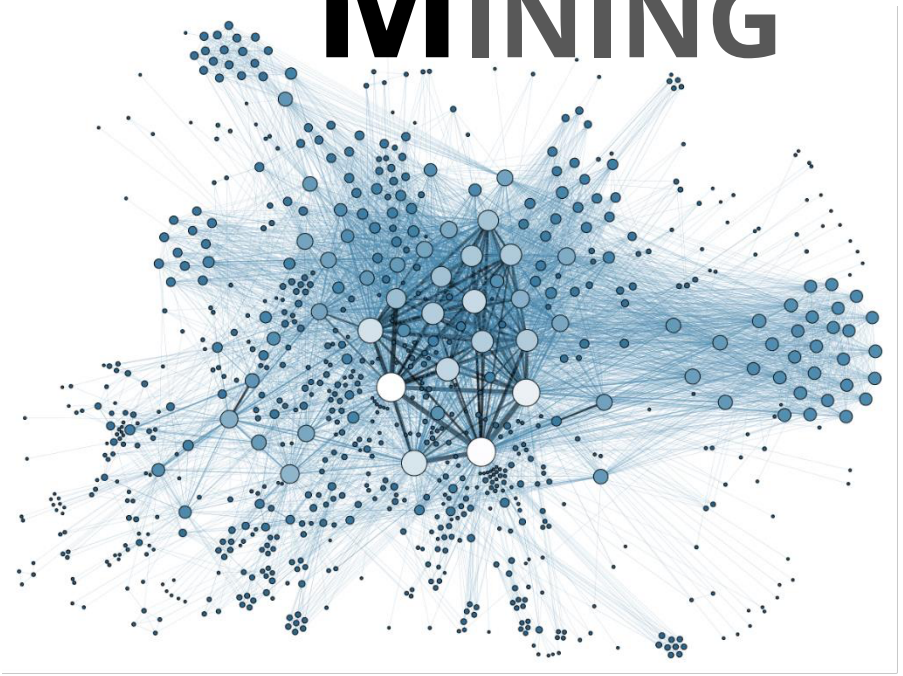
Social Media Mining

Network Measures



Dear instructors/users of these slides:

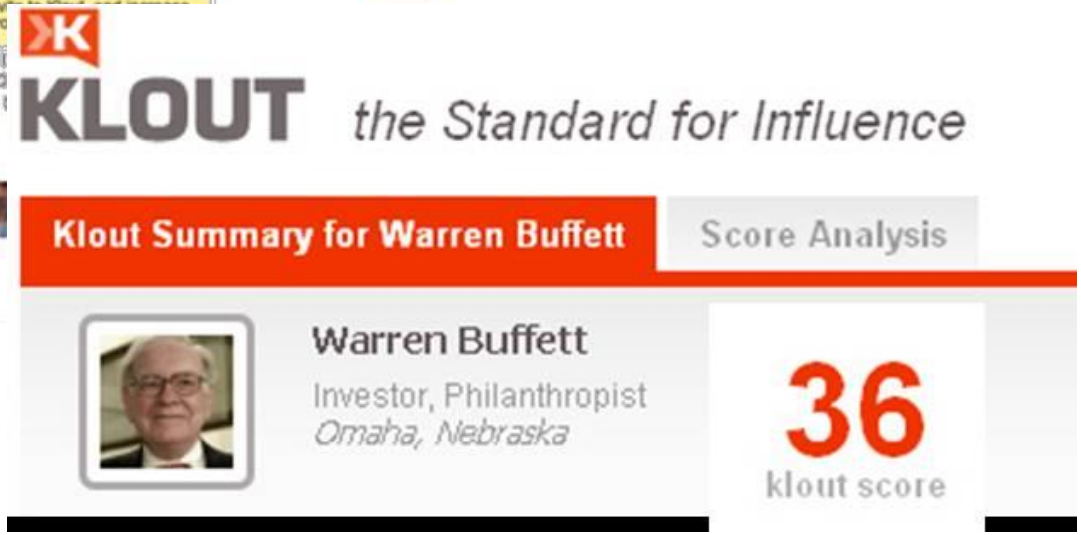
Please feel free to include these slides in your own material, or modify them as you see fit. If you decide to incorporate these slides into your presentations, please include the following note:

R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*, Cambridge University Press, 2014.
Free book and slides at **http://socialmediamining.info/**

or include a link to the website:
http://socialmediamining.info/

Klout

It is difficult to measure influence!



