Machine Learning with Graphs (MLG)

Centrality Analysis in Graphs

Cheng-Te Li (李政德)

Institute of Data Science

National Cheng Kung University

chengte@mail.ncku.edu.tw

Network Centrality

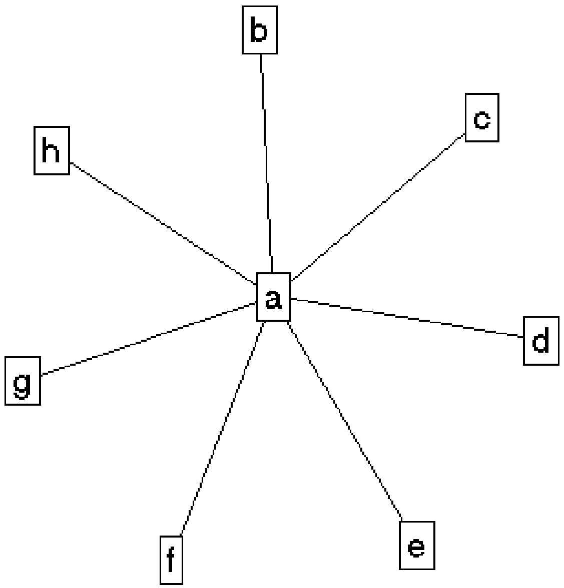
- We want to know which nodes are the most "central" or important in a social network
- Centrality Measures
  - Local: Degree
  - Global: Closeness, Betweenness, Eigenvector
- Centralization
  - How evenly is centrality distributed among nodes?

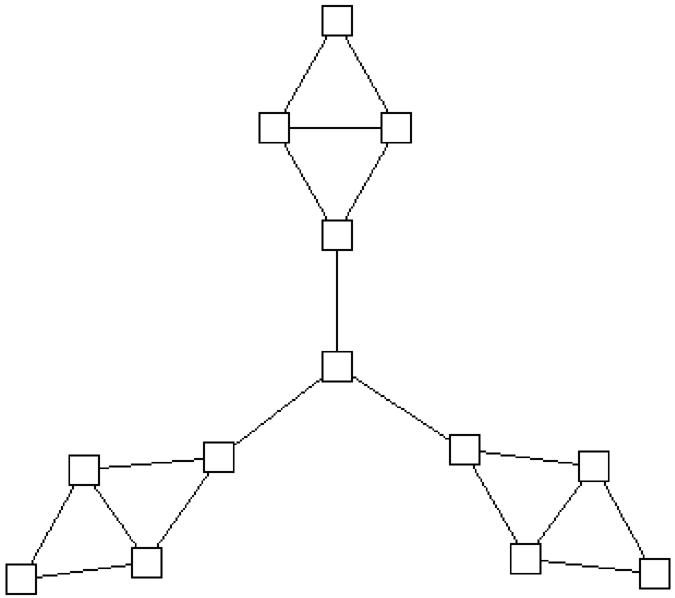
References

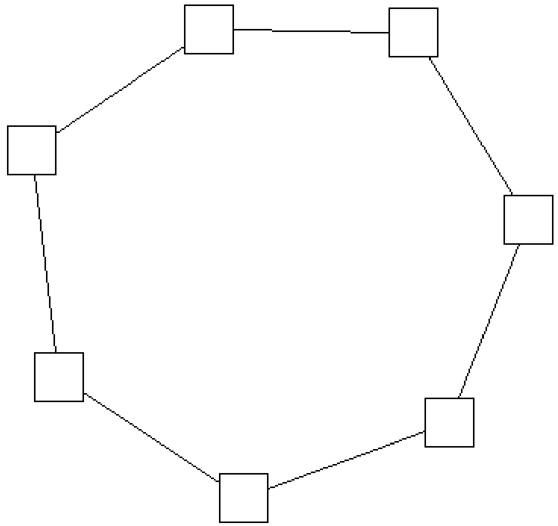
- Chapter 10 of “Introduction to social network methods”
  - http://www.faculty.ucr.edu/~hanneman/nettext/index.html
- Chapter 4 “Extending Centrality” of the book “Models and methods in social network analysis”. Peter J. Carrington, John Scott, Stanley Wasserman. 2005.
- Chapter 5 of “Social Network Analysis: Methods and Applications” S. Wasserman, K. Faust. 1994.
- L. C. Freeman. “Centrality in Social Networks: Conceptual Clarification”, Social Networks, 1: 215-239, 1978. (16361 cites)

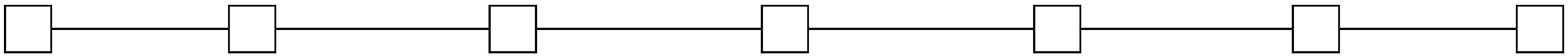
Centrality in Networks

- A measure allowing us to identify the "central" actors





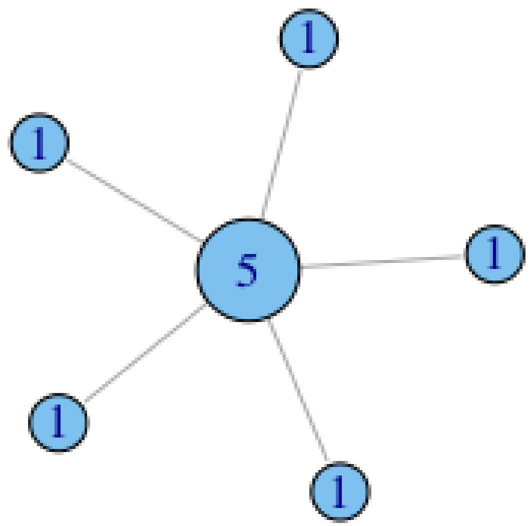




Degree Centrality

- Intuition
  - Those have many friends are centrality nodes
- Degree centrality: ranks nodes with more connections higher in terms of centrality
  - d_i is the degree (number of friends) for node v_i

$$C_d(v_i) = d_i$$



Degree Centrality in Directed Graphs

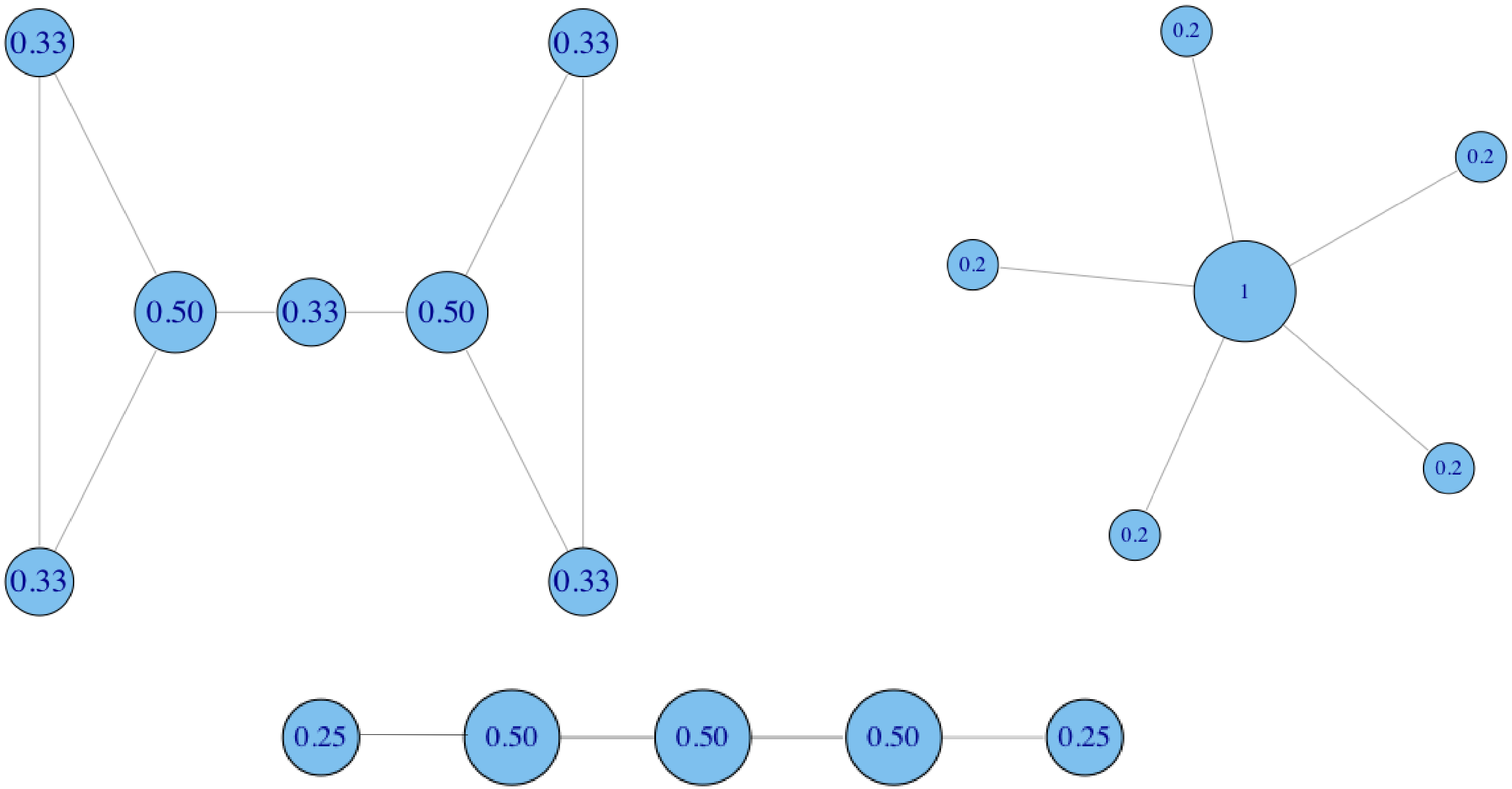
- In directed graphs, we can either use the in-degree, the out-degree, or the combination as the degree centrality value
- In practice, mostly in-degree is used

