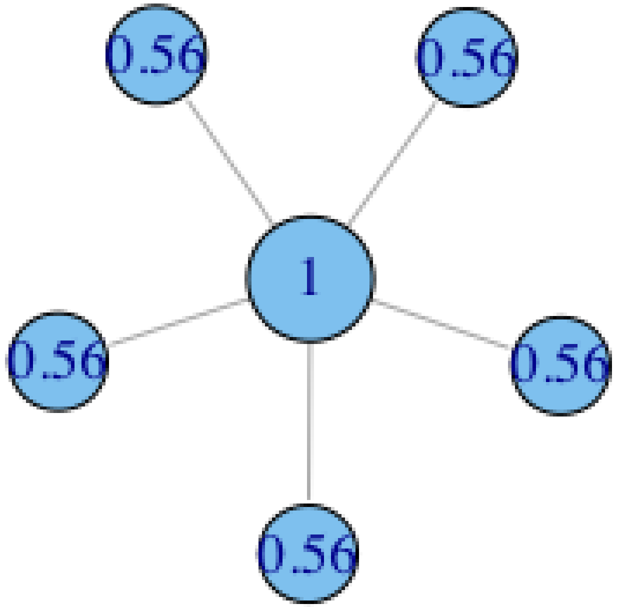
Toy Example:

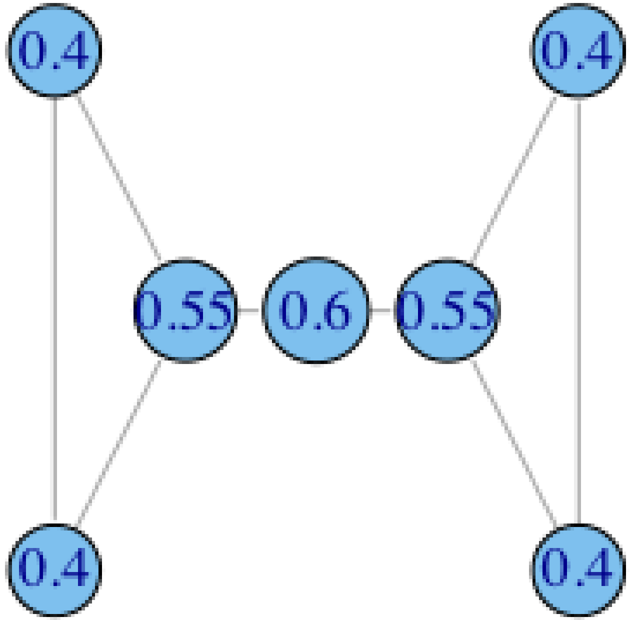


$$C_c'(A) = \left[\frac{\sum_{j=1}^{N} d(A,j)}{N-1}\right]^{-1} = \left[\frac{1+2+3+4}{4}\right]^{-1} = \left[\frac{10}{4}\right]^{-1} = 0.4$$

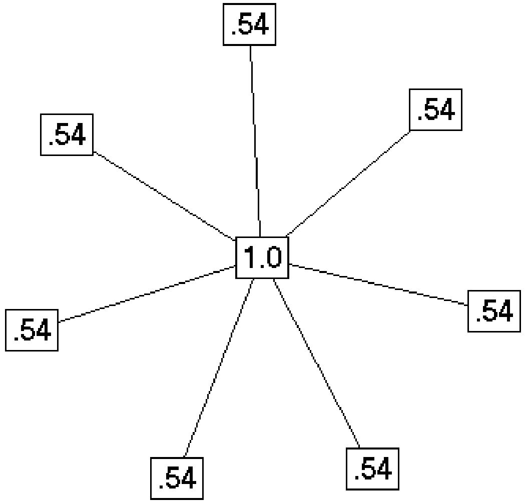
Closeness Centrality

More Toy Example:

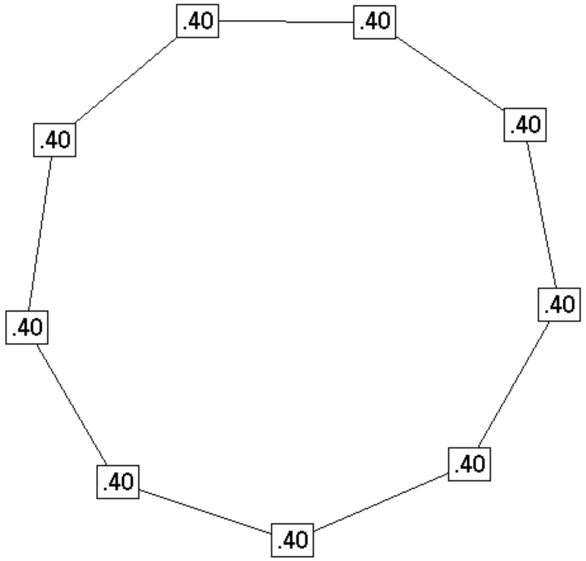




Closeness Centrality in Social Networks



Distance	Closeness	normalized
0 1 1 1 1 1 1 1	.143	1.00
1 0 2 2 2 2 2 2	.077	.538
1 2 0 2 2 2 2 2	.077	.538
1 2 2 0 2 2 2 2	.077	.538
1 2 2 2 0 2 2 2	.077	.538
1 2 2 2 2 0 2 2	.077	.538
1 2 2 2 2 2 0 2	.077	.538
1 2 2 2 2 2 2 0	.077	.538