Why Do We Need Measures?

- Who are the central figures (influential individuals) in the network?
  - **Centrality**
- What interaction patterns are common in friends?
  - **Reciprocity and Transitivity**
  - **Balance and Status**
- Who are the like-minded users and how can we find these similar individuals?
  - **Similarity**
- To answer these and similar questions, one first needs to define measures for quantifying **centrality**, **level of interactions**, and **similarity**, among others.

Centrality

Centrality defines how important a node is within a network

Centrality in terms of those who you are connected to

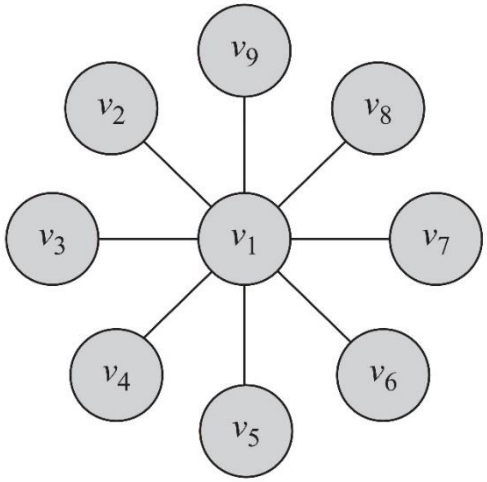
Degree Centrality

- **Degree centrality**: ranks nodes with more connections higher in terms of centrality

$$C_d(v_i) = d_i$$

- d_i is the degree (number of friends) for node v_i
  - i.e., the number of length-1 paths (can be generalized)

In this graph, degree centrality for node v_1 is d_1=8 and for all others is $d_j = 1, j \neq 1$



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.

$$C_d(v_i) = d_i^{\text{in}} \quad (prestige)$$
$$C_d(v_i) = d_i^{\text{out}} \quad (gregariousness)$$
$$C_d(v_i) = d_i^{\text{in}} + d_i^{\text{out}}$$