$$
\begin{aligned}
C_d(v_i) &= d_i^{\text{in}} && (prestige) \\
C_d(v_i) &= d_i^{\text{out}} && (gregariousness) \\
C_d(v_i) &= d_i^{\text{in}} + d_i^{\text{out}}
\end{aligned}
$$

d_i^{out} is the number of outgoing links for node v_i

Normalized Degree Centrality

- When comparing to nodes of other networks, it needs to be normalized by network size



Normalized Degree Centrality

- Normalized by the maximum possible degree

$$C_d^{\text{norm}}(v_i) = \frac{d_i}{n-1}$$

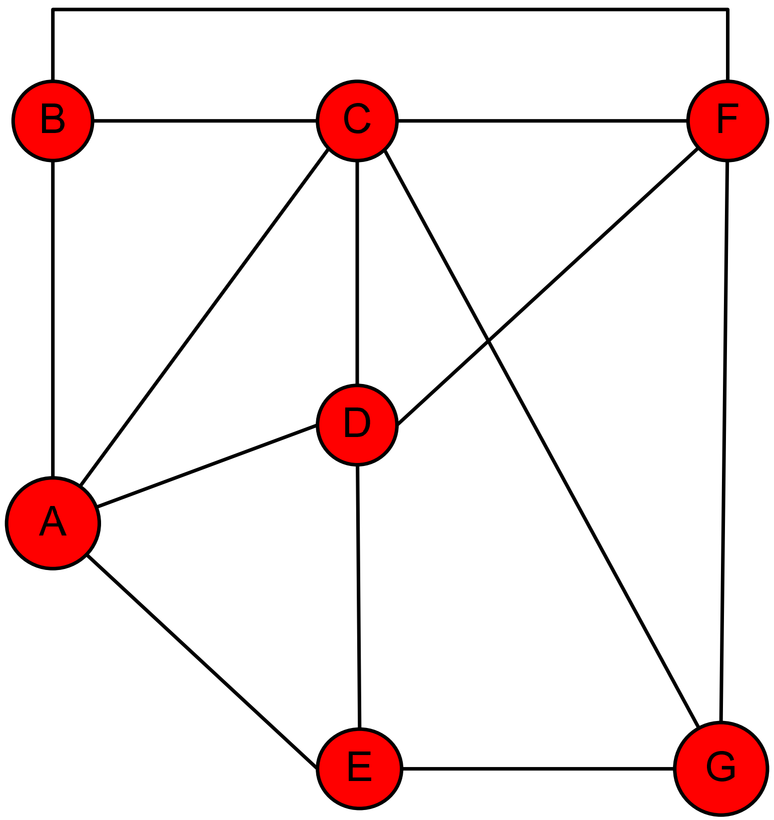
- Normalized by the maximum degree

$$C_d^{\max}(v_i) = \frac{d_i}{\max_j d_j}$$

- Normalized by the degree sum

$$C_d^{\text{sum}}(v_i) = \frac{d_i}{\sum_j d_j} = \frac{d_i}{2|E|} = \frac{d_i}{2m}$$

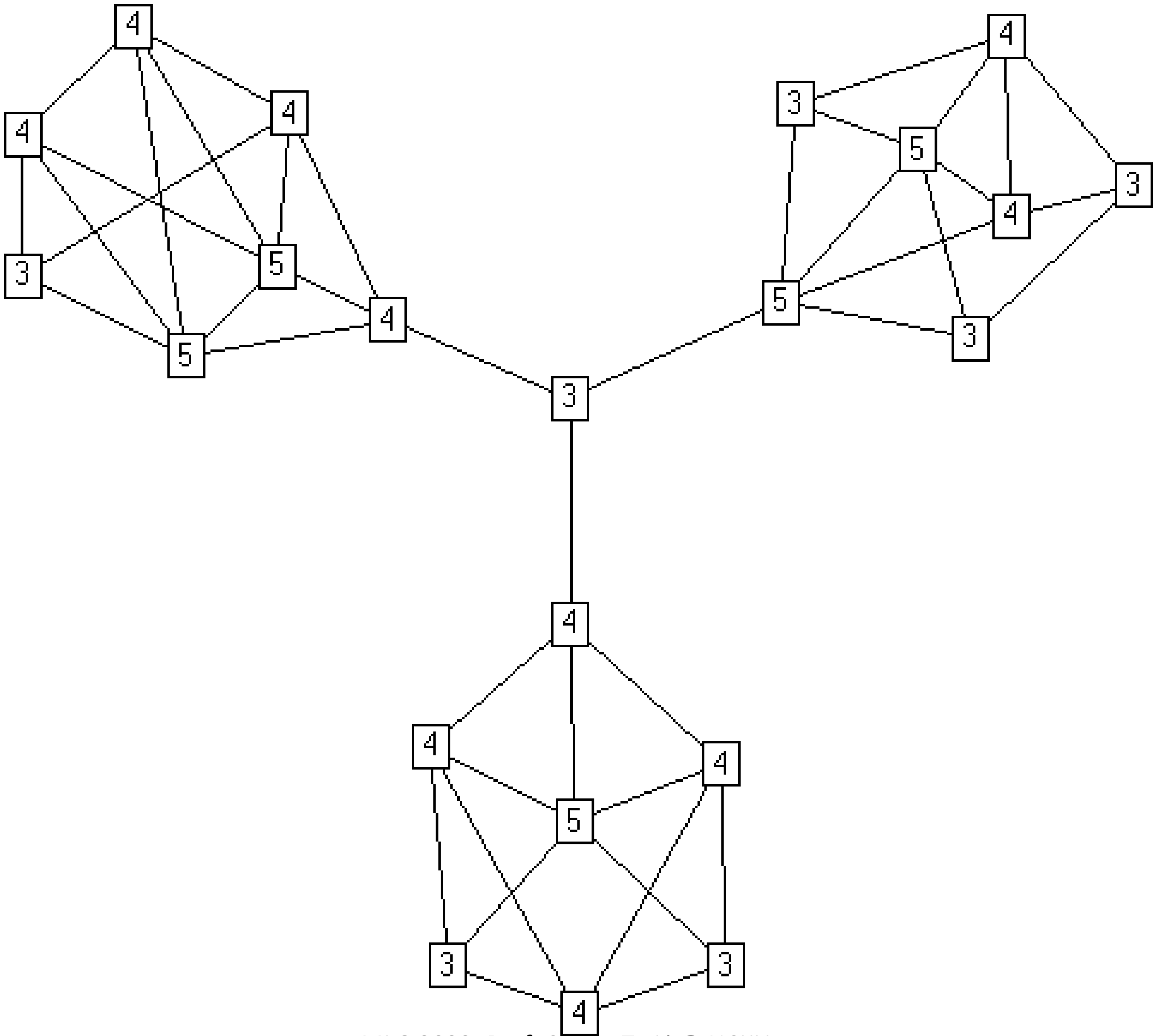
Degree Centrality (Undirected Graph) Example