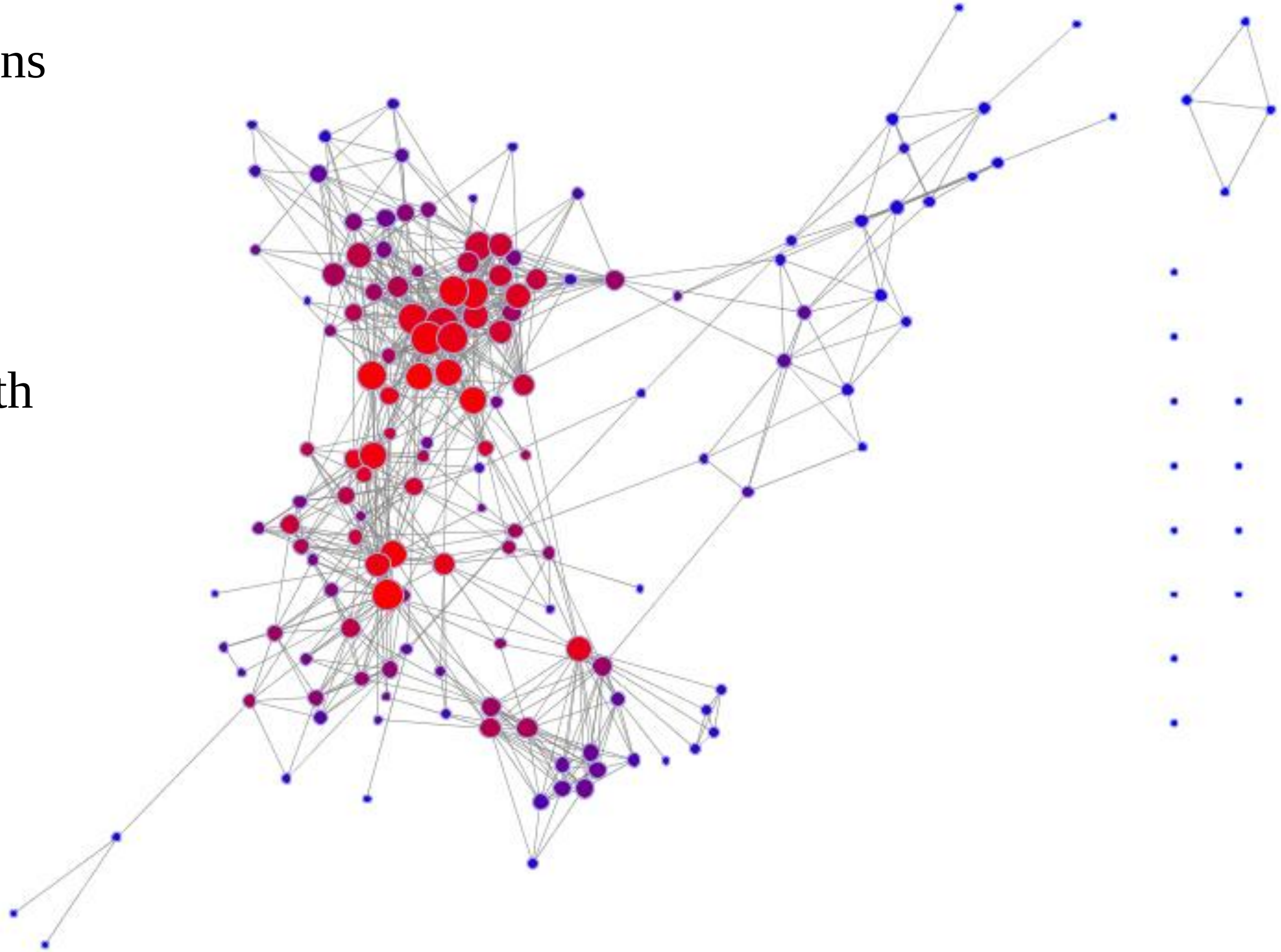


Distance	Closeness	normalized
0 1 2 3 4 4 3 2 1	.050	.400
1 0 1 2 3 4 4 3 2	.050	.400
2 1 0 1 2 3 4 4 3	.050	.400
3 2 1 0 1 2 3 4 4	.050	.400
4 3 2 1 0 1 2 3 4	.050	.400
4 4 3 2 1 0 1 2 3	.050	.400
3 4 4 3 2 1 0 1 2	.050	.400
2 3 4 4 3 2 1 0 1	.050	.400
1 2 3 4 4 3 2 1 0	.050	.400

How closely do degree and betweenness correspond to closeness?

- **Degree**
  - number of connections
  - denoted by size

- **Closeness**
  - length of shortest path to all others
  - denoted by color



Closeness Centrality

Some Problems with Closeness Centrality

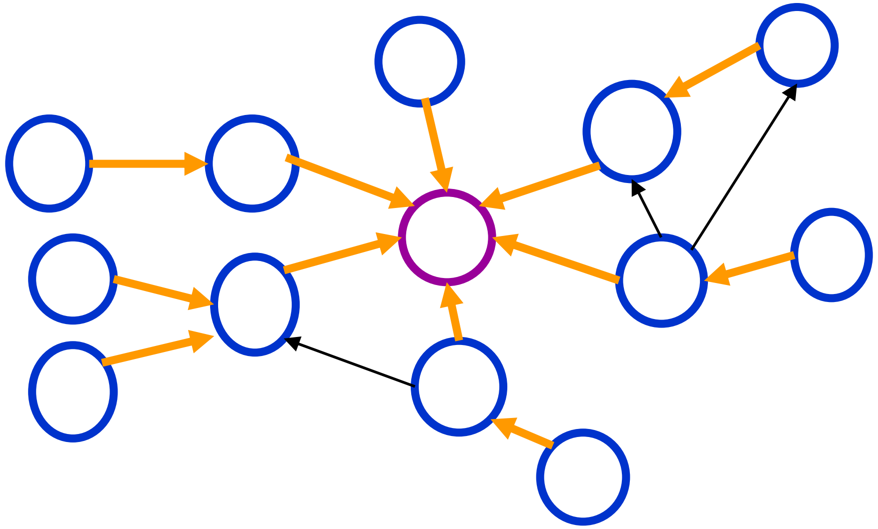
- Values tend to span a rather small dynamic range
  - typical distance increases logarithmically with network size
- In a typical network the closeness centrality C might span a factor of five or less
  - It is difficult to distinguish between central and less central vertices
  - A small change in network might considerably affect the centrality order
- Alternative computations exist but they have their own problems

Closeness Centrality

Extensions of undirected closeness centrality

- Closeness Centrality usually implies
  - all paths should lead to you
  - paths should lead from you to everywhere else