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

Normalized Degree Centrality

- Normalized by the maximum possible degree

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

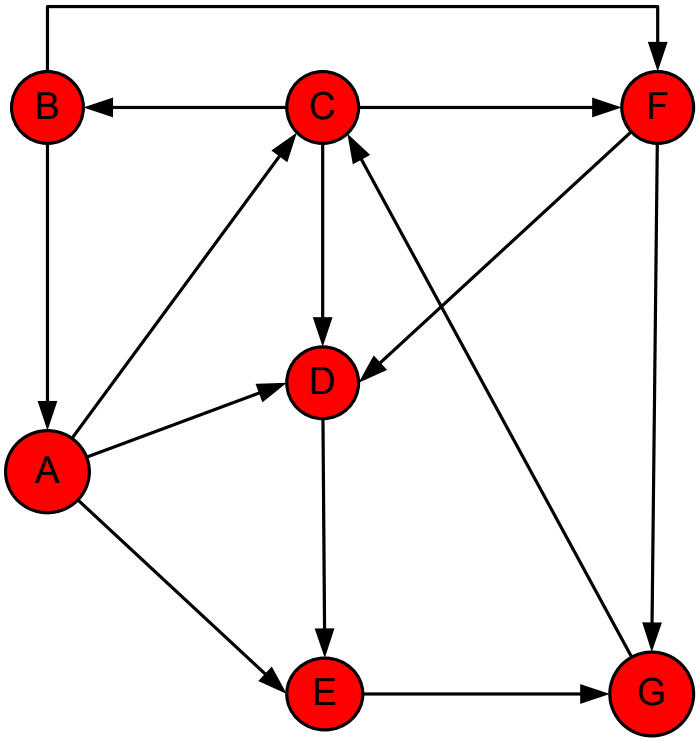
- Normalized by the maximum degree

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

- 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 (Directed Graph)Example

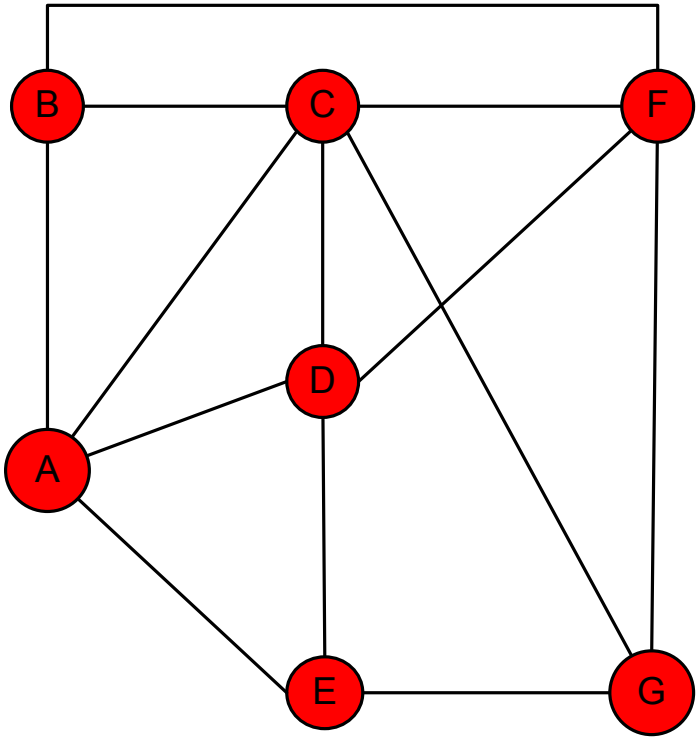


Node	In-Degree	Out-Degree	Centrality	Rank
A	1	3	1/2	**1**
B	1	2	1/3	**3**
C	2	3	1/2	**1**
D	3	1	1/6	**5**
E	2	1	1/6	**5**
F	2	2	1/3	**3**
G	2	1	1/6	**5**

Normalized by the maximum possible degree

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

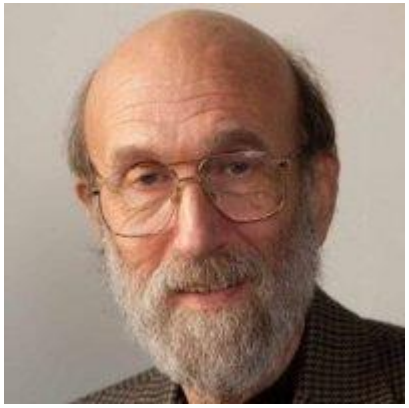
Degree Centrality (undirected Graph) Example



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**

Eigenvector Centrality

- Having more friends does not by itself guarantee that someone is more important
  - Having more **important friends** provides a stronger signal



Phillip Bonacich

- Eigenvector centrality generalizes degree centrality by incorporating the importance of the neighbors (undirected)

- For directed graphs, we can use incoming or outgoing edges

Formulation

- Let's assume the eigenvector centrality of a node is $c_e(v_i)$ (unknown)

- We would like $c_e(v_i)$ to be higher when important neighbors (node v_j with higher $c_e(v_j)$) point to us
  - Incoming or outgoing neighbors?
  - For incoming neighbors $A_{j,i} = 1$

- We can 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 (Matrix Formulation)

- Let $\mathbf{C}_e = (C_e(v_1), C_e(v_2), \ldots, C_e(v_n))^T$

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

- 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?

Finding the eigenvalue by finding a fixed point…

- Start from an initial guess $C_e(0)$ (e.g., all centralities are 1) and iterative t times

$$C_e(t) = (A^T)^t C_e(0)$$

- We can write $C_e(0)$ as a linear combination of eigenvectors v_i's of the A^T

$$C_e(0) = \sum_i \alpha_i v_i$$

- Substituting this, we get

$$C_e(t) = (A^T)^t \sum_i \alpha_i v_i = \sum_i \alpha_i {\lambda_i}^t v_i = {\lambda_1}^t \sum_i \alpha_i (\frac{\lambda_i}{\lambda_1})^t v_i$$

λ_1 is the largest eigenvalue

Finding the eigenvalue by finding a fixed point…

- As t grows, we will have in the limit

$$C_e(t) \rightarrow \alpha_1 {\lambda_1}^t v_1$$

- Or equivalently

$$A^T C_e(t) = A^T C_e = \lambda_1 C_e$$

- If we start with an all positive $C_e(0)$ all $C_e(t)$'s will be positive (why?)
  - All the centrality values would be positive
  - We need an eigenvalue-eigenvector pair that guarantees all centralities have the same sign
    - E.g., for comparison purposes

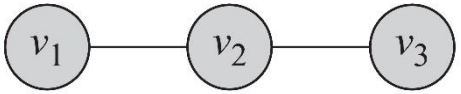
Eigenvector Centrality, cont.