Node	Degree	Centrality	Rank
A	4	2/3	2
B	3	1/2	5
C	5	5/6	1
D	4	2/3	2
E	3	1/2	5
F	4	2/3	2
G	3	1/2	5

Problems for Degree Centrality

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



Centralization

- How much variation is there in the centrality scores among the nodes?
- Simple variance of the individual centrality

$$S_D^2 = \frac{\sum_{i=1}^{N}(C_D(v_i) - \boxed{\bar{C}_D})^2}{N}$$

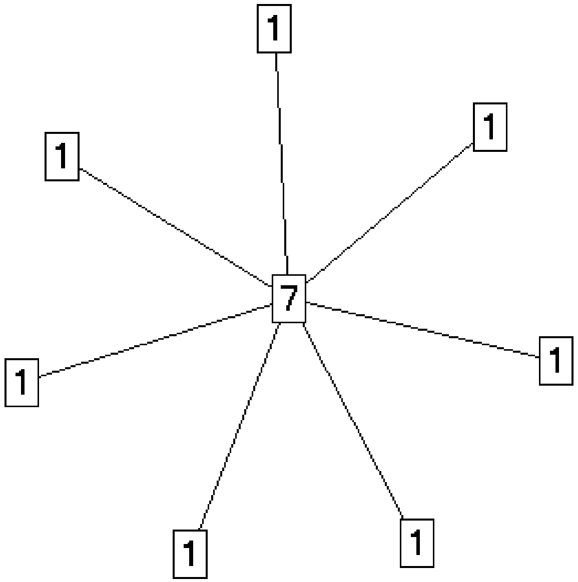
average degree

- Freeman's general formula for centralization

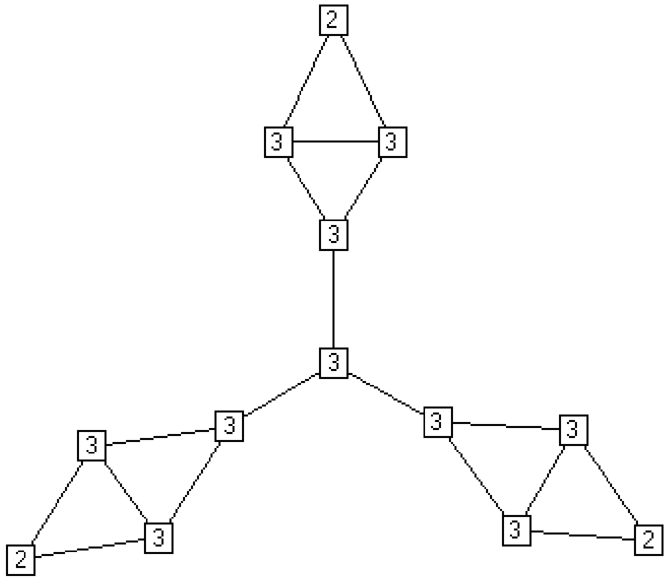
maximum centrality value in the network

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

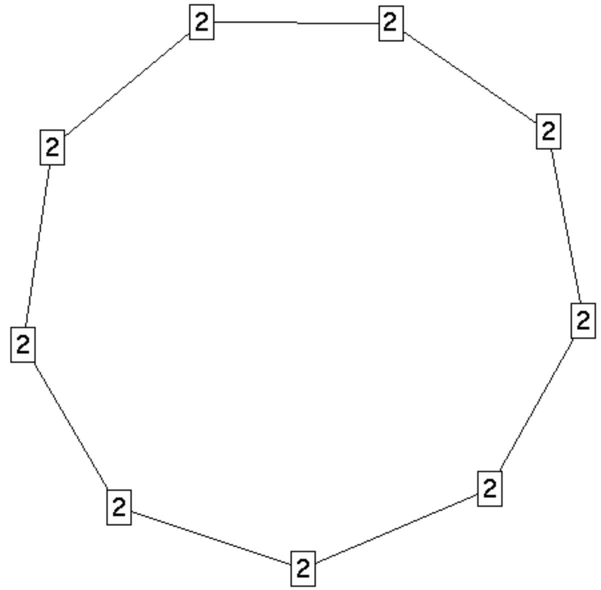
Example for Degree Centralization (1/2)



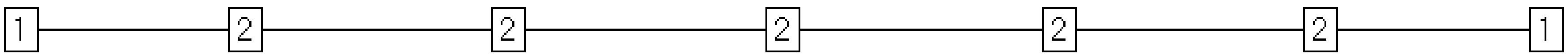
Freeman: 1.0
Variance: 3.9



Freeman: 0.02
Variance: 0.17



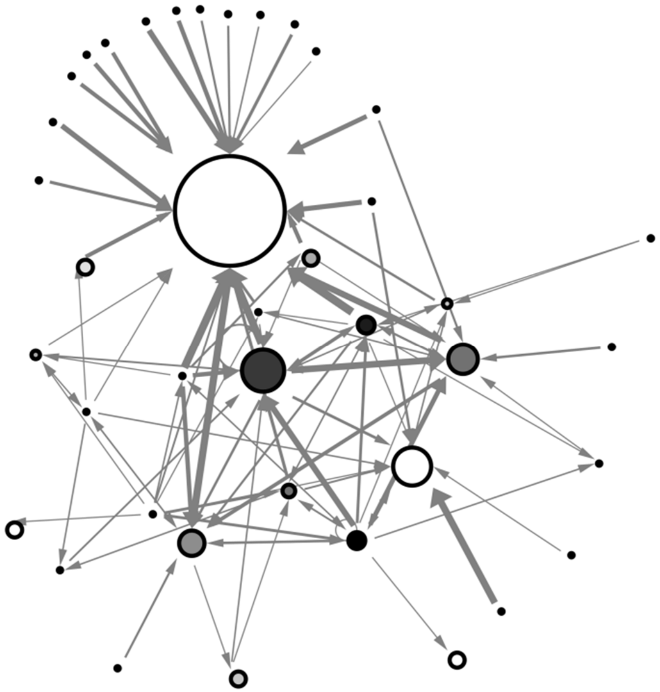
Freeman: 0.0
Variance: 0.0

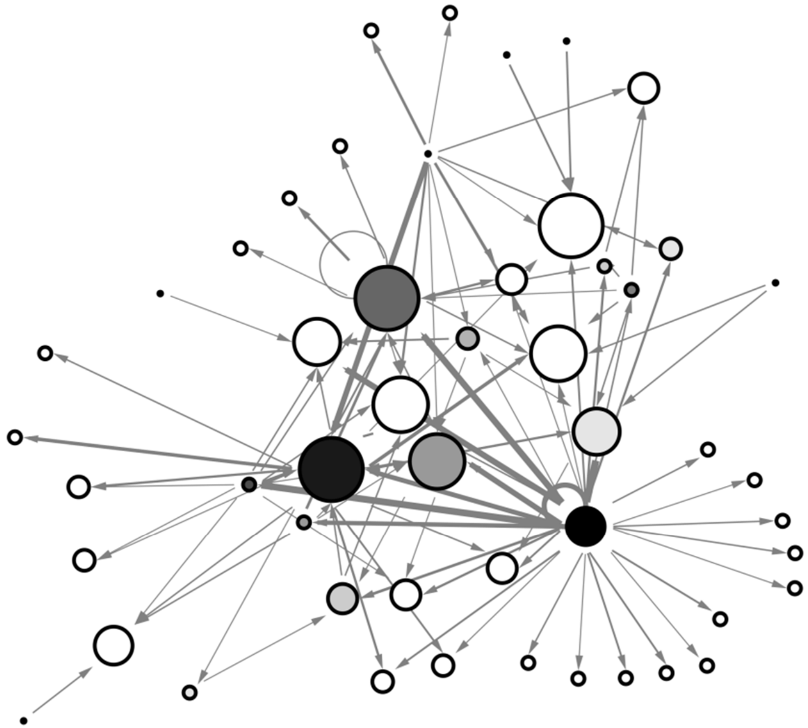


Freeman: 0.07
Variance: 0.20

Example for Degree Centralization (2/2)