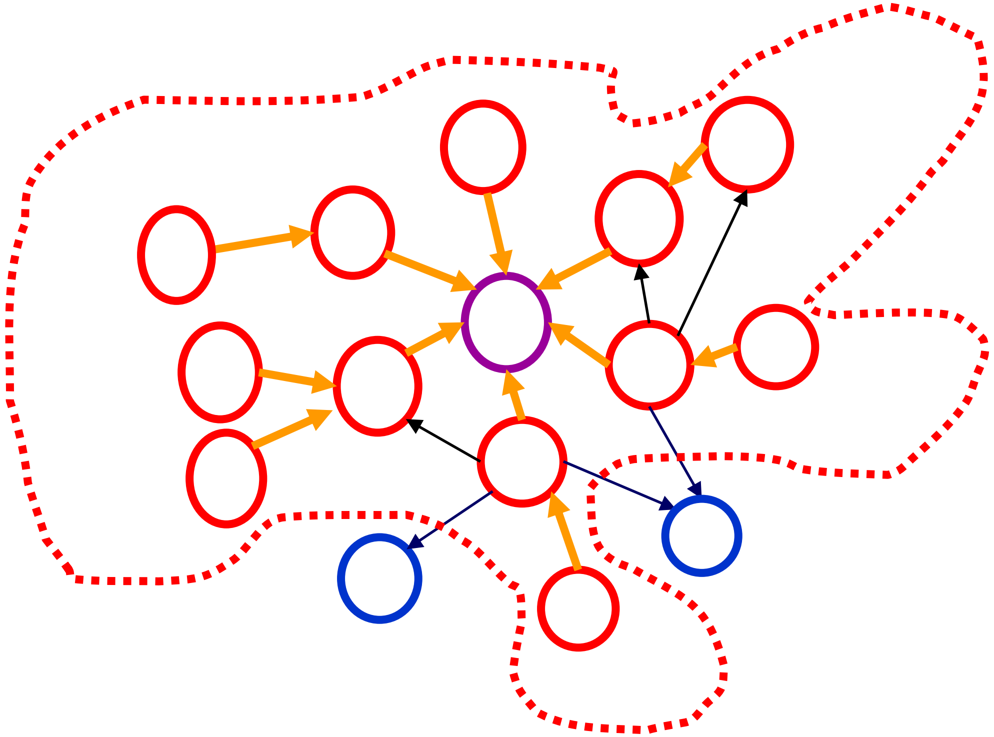
- usually consider only vertices from which the node *i* in question can be reached



Closeness Centrality

Influence range

- The influence range of *i* is the set of vertices which are reachable from the node *i*



Centrality in Networks

Comparing across these 3 centrality measures

- Generally, the 3 centrality types will be positively correlated
- When they are not (low) correlated, it probably tells you something interesting about the network.

	Low Degree	Low Closeness	Low Betweenness
High Degree		Embedded in cluster that is far from the rest of the network	Node’s connections are redundant - communication bypasses the node
High Closeness	Tied to important/active nodes		Probably multiple paths in the network, node is near many nodes, but so are many others
High Betweenness	Node’s few ties are crucial for network flow	Would mean that node monopolizes the ties from a small number of nodes to many others.	

Eigenvalues and eigenvectors

- **Eigenvalues and eigenvectors** have their origins in physics
    - in particular in problems where motion is involved
    - their uses extend from solutions to stress and strain problems to differential equations and quantum mechanics
- **Eigenvectors are vectors that point in directions where there is no rotation**
    - Eigenvalues are the change in length of the eigenvector from the original length
- The basic equation in eigenvalue problems is:

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$$

Eigenvalues and eigenvectors

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

- In words, this simple equation says that for the square matrix **A**, there is a vector **x** such that the product of **Ax** is equal to a SCALAR λ multiplied by **x**
- The multiplication of vector **x** by a scalar constant is the same as stretching or shrinking the coordinates by a constant value
- The vector **x** is called an **eigenvector** and the scalar λ, is called an **eigenvalue**.

Eigenvector Centrality

In many circumstances a vertex's importance in a network is increased by having connections to other vertices that are *themselves important*. Instead of awarding vertices just one point for each neighbor, *eigenvector centrality* gives each vertex a score proportional to the sum of the scores of its neighbors. This is a natural extension of *degree centrality*.

- ➢ Adjacency matrix redistributes vertex contents
- ➢ Some vector of contents is in equilibrium
- ➢ These are the eigenvector centralities

Eigenvector Centrality

The eigenvector centrality can be obtained starting from an initial guess of the centrality of the nodes $x_i^{(0)}$ and by considering the following recursive process:

$$x_i^{(t)} = \sum_j A_{ij} x_j^{(t-1)} = \sum_j A_{ij}^t x_j^{(0)}$$

We can now write this in matrix form,

$$\boldsymbol{x}^{(t)} = \boldsymbol{A}^t \boldsymbol{x}^{(0)}$$

Here $\boldsymbol{A}$ is the adjacency matrix. We now write $\boldsymbol{x}^{(0)} = \sum_i c_i \boldsymbol{v}_i$ where $\boldsymbol{v}_i$ are the eigenvectors and c_i are some constants.

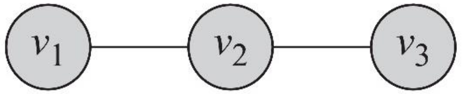
Assume that c_1 is the largest constant, we can now write the other c_i's as $c_i = c_1 k_i$, with all k_i's less than 1. As t → ∞ , $\boldsymbol{x}^{(t)} = c_1 k_1^t \boldsymbol{v}_1$. In other words, the limiting vector of centralities is simply proportional to the leading eigenvector of the adjacency matrix. Hence, we could say that the centrality $\boldsymbol{x}$ satisfies

$$\boldsymbol{A}\boldsymbol{x} = k_1 \boldsymbol{x}$$

The centrality of vertex $\boldsymbol{i}$ is proportional to the sum of the centralities of *i's* neighbors, $x_i = \frac{1}{k_1} \sum_j A_{ij} x_j$

***It can be large either because a vertex has many neighbors or because it has important neighbors (or both).

Eigenvector Centrality: Example 1



$$\lambda \mathbf{C}_e = A\mathbf{C}_e \qquad (A - \lambda I)\mathbf{C}_e = 0 \quad \mathbf{C}_e = [u_1 \; u_2 \; u_3]^T$$

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

$$\begin{bmatrix} 0-\lambda & 1 & 0 \\ 1 & 0-\lambda & 1 \\ 0 & 1 & 0-\lambda \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$
$$= aei + bfg + cdh - ceg - bdi - afh.$$

$$det(A - \lambda I) = \begin{vmatrix} 0-\lambda & 1 & 0 \\ 1 & 0-\lambda & 1 \\ 0 & 1 & 0-\lambda \end{vmatrix} = 0$$

$$(-\lambda)(\lambda^2 - 1) - 1(-\lambda) = 2\lambda - \lambda^3 = \lambda(2 - \lambda^2) = 0$$