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 $\forall v_i > 0$.*

So, 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 $\mathbf{C_e}$.
4. Based on the Perron-Frobenius theorem, all the components of $\mathbf{C_e}$will be positive
5. The components of $\mathbf{C_e}$ are the eigenvector centralities for the graph.

Eigenvector Centrality: Example 1



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

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

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

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

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

Eigenvalues are

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

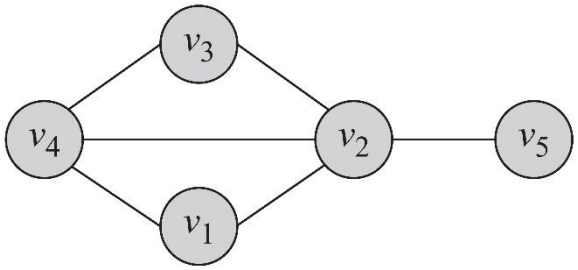
Largest Eigenvalue

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

$$\begin{bmatrix} 0-\sqrt{2} & 1 & 0 \\ 1 & 0-\sqrt{2} & 1 \\ 0 & 1 & 0-\sqrt{2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

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

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 Vector

$\lambda_{max} = 2.68 \Rightarrow$

$$C_e = \begin{bmatrix} 0.4119 \\ 0.5825 \\ 0.4119 \\ 0.5237 \\ 0.2169 \end{bmatrix}$$

Katz Centrality

- A major problem with eigenvector centrality arises when it deals with directed graphs
- Centrality only passes over *outgoing* edges and in special cases such as when a node is in a directed acyclic graph centrality becomes zero
  - The node can have many edge connected to it

Elihu Katz

- To resolve this problem we add bias term β to the centrality values for all nodes

Eigenvector Centrality

$$C_{\text{Katz}}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{\text{Katz}}(v_j) + \beta$$

Katz Centrality, cont.

$$C_{\text{Katz}}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{\text{Katz}}(v_j) + \beta$$

Controlling term

Bias term

Rewriting equation in a vector form

$$\mathbf{C}_{\text{Katz}} = \alpha A^T \mathbf{C}_{\text{Katz}} + \beta \mathbf{1}$$

vector of all 1's

Katz centrality: $\mathbf{C}_{\text{Katz}} = \beta(\mathbf{I} - \alpha A^T)^{-1} \cdot \mathbf{1}$

Katz Centrality, cont.

$$C_{\text{Katz}}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{\text{Katz}}(v_j) + \beta$$

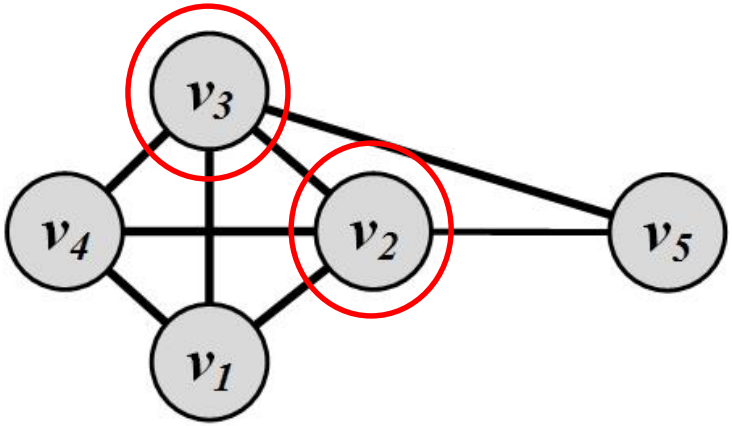
- When α=0, the eigenvector centrality is removed and all nodes get the same centrality value β
  - As α gets larger the effect of β is reduced

- For the matrix $(I - \alpha A^T)$ to be invertible, we must have
  - $det(I - \alpha A^T) \neq 0$
  - By rearranging we get $det(A^T - \alpha^{-1} I) = 0$
  - This is basically the characteristic equation,
  - The characteristic equation **first** becomes zero when the largest eigenvalue equals α^{-1}

The largest eigenvalue is easier to compute (power method)

In practice we select $\boldsymbol{\alpha < 1/\lambda}$, where λ is the largest eigenvalue of $\boldsymbol{A^T}$

Katz Centrality Example



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

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

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

Most important nodes!