Example on financial trading networks



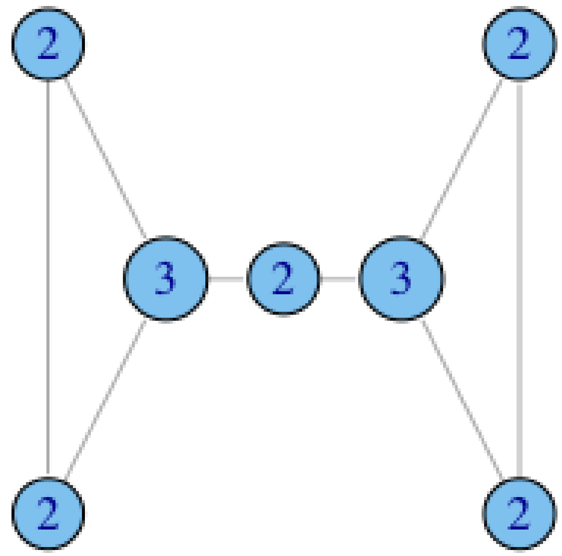


high centralization: one node trading with many others

low centralization: trades are more evenly distributed

Global Centrality Measures

- Sometimes degree fails to capture centrality...
  - Example

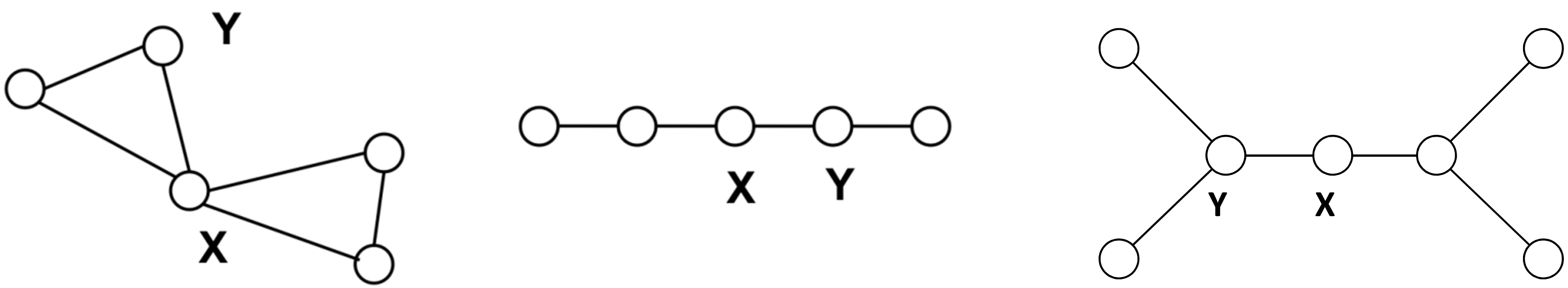


Brokers between groups

- Global measure
  - Betweenness
  - Closeness
  - Eigenvector

Betweenness Centrality

- Intuition
  - How many pairs of individuals would have to go through the node in order to reach one another within the minimum number of hops?
- Example
  - Who has higher betweenness, X or Y?



Betweenness Centrality

$$C_B(i) = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

the number of shortest paths connecting jk *that passes node* i

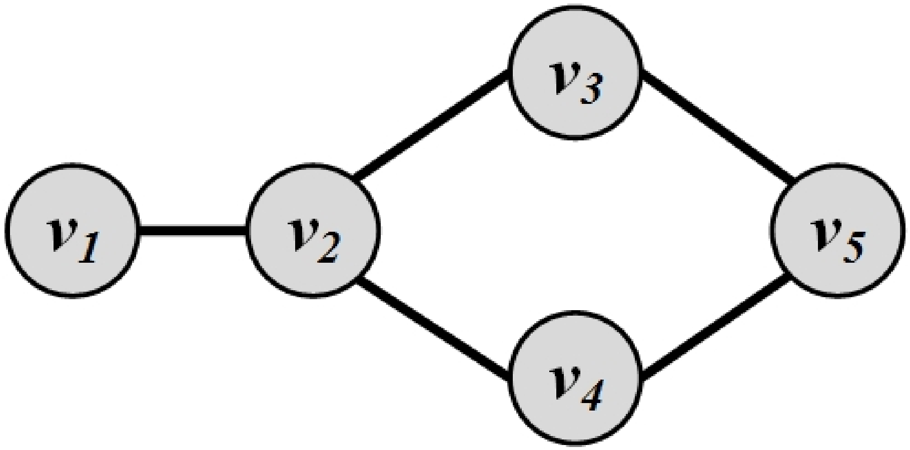
the number of shortest paths Connecting node j and node k

- It can be normalized by

$$\hat{C}_B = \frac{C_B(i)}{\boxed{(N-1)(N-2)}}$$

number of pairs of nodes excluding the node itself

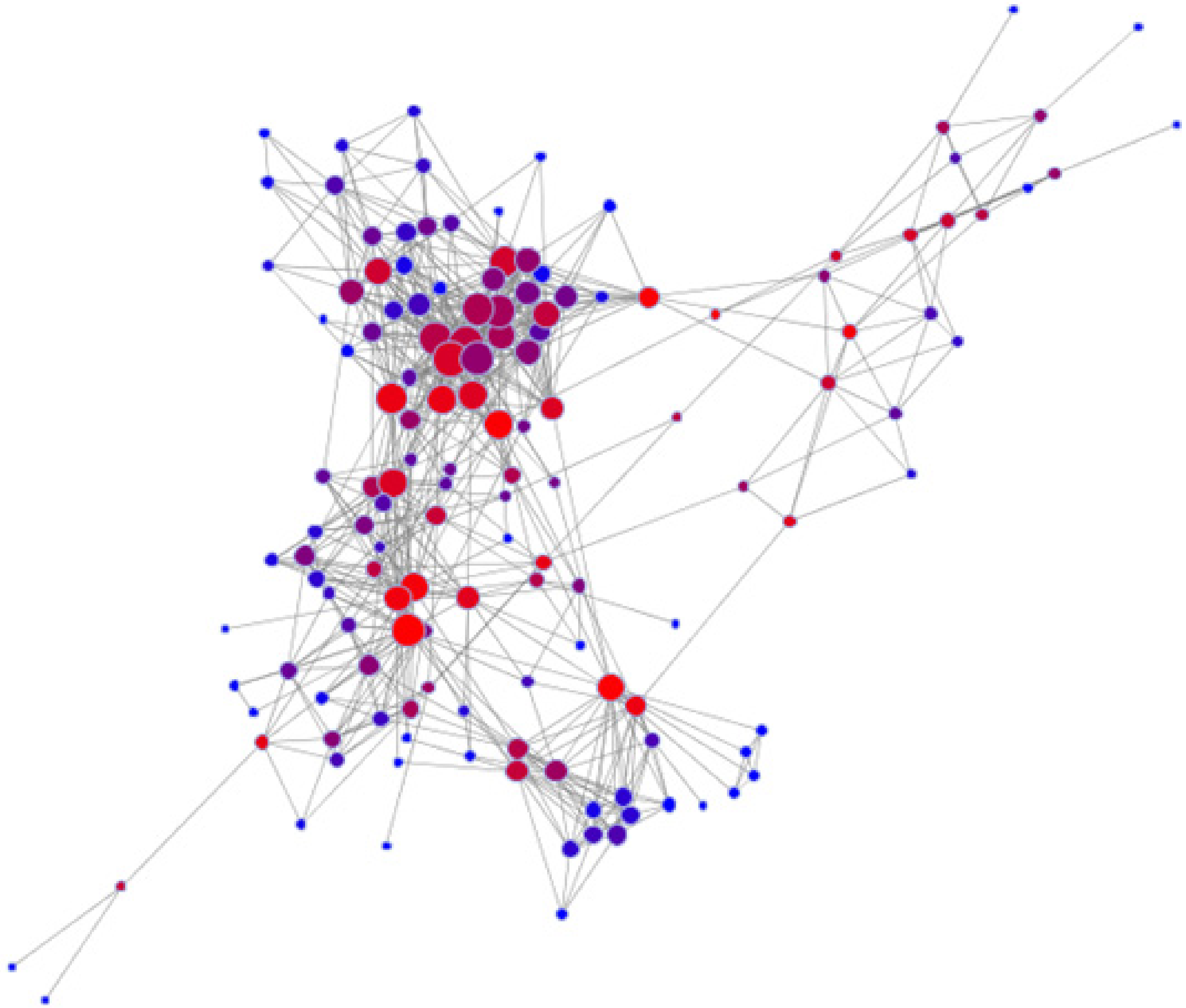
Example (1/2)