Eigenvalues are

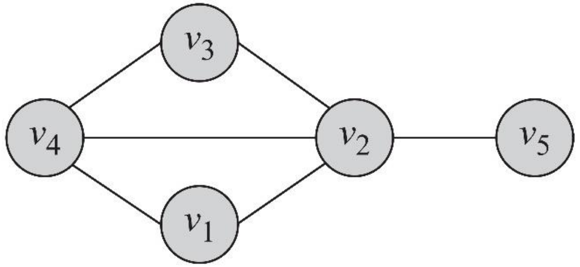
$$(-\sqrt{2}, 0, \boxed{+\sqrt{2})}$$

Largest Eigenvalue

Corresponding eigenvector

$$\begin{bmatrix} 0-\sqrt{2} & 1 & 0 \\ 1 & 0-\sqrt{2} & 1 \\ 0 & 1 & 0-\sqrt{2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{C}_e = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 1/2 \\ \sqrt{2}/2 \\ 1/2 \end{bmatrix}$$

Eigenvector Centrality: Example 2



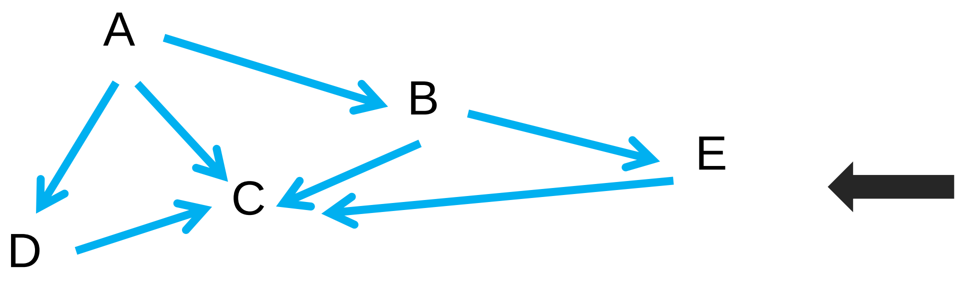
$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \lambda = (2.68,\ -1.74, -1.27,\ 0.33, 0.00)$$

Eigenvalues

$$\lambda_{\text{max}} = 2.68 \Rightarrow C_e = \begin{bmatrix} 0.4119 \\ 0.5825 \\ 0.4119 \\ 0.5237 \\ 0.2169 \end{bmatrix}$$

Eigenvector Centrality

- Eigenvector centrality can also be calculated for ***directed graphs*** but there are problems
  - We need to decide between incoming or outgoing edges



All the nodes in this graph have zero centrality

- A has no incoming edges, hence a centrality of 0
- B has only an incoming edge from A
  - hence its centrality is also 0
- Only vertices that are in a strongly connected component of two or more vertices or the out-component of such a component have non-zero centrality
- Acyclic networks such as citation networks, have no strongly connected components of more than one vertex, so all vertices will have centrality zero.

Katz Centrality

The ***Katz centrality*** has been proposed to solve the undesired property of the eigenvector centrality , i.e. the vanishing of the *eigenvector centrality* for the nodes in the in-component of directed network. It assigns to each node a small centrality value just regardless of its position in the network or the centrality of its neighbors, then the centrality of the node increases if many important nodes point to it.

The Katz centrality *x* satisfies the following equation

$$x_i = \alpha \sum_{j=1}^{N} A_{ij} x_j + \beta$$

where β > 0 and $\alpha \in (0, \frac{1}{\lambda_1})$ is a scaling constant, λ_1 is the Perron-Frobenius eigenvalue of the adjacency matrix $\boldsymbol{A}$.

In matrix form,

$$\boldsymbol{x} = \alpha \boldsymbol{A} \boldsymbol{x} + \beta \boldsymbol{1}$$

where **1** is the column vector with $1_i = 1;\ \forall i = 1, 2, \ldots, N$.

Katz Centrality

Rearranging, we have

$$\boldsymbol{x} = (\mathbf{I} - \alpha\boldsymbol{A})^{-1}\beta\mathbf{1}$$

The matrix diverges for det($\mathbf{I}$ - $\alpha\boldsymbol{A}$) = 0. This is the reason to set α to take the value in the above range. We can also set β=1 for convenience since one only cares about the ratio between the eigenvector components.

Most researchers have employed α to take values close to the maximum of $1/\lambda_1$, which places the maximum amount of weight on the eigenvector term and the smallest amount on the constant term, resulting in a centrality that is numerically quite close to the ordinary eigenvector centrality, but gives small non-zero values to vertices that are not in the strongly connected components or their out-components.

Extend to the case with different β's, e.g., in a social network the importance of an individual might depend on non-network factors such as their age, income, etc.

$$x_i = \alpha \sum_{j=1}^{N} A_{ij} x_j + \beta_i$$

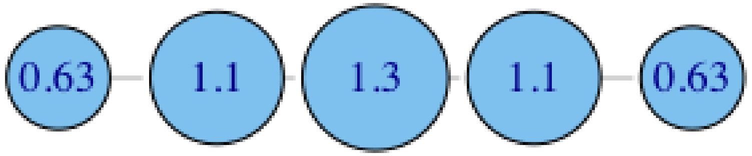
We then have

$$\boldsymbol{x} = (\mathbf{I} - \alpha\boldsymbol{A})^{-1}\boldsymbol{\beta}$$

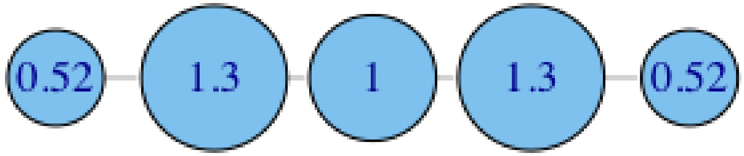
where β is now a column vector.

Katz Centrality: Examples

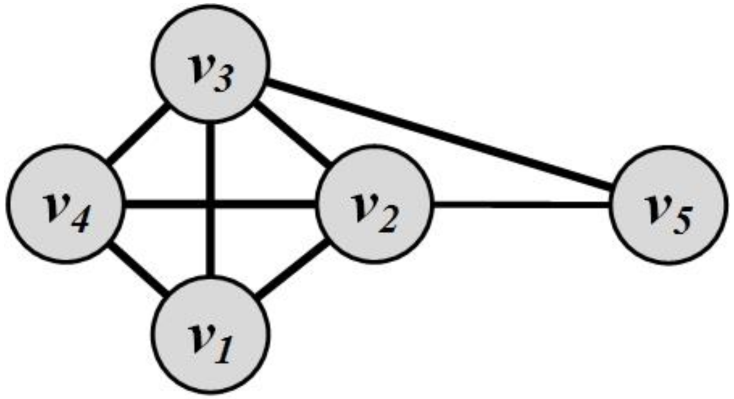
β=.25



β=-.25



Katz Centrality: Examples



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

- The Eigenvalues are -1.68, -1.0, -1.0, 0.35, 3.32
- We assume α=0.25 < 1/3.32 and $\beta = 0.2$

$$\boldsymbol{x}_{Katz} = \beta(\mathbf{I} - \alpha \boldsymbol{A}^T)^{-1}.\mathbf{1} = \begin{bmatrix} 1.14 \\ 1.31 \\ 1.31 \\ 1.14 \\ 0.85 \end{bmatrix}$$

Most important nodes!