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, cont.

$$C_p(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} \frac{C_p(v_j)}{d_j^{\text{out}}} + \beta$$

What if the degree is zero?

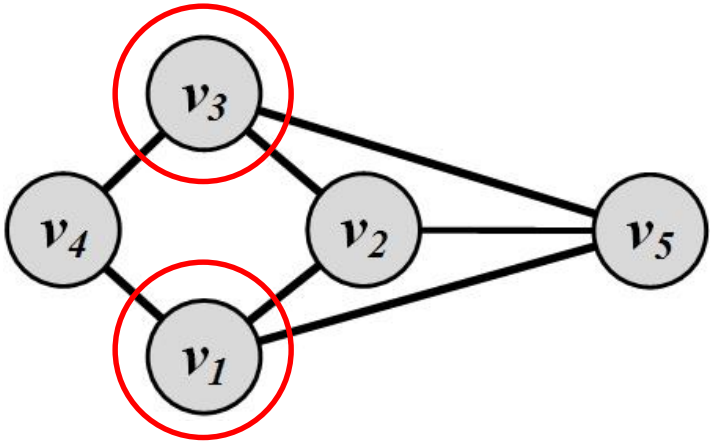
$$\begin{cases} d_j^{out} > 0 \\ D = diag(d_1^{\text{out}}, d_2^{\text{out}}, \ldots, d_n^{\text{out}}) \end{cases} \Rightarrow \mathbf{C}_p = \alpha A^T D^{-1} \mathbf{C}_p + \beta \mathbf{1}$$

$$\mathbf{C}_p = \beta(\mathbf{I} - \alpha A^T D^{-1})^{-1} \cdot \mathbf{1}$$

Similar to Katz Centrality, in practice, $\boldsymbol{\alpha < 1/\lambda}$, where λ is the largest eigenvalue of $A^T D^{-1}$. In undirected graphs, the largest eigenvalue of $A^T D^{-1}$ is $\boldsymbol{\lambda = 1}$; therefore, $\boldsymbol{\alpha < 1}$.

PageRank Example

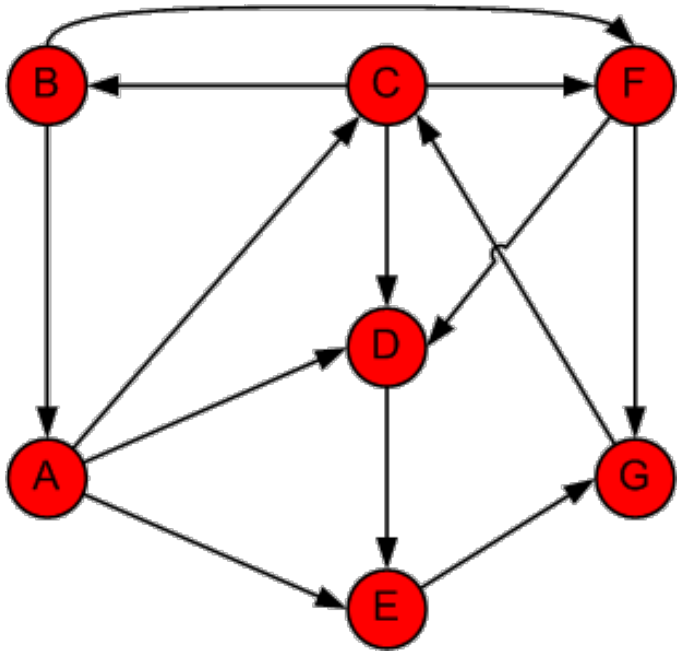
- We assume α=0.95 < 1 and and $\beta = 0.1$



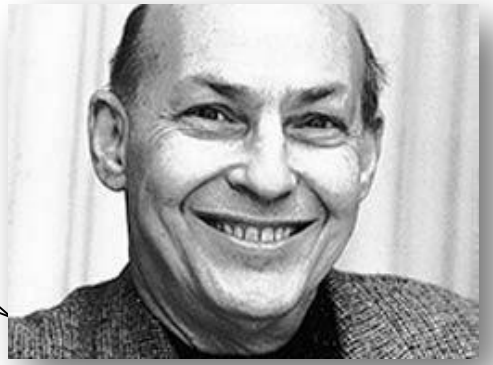
$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{C}_p = \beta(\mathbf{I} - \alpha A^T D^{-1})^{-1} \cdot \mathbf{1} = \begin{bmatrix} 2.14 \\ 2.13 \\ 2.14 \\ 1.45 \\ 2.13 \end{bmatrix}$$