$$C_b(v_2) = 2 \times (\underbrace{(1/1)}_{s=v_1,t=v_3} + \underbrace{(1/1)}_{s=v_1,t=v_4} + \underbrace{(2/2)}_{s=v_1,t=v_5} + \underbrace{(1/2)}_{s=v_3,t=v_4} + \underbrace{0}_{s=v_3,t=v_5} + \underbrace{0}_{s=v_4,t=v_5})$$
$$= 2 \times 3.5 = 7,$$
$$C_b(v_3) = 2 \times (\underbrace{0}_{s=v_1,t=v_2} + \underbrace{0}_{s=v_1,t=v_4} + \underbrace{(1/2)}_{s=v_1,t=v_5} + \underbrace{0}_{s=v_2,t=v_4} + \underbrace{(1/2)}_{s=v_2,t=v_5} + \underbrace{0}_{s=v_4,t=v_5})$$
$$= 2 \times 1.0 = 2,$$
$$C_b(v_4) = C_b(v_3) = 2 \times 1.0 = 2,$$
$$C_b(v_5) = 2 \times (\underbrace{0}_{s=v_1,t=v_2} + \underbrace{0}_{s=v_1,t=v_3} + \underbrace{0}_{s=v_1,t=v_4} + \underbrace{0}_{s=v_2,t=v_3} + \underbrace{0}_{s=v_2,t=v_4} + \underbrace{(1/2)}_{s=v_3,t=v_4})$$
$$= 2 \times 0.5 = 1,$$

Example (2/2)

- Nodes are sized by degree, colored by betweenness
  - What about high degree but relatively low betweenness?

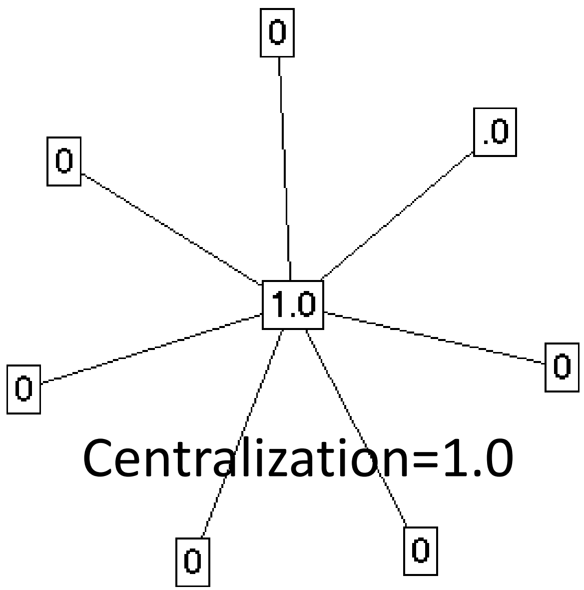


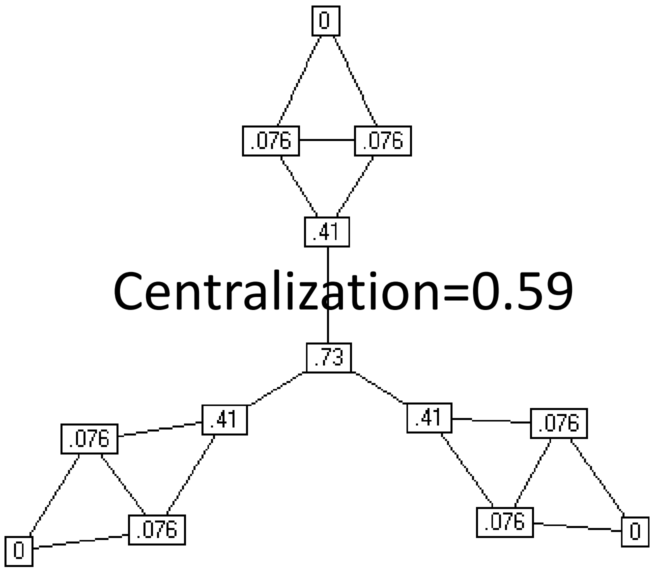
Betweenness Centralization

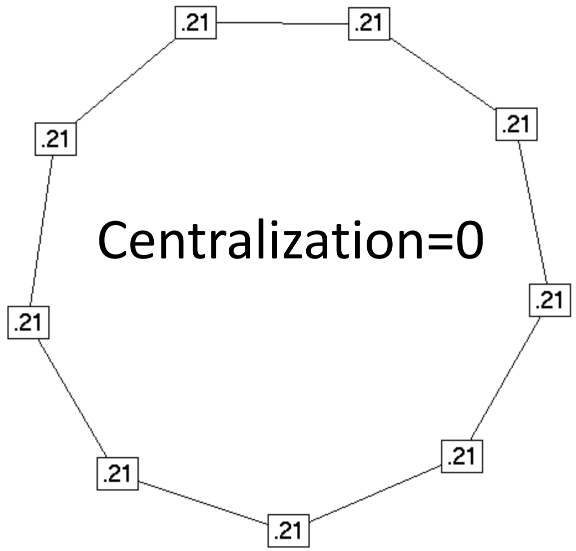
- Simple variance of the individual centrality