Centrality in Social Networks

There are other options, usually based on generalizing some aspect of those above:

•***Random Walk Betweenness*** Looks at the number of times you would expect node *i* to be on the path between *k* and *j* if information traveled a 'random walk' through the network.

•***Peer Influence*** based on the assumed network autocorrelation model of peer influence. In practice it's a variant of the eigenvector centrality measures.

•***Subgraph centrality*** Counts the number of cliques of size 2, 3, 4, … *n*-1 that each node belongs to. Reduces to (another) function of the eigenvalues. Similar to influence & information centrality, but does distinguish some unique positions.

•***Fragmentation centrality*** Key Player idea where nodes are central if they can easily break up a network.

•***Embeddedness*** measure is a group-level index, but captures the extent to which a given set of nodes are nested inside a network

•***Removal Centrality*** effect on the rest of the (graph for any given statistic) with the removal of a given node. Gets at the system-contribution of a particular actor.

PageRank

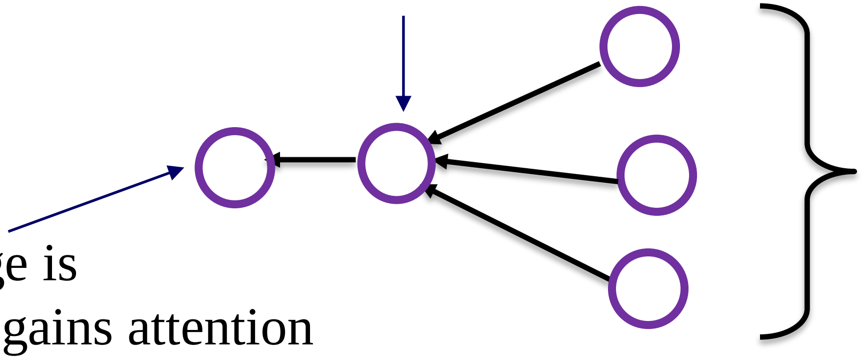
- Problem with Katz Centrality:
  - In directed graphs, once a node becomes an authority (high centrality), it passes **all** its centrality along **all** of its out-links. This is less desirable since not everyone known by a well-known person is well-known

- **Solution?**
  - We can divide the value of passed centrality by the number of outgoing links, i.e., out-degree of that node
  - Each connected neighbor gets a fraction of the source node's centrality

PageRank: Bringing order to the web

- It’s in the links:
  - links to URLs can be interpreted as endorsements or recommendations
  - the more links a URL receives, the more likely it is to be a good/entertaining/provocative/authoritative/interesting information source
  - but not all link sources are created equal
    - a link from a respected information source
    - a link from a page created by a spammer

an important page, e.g. twitter



Many webpages scattered across the web

if a web page is twittered, it gains attention

PageRank

The PageRank centrality *x* satisfies the following equation

$$x_i = \alpha \sum_{j=1}^{N} A_{ij} \frac{1}{K_j^{out}} x_j + \beta$$

where $K_j^{out} = \max\left(k_j^{out}, 1\right)$, $\beta > 0$ and $\alpha \in (0, \frac{1}{\lambda_1})$ is a scaling constant, λ_1 is the Perron-Frobenius eigenvalue of the matrix ***A*/*D*** with ***D*** given by the diagonal matrix of elements $D_{ii} = K_j^{out} = \max\left(k_j^{out}, 1\right)$.

In matrix formalism, one has

$$\boldsymbol{x} = \alpha \boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{x} + \beta\boldsymbol{1}$$

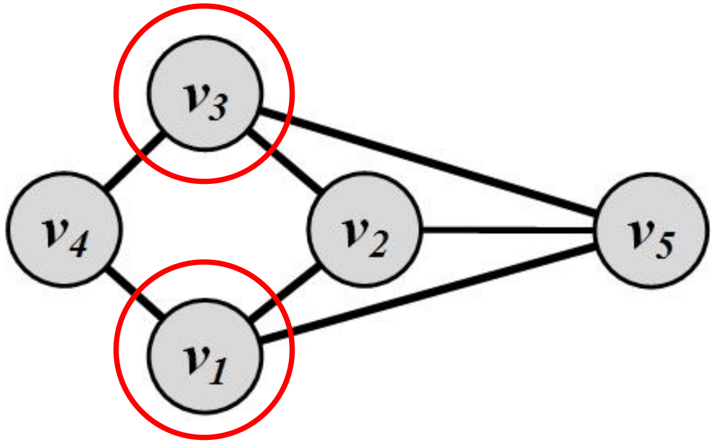
Rearranging, we have

$$\boldsymbol{x} = (\mathbf{I} - \alpha\boldsymbol{A}\boldsymbol{D}^{-1})^{-1}\beta\boldsymbol{1}$$

The condition $\alpha < 1/\lambda_1$ guarantees that the PageRank centrality is well defined for every node of the network. In undirected network, λ_1=1. Therefore $\alpha \in (0,1)$, in directed network λ_1 is order one. In the original PageRank algorithm of Google $\alpha \approx 0.85$. Again, one can set $\beta = 1$ for convenience.