PageRank Example – Alternative Approach [Markov Chains]



"You don't understand anything until you learn it more than one way"



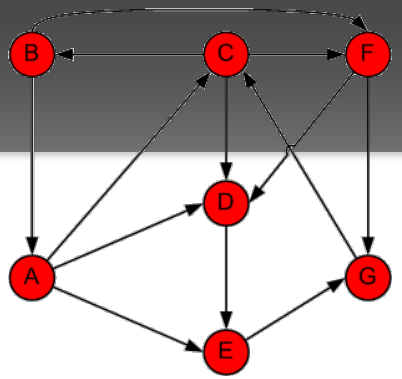
Marvin Minsky (1927-2016)

Using Power Method

α=1 and β =0?

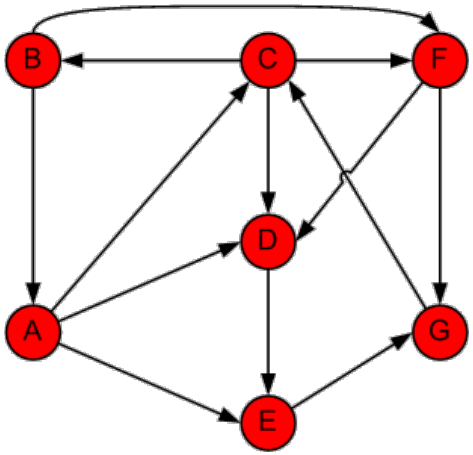
Step	A	B	C	D	E	F	G
0	1/7	1/7	1/7	1/7	1/7	1/7	1/7
1	B/2	C/3	A/3 + G	A/3 + C/3 + F/2	A/3 + D	C/3 + B/2	F/2 + E
	0.071	0.048	0.190	0.167	0.190	0.119	0.214

PageRank: Example



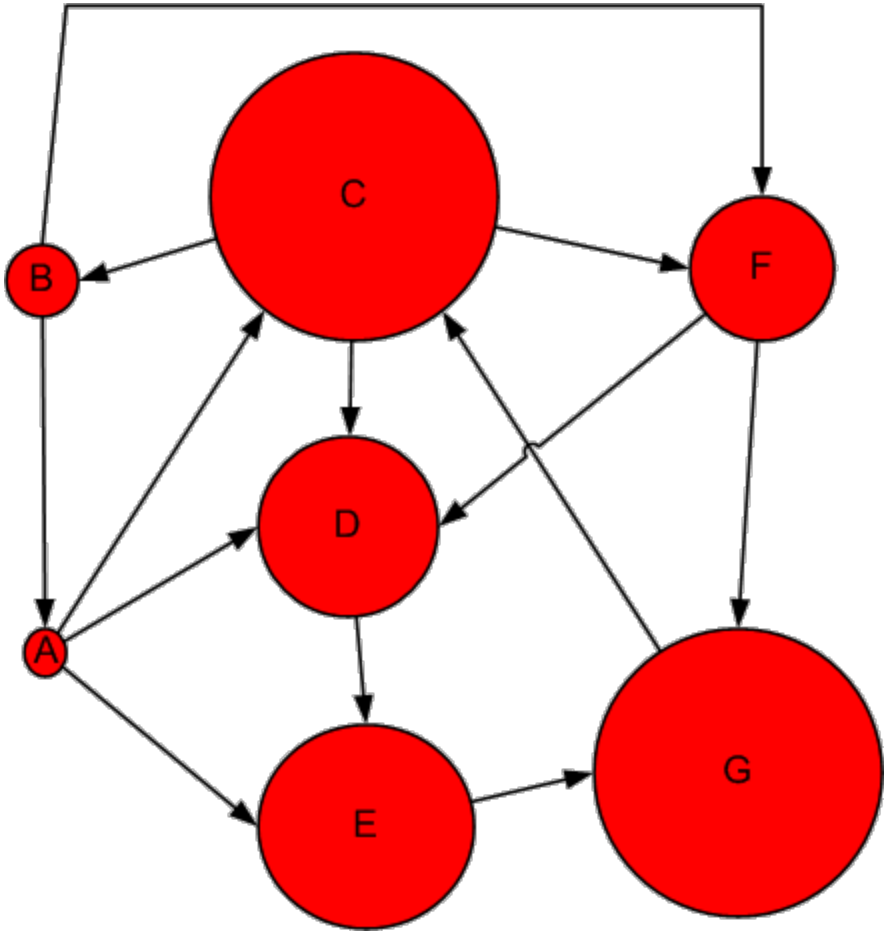
Step	A	B	C	D	E	F	G	Sum
1	0.143	0.143	0.143	0.143	0.143	0.143	0.143	1.000
2	0.071	0.048	0.190	0.167	0.190	0.119	0.214	1.000
3	0.024	0.063	0.238	0.147	0.190	0.087	0.250	1.000
4	0.032	0.079	0.258	0.131	0.155	0.111	0.234	1.000
5	0.040	0.086	0.245	0.152	0.142	0.126	0.210	1.000
6	0.043	0.082	0.224	0.158	0.165	0.125	0.204	1.000
7	0.041	0.075	0.219	0.151	0.172	0.115	0.228	1.000
8	0.037	0.073	0.241	0.144	0.165	0.110	0.230	1.000
9	0.036	0.080	0.242	0.148	0.157	0.117	0.220	1.000
10	0.040	0.081	0.232	0.151	0.160	0.121	0.215	1.000
11	0.040	0.077	0.228	0.151	0.165	0.118	0.220	1.000
12	0.039	0.076	0.234	0.148	0.165	0.115	0.223	1.000
13	0.038	0.078	0.236	0.148	0.161	0.116	0.222	1.000
14	0.039	0.079	0.235	0.149	0.161	0.118	0.219	1.000
15	0.039	0.078	0.232	0.150	0.162	0.118	0.220	1.000
Rank	**7**	**6**	**1**	**4**	**3**	**5**	**2**	