$$S_B^2 = \frac{\sum_{i=1}^{N}(C_B(v_i) - \boxed{\bar{C}_B})^2}{N}$$

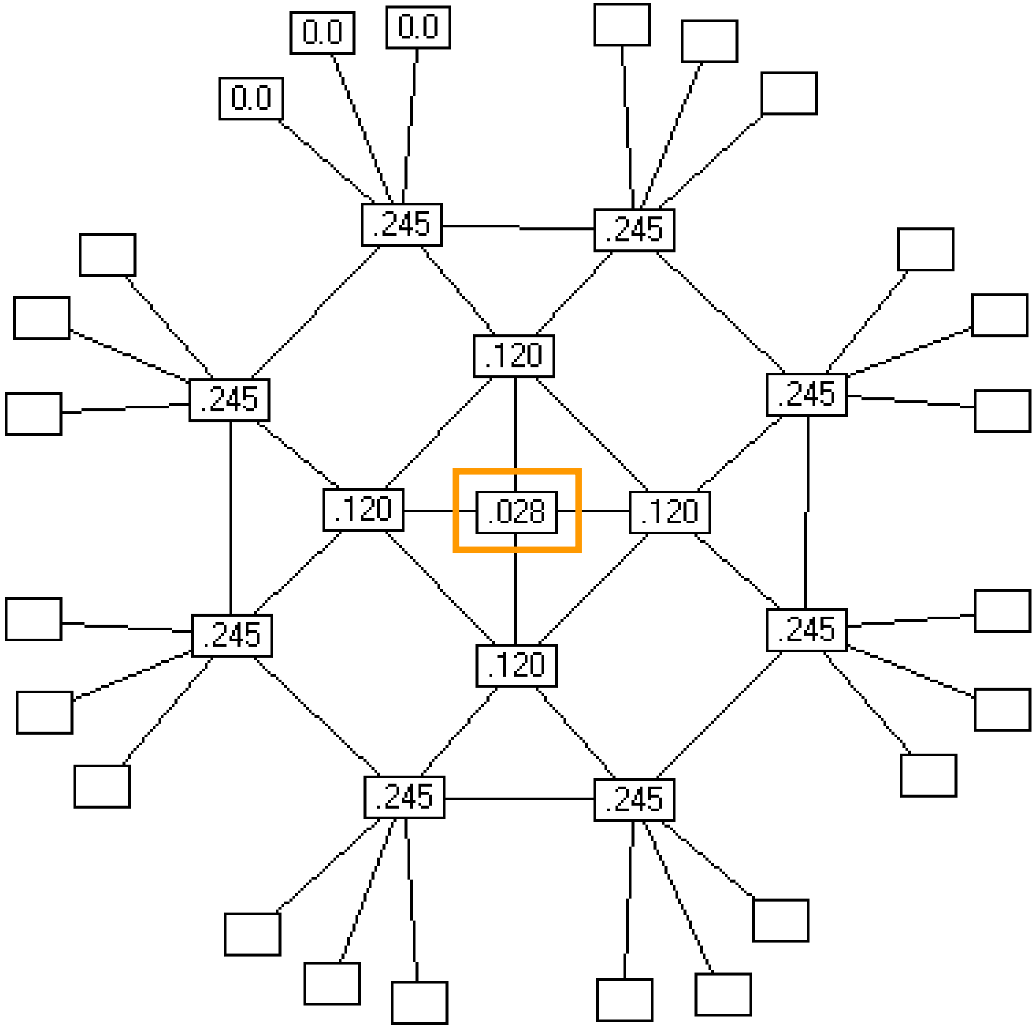
average betweenness







Between Centrality and Centralization



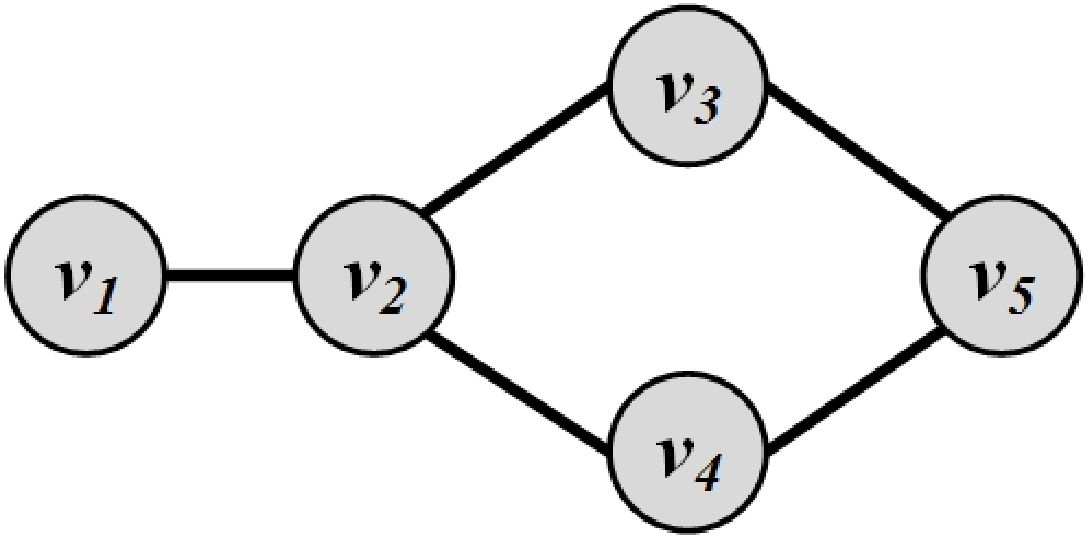
Centralization=0.183

Closeness Centrality

- Motivation
  - What if it's not so important to have many direct friends or to be "between" others?
  - But one still wants to be in the "middle" of things, not too far from the center
- Closeness is based on the length of average shortest path between a node and all nodes in the graph
  - The intuition is that influential and central nodes can quickly reach other nodes

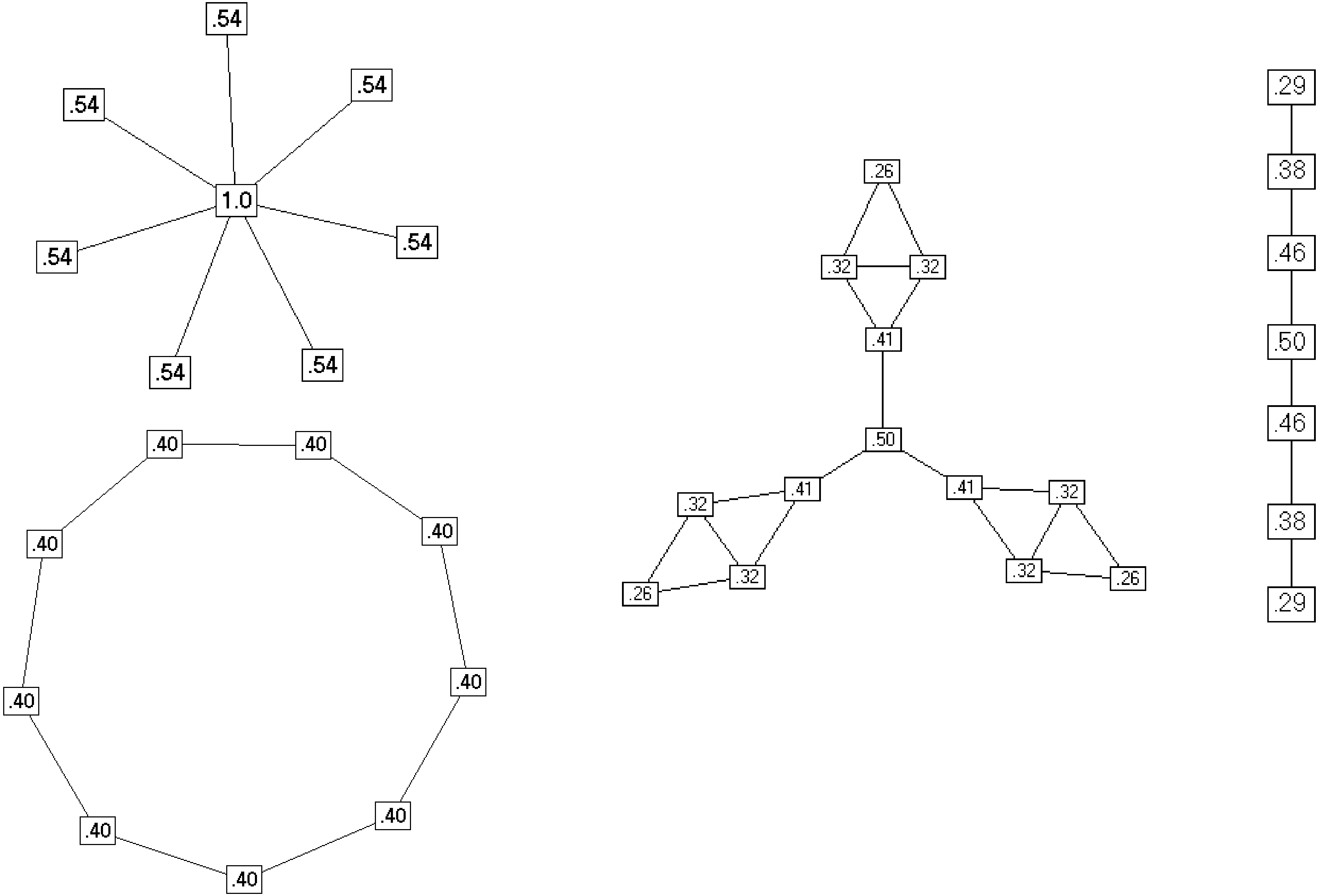
$$C_C(i) = \frac{1}{\bar{l}(i)} \qquad \bar{l}(i) = \frac{1}{N-1}\sum_{j \neq i} l(i,j)$$

Example (1/3)



$$
C_c(v_3) = \begin{aligned}
C_c(v_1) &= 1\ /\ (\ (1+2+2+3)/4\ ) = 0.5, \\
C_c(v_2) &= 1\ /\ (\ (1+1+1+2)/4\ ) = 0.8, \\
C_b(v_4) &= 1\ /\ (\ (1+1+2+2)/4\ ) = 0.66, \\
C_c(v_5) &= 1\ /\ (\ (1+1+2+3)/4\ ) = 0.57.
\end{aligned}
$$