PageRank: Example

- Assume $\alpha = 0.95$ (<1) and $\beta = 0.1$



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

$$\boldsymbol{x}_{pagerank} = (\mathbf{I} - \alpha \boldsymbol{A}\boldsymbol{D}^{-1})^{-1} \beta \mathbf{1} = \begin{bmatrix} 2.14 \\ 2.13 \\ 2.14 \\ 1.45 \\ 2.13 \end{bmatrix}$$

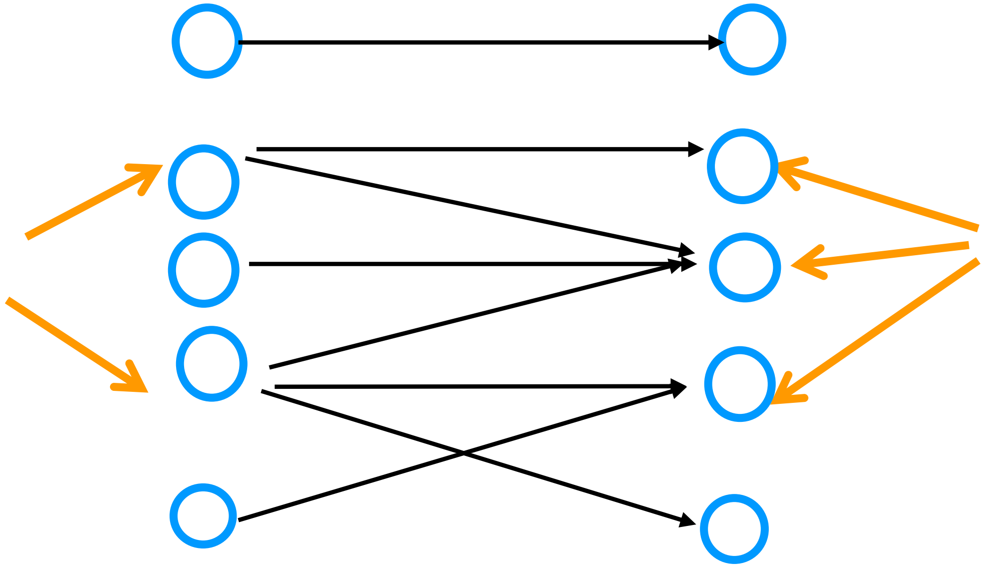
Matrix-based Centrality measures

	without constant term	with constant term
Divide by out-degree	Degree centrality	PageRank
No division	Eigenvector centrality	Katz centrality

Hubs and Authorities

- In directed networks, vertices that point to important resources should also get a high centrality
    - e.g. review articles, web indexes

- recursive definition:



hubs are nodes
that links to
good authorities

authorities are nodes
that are linked to by
good hubs

Authorities are nodes that contain useful information on a topic of interest; *hubs* are nodes that tell us where the best authorities are to be found. An *authority* may also be a *hub*, and vice versa: review articles often contain useful discussions of the topic at hand as well as citations to other discussions.

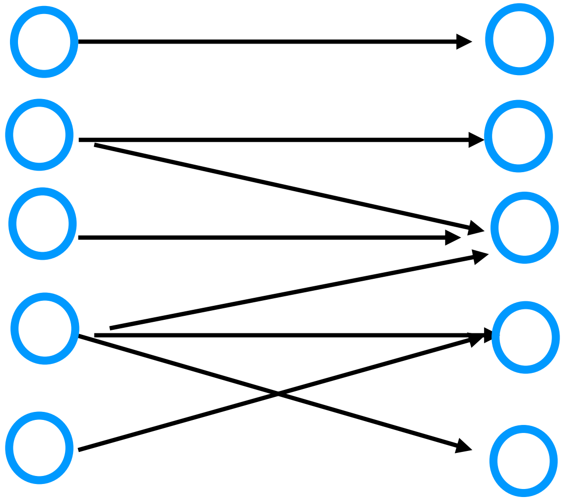
Hyperlink-Induced Topic Search

- HITS algorithm
  - start with a set of pages matching a query
  - expand the set by following forward and back links
  - take transition matrix E, where the i,jth entry $E_{ij}=1/n_i$
    - where i links to j, and n_i is the number of links from i
  - then one can compute the *authority scores* ***a***, and *hub scores* ***h*** through an iterative approach:

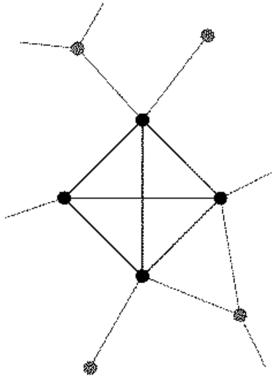
$$\underline{a}' = E^T \underline{h} \qquad \underline{h}' = Ea$$

$$A(p) = \sum_{q \in B | q \to p} H(q)$$

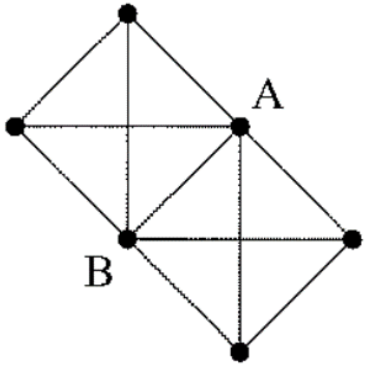
$$H(p) = \sum_{q \in B | p \to q} A(q)$$



Cliques, Plexes and Cores



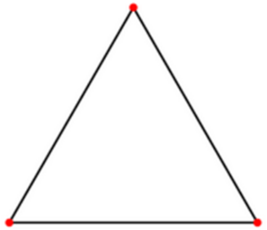
A clique of four vertices within a network.



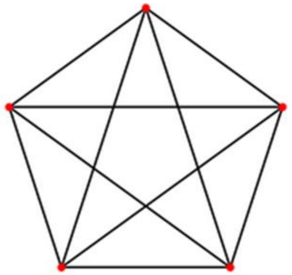
Two overlapping cliques. Vertices A and B in this network both belong to two cliques of four vertices.

A ***clique*** is a maximal subset of the vertices in an undirected network such that every member of the set is connected by an edge to every other.

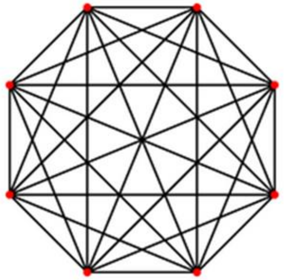
- Alternatively, we can also say that K_n is the complete graph (clique) with K vertices
  - each vertex is connected to every other vertex
  - there are $n(n-1)/2$ undirected edges



K_3

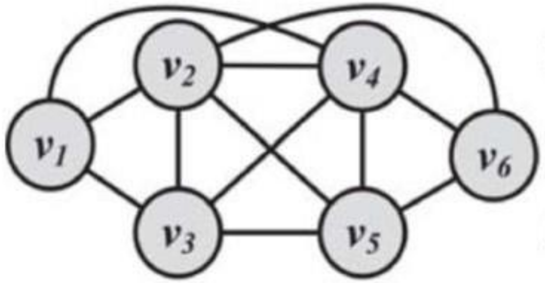


K_5



K_8

Cliques, Plexes and Cores



1-plex :$\{v_2, v_3, v_4, v_5\}$

2-plex :$\{v_1, v_2, v_3, v_4, v_5\}, \{v_2, v_3, v_4, v_5, v_6\}$

3-plex :$\{v_1, v_2, v_3, v_4, v_5, v_6\}$

- A k-plex is maximal if it is not contained in a larger k-plex (i.e., with more nodes)
- Finding the maximum k-plex is still NP-hard
  - In practice it is easier to due to smaller search space

A ***k-plex*** of size *n* is a maximal subset of *n* vertices within a network such that each vertex is connected to at least *n* - *k* of the others. If *k* = 1, we recover the definition of an ordinary clique -- a *1-plex* is the same as a clique. If *k* = 2, then each vertex must be connected to all or all-but-one of the others.