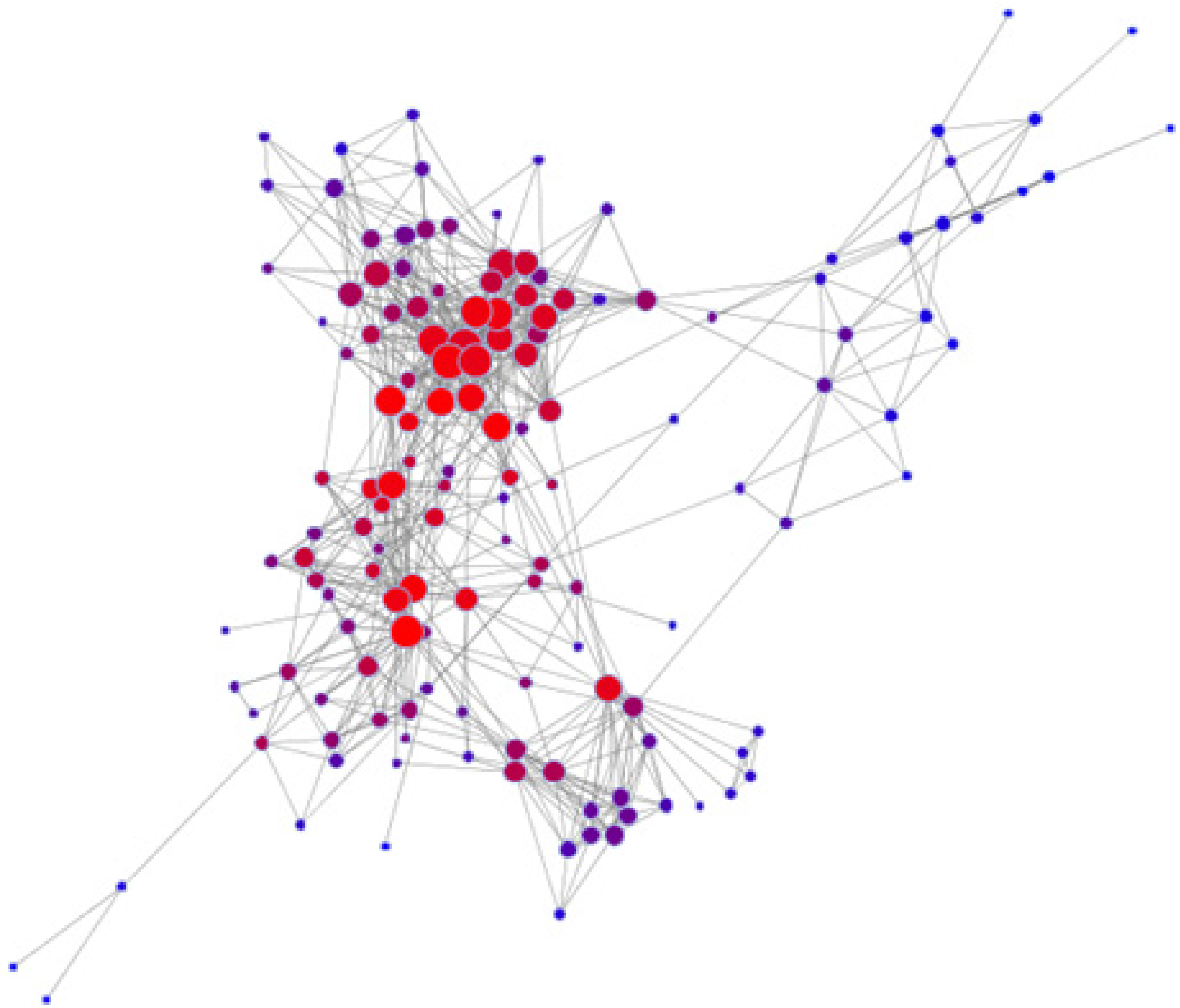
Example (2/3)



.54
.54
.54
1.0
.54
.54
.54
.54
.40
.40
.40
.40
.40
.40
.40
.40
.40
.40
.26
.32
.32
.41
.50
.41
.32
.41
.32
.26
.32
.32
.26
.29
.38
.46
.50
.46
.38
.29

Example (3/3)

- Nodes are sized by degree, colored by closeness
  - What about high degree but relatively low closeness?



Short Summary

- Generally different centrality metrics will be **positively correlated**
  - When they are not, then there is likely 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 偏遠領袖	A node's connections are redundant - communication bypasses him/her 小眾領袖
High Closeness	Key player tied to important or active alters 上位者		Probably multiple paths in the network, ego is near many people, but so are many others 主流平民
High Between	Ego's few ties are crucial for network flow 橋梁	Very rare cell, Would mean that ego monopolizes the ties from a small number of people to many others 偏遠橋梁	

Eigenvector Centrality

- Intuition: A node is more central
  if it is connected to some central nodes
- Having more friends does not by itself
  guarantee that someone is more important
  - Having more important friends
    provides a stronger signal
- Eigenvector centrality generalizes degree centrality by incorporating the importance of the neighbors in undirected graphs
  - In directed graphs,
    we can use incoming and outgoing edges

Formulation of Eigenvector Centrality

- Let $c_e(v_i)$ the eigenvector centrality of node v_i (unknown)
- We want $c_e(v_i)$ to be higher when important neighbors (node v_j with higher $c_e(v_j)$) point to v_i
  - Incoming or outgoing neighbors?
  - For incoming neighbors $A_{j,i} = 1$
- Assume that v_i's centrality is the summation of its neighbors' centralities

$$c_e(v_i) = \sum_{j=1}^{n} A_{j,i} c_e(v_j)$$

- Is this summation bounded?
  - We have to normalize!
  - λ: some fixed constant

$$c_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{j,i} c_e(v_j)$$

Eigenvector Centrality via Matrix Formulation

- Let $\boldsymbol{C}_e = \left(c_e(v_1), c_e(v_2), \dots, c_e(v_n)\right)^T$

$$c_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{j,i} c_e(v_j)$$

$$\rightarrow \lambda \boldsymbol{C}_e = A^T \boldsymbol{C}$$

- This means that $\boldsymbol{C}_e$ is an eigenvector of adjacency matrix A^T (or A when undirected) and λ is the corresponding eigenvalue
- Which eigenvalue-eigenvector pair should we choose?

Eigenvector Centrality (cont.)

Theorem 1 (Perron-Frobenius Theorem). *Let $A \in \mathbb{R}^{n \times n}$ represent the adjacency matrix for a [strongly] connected graph or $A : A_{i,j} > 0$ (i.e. a positive n by n matrix). There exists a positive real number (Perron-Frobenius eigenvalue) $\lambda_{\max}$, such that $\lambda_{\max}$ is an eigenvalue of A and any other eigenvalue of A is strictly smaller than $\lambda_{\max}$. Furthermore, there exists a corresponding eigenvector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ of A with eigenvalue $\lambda_{\max}$ such that $\forall v_i > 0$.*

To compute eigenvector centrality of A

1) We compute the eigenvalues of A
2) Select the largest eigenvalue λ
3) The corresponding eigenvector of λ is $\boldsymbol{C_e}$
4) Based on the Perron-Frobenius theorem, all the components of $\boldsymbol{C_e}$ will be positive
5) $\boldsymbol{C_e}$ components are the eigenvector centrality values for all nodes in the graph

Example

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

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

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

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

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

Eigenvalues are

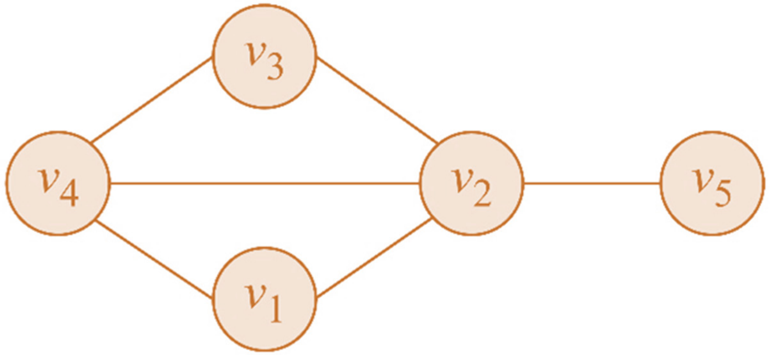
$$(-\sqrt{2}, 0, \boxed{+\sqrt{2}})$$

Largest Eigenvalue

Corresponding eigenvector (assuming $\mathbf{C_e}$ has norm 1)

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

Example

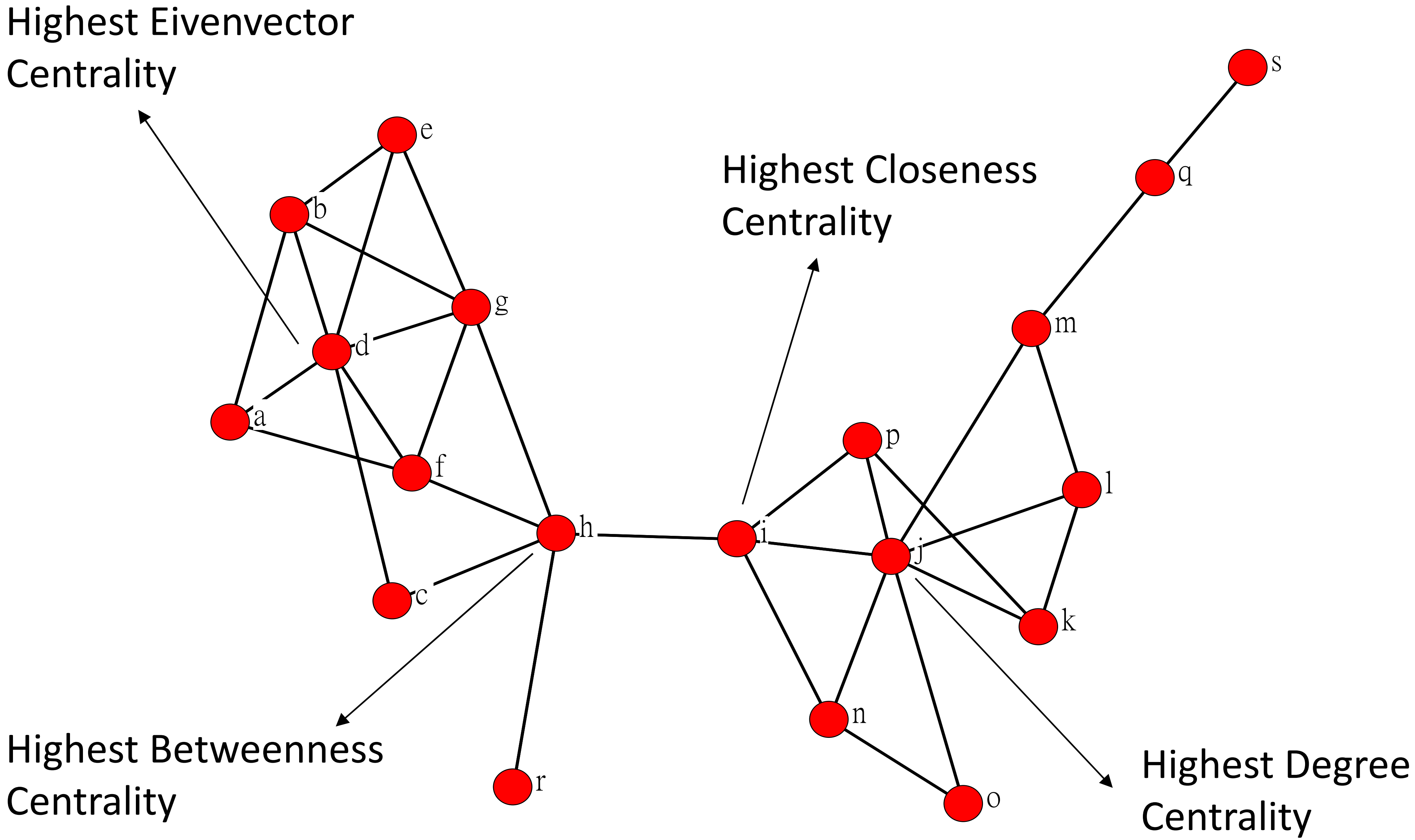


$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \lambda = (2.68,\ -1.74,\ -1.27,\ 0.33,\ 0.00)$$

Eigenvalues Vector

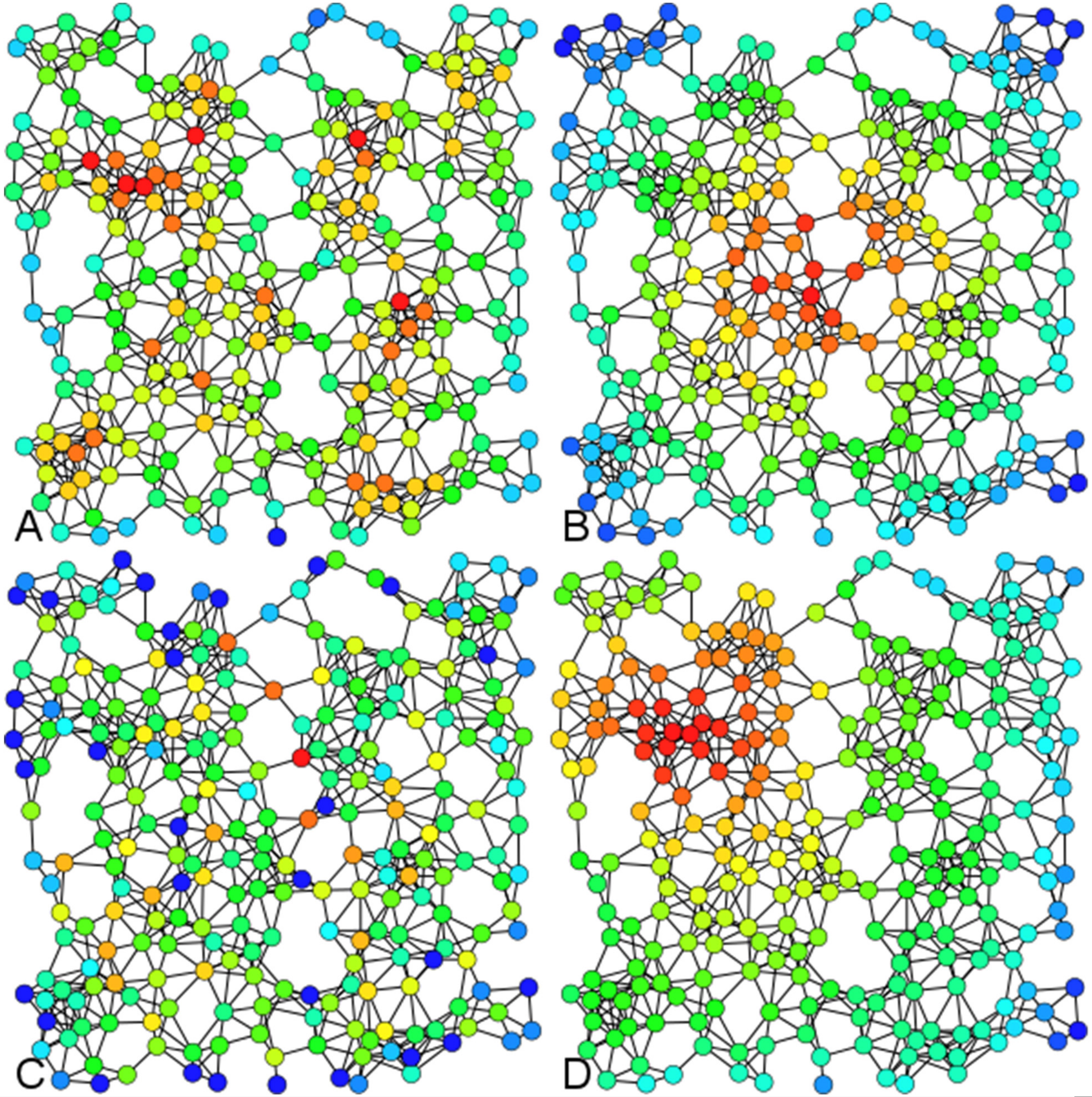
$$\lambda_{max} = 2.68 \Rightarrow C_e = \begin{bmatrix} 0.4119 \\ 0.5825 \\ 0.4119 \\ 0.5237 \\ 0.2169 \end{bmatrix}$$

Comparison of Centrality Measures



Comparison of Centrality Measures

A) Degree centrality
B) Closeness centrality
C) Betweenness centrality
D) Eigenvector centrality



https://en.wikipedia.org/wiki/Centrality

Network Centrality

- We want to know which nodes are the most "central" or important in a social network
- **Centrality Measures**
  - Local: Degree
  - Global: Closeness, Betweenness, Eigenvector
- **Centralization**
  - How evenly is centrality distributed among nodes?
- Centrality scores can be considered as features for node classification