Like cliques, *k-plexes* can overlap one another; a single vertex can belong to more than one *k-plex*.

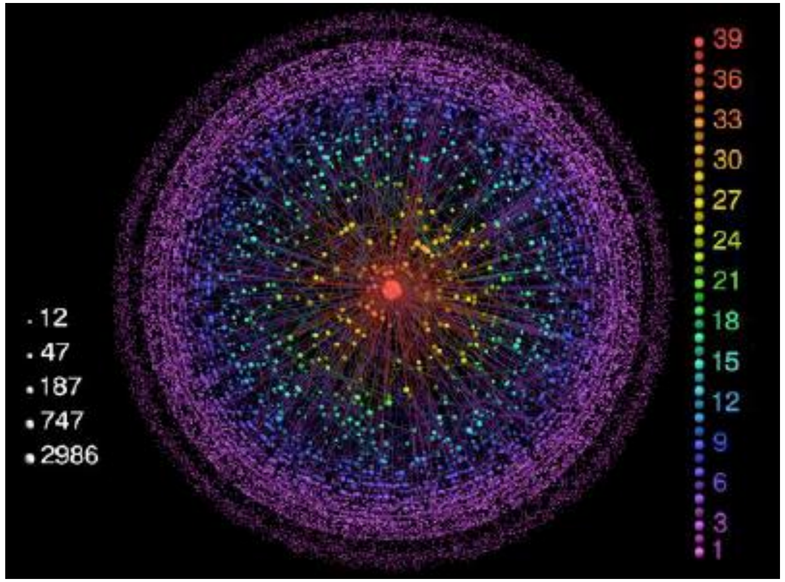
Cliques, Plexes and Cores

A network is organized as a set of successively enclosed *k-cores* (like a Russian nesting doll).

A *k-core* is a maximal subset of vertices such that each is connected to at least *k* others in the subset. (**Prove for yourself that a *k-core* of *n* vertices is also an (*n* - *k*)-*plex*).

K-core algorithm:

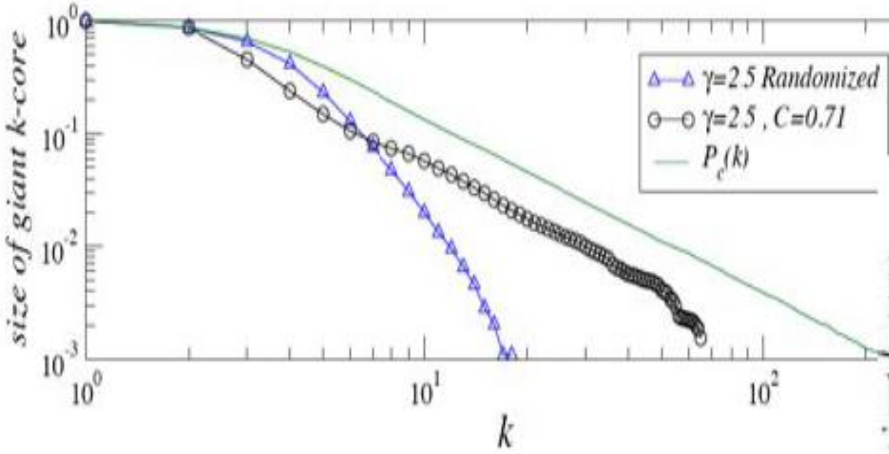
1. Remove from the original graph all vertices (and their connections) with degree less than k
2. Remove from the remaining graph all vertices (and their connections) with degree less than k
3. Repeat step (2) until no further removal is possible

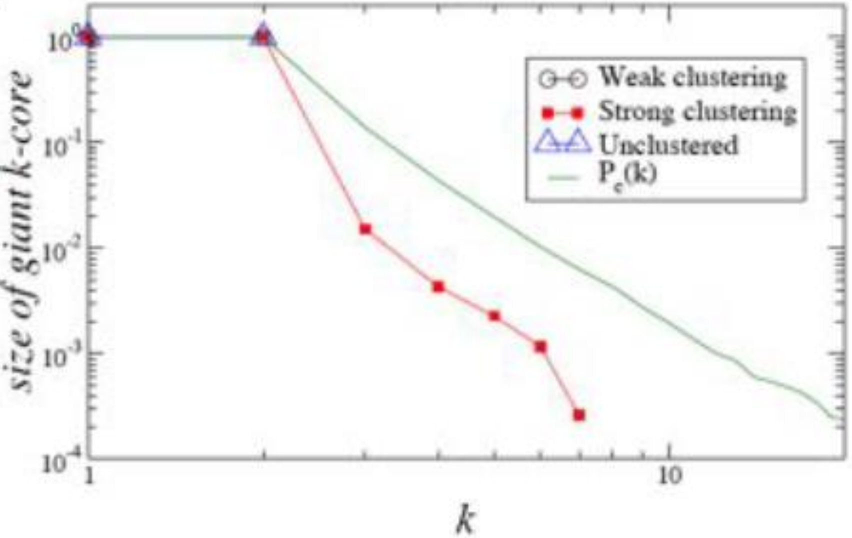


Cliques, Plexes and Cores

- The *k*-core organization of a network gives idea of its sparsity/tree likeness.

- ***Strong clustering*** produces a deeper hierarchy of *k*-cores.





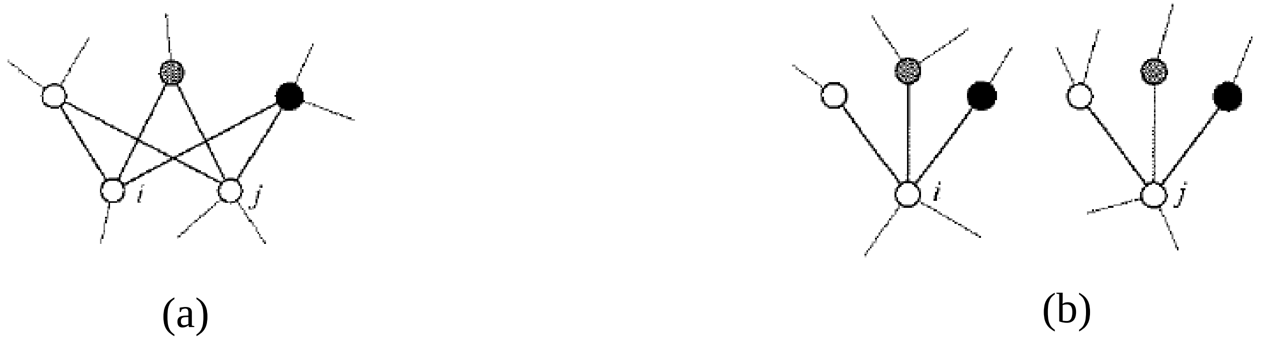
Similarity

Another central concept in social network analysis is *similarity between vertices*. In what ways can vertices in a network be similar, and how can we *quantify* that similarity?

Two fundamental approaches to constructing measures of network similarity, called *structural equivalence* and *regular equivalence*.

Two vertices in a network are *structurally equivalent* if they share many of the same network neighbors.

Two *regularly equivalent* vertices do not necessarily share the same neighbors, but they have neighbors *who are themselves similar*.



Structural equivalence and regular equivalence. (a) Vertices *i* and *j* are structurally equivalent if they share many of the same neighbors. (b) Vertices *i* and *j* are regularly equivalent if their neighbors are themselves equivalent (indicated here by the different shades of vertices).

Similarity: Structural Equivalence

Cosine Similarity

The simplest and most obvious measure of structural equivalence would be just a count of the number of common neighbors two vertices have. In an undirected network the number n_{ij} common neighbors of vertices *i* and j is given by

$$n_{ij} = \sum_k A_{ik} A_{kj}$$

which is the *ij*-th element of $\boldsymbol{A}^2$. This is essentially the same thing as the "co-citation" measure introduced for directed networks. One should normalize this with the varying degrees of the vertices. A common measure is the *cosine similarity* (or *Salton's cosine*).

$$\sigma_{ij} = \cos\theta = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}}$$

For an unweighted simple graph, the elements of the adjacency matrix take only values 0 and 1, so that $\sum_k A_{ik}^2 = \sum_k A_{ik} = k_i$, where k_i is the degree of vertex *i.* Then

$$\sigma_{ij} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{k_i}\sqrt{k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

The cosine similarity for figure (a) above, $\sigma_{ij} = \frac{3}{\sqrt{4 \times 5}} \cong 0.671$. Notice that σ_{ij} is always between 0 and 1.

Similarity: Structural Equivalence

Pearson Coefficients

Suppose vertices i and j have degrees k_i and k_j respectively. How many common neighbors should we expect them to have?

If they choose their neighbors purely at random, the answer is $k_i k_j / n$, where n is the total number of vertices in the graph. A reasonable measure of similarity between two vertices is the *actual* number of common neighbors they have minus the *expected* number that they *would* have if they chose their neighbors at random, i.e.

$$\sum_k A_{ik} A_{kj} - \frac{k_i k_j}{n} = \sum_k A_{ik} A_{kj} - \frac{1}{n} \sum_k A_{ik} \sum_l A_{jl} = \sum_k A_{ik} A_{kj} - n < A_i >< A_j >$$

$$= \sum_k [\, A_{ik} A_{kj} - < A_i >< A_j >] = \sum_k (\, A_{ik} - < A_i >)(A_{kj} - < A_j >) = n \operatorname{cov}(A_i, A_j)$$

where $< A_i >$ denotes the mean $\frac{1}{n} \sum_k A_{ik}$ of the elements of the i-th row of the adjacency matrix, $\operatorname{cov}(A_i, A_j)$ is the covariance of the two rows of the adjacency matrix.

Similarity: Structural Equivalence

Pearson Coefficients

Normalizing by the variances of A_i and A_j gives the standard Pearson correlation coefficient:

$$r_{ij} = \frac{\mathrm{cov}(A_i, A_j)}{\sigma_i \sigma_j} = \frac{\sum_k (A_{ik} - < A_i >)(A_{kj} - < A_j >)}{\sqrt{\sum_k (A_{ik} -< A_i >)^2} \sqrt{\sum_k (A_{jk} -< A_j >)^2}}$$

where -1 < r_{ij} < 1. If r_{ij} is positive, the vertices are more similar, otherwise they are dissimilar.

Similarity: Structural Equivalence

Other measures:

Another measure of structural equivalence is the so-called ***Euclidean distance*** d_{ij}, which is equal to the number of neighbors that differ between two vertices. (Note: this way of defining the measure is essentially the same as the *Hamming distance* in computer science.)

$$d_{ij} = \sum_k (A_{ik} - A_{jk})^2 \, .$$

Euclidean distance is really a dissimilarity measure, since it is larger for vertices that differ more.
The maximum value of d_{ij} is $k_i + k_j$. Normalizing by this maximum value gives

$$\frac{\sum_k (A_{ik} - A_{jk})^2}{k_i + k_j} = \frac{\sum_k (A_{ik} + A_{jk} - 2A_{ik}A_{jk})}{k_i + k_j} = 1 - 2\frac{n_{ij}}{k_i + k_j}$$

where we have used $A_{ij}^2 = A_{ij}$, and n_{ij} is again the number of neighbors that *i* and *j* have in common.
Within additive and multiplicative constants, this *normalized Euclidean distance* can thus be regarded as just another alternative normalization of the number of common neighbors.

Similarity: Regular Equivalence

Quantitative measures of regular equivalence are less well developed than measures of structural equivalence. One way is to define a similarity score σ_{ij} such that *i* and *j* have high similarity if they have neighbors *k* and *l* that themselves have high similarity. For an undirected network we can write this as

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik}\sigma_{kj} + \delta_{ij}$$

In matrix form,

$$\boldsymbol{\sigma} = \alpha \boldsymbol{A}\boldsymbol{\sigma} + \boldsymbol{I} = (\boldsymbol{I} - \alpha\boldsymbol{A})^{\mathbf{-1}}$$

This equation is reminiscent of the formula for the Katz centrality.