Effect of PageRank



PageRank

Node	Rank
A	**7**
B	**6**
C	**1**
D	**4**
E	**3**
F	**5**
G	**2**

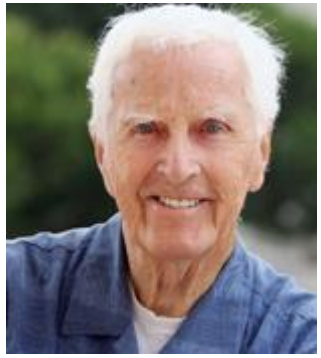


Centrality in terms of how you connect others

(information broker)

Betweenness Centrality

Another way of looking at centrality is by considering how important nodes are in connecting other nodes



Linton Freeman

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

σ_{st} The number of shortest paths from vertex s to t – *a.k.a.* information pathways

$\sigma_{st}(v_i)$ The number of **shortest paths** from s to t that pass through v_i

Normalizing Betweenness Centrality

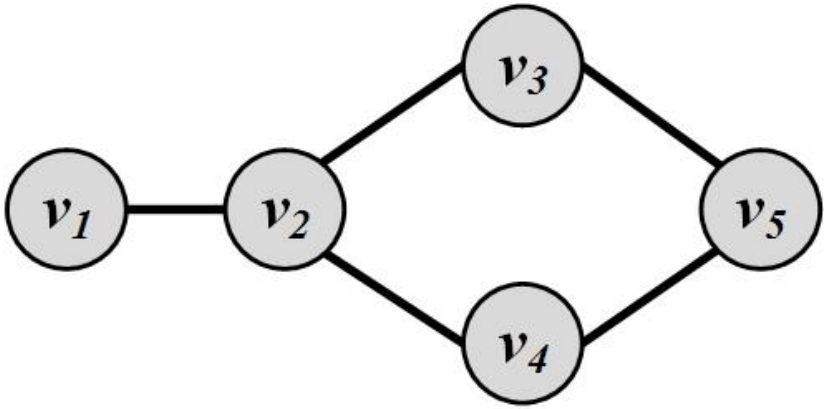
- In the best case, node v_i is on all shortest paths from s to t, hence, $\frac{\sigma_{st}(v_i)}{\sigma_{st}} = 1$

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$
$$= \sum_{s \neq t \neq v_i} 1 = 2\binom{n-1}{2} = (n-1)(n-2)$$

Therefore, the maximum value is $(n-1)(n-2)$

Betweenness centrality: $C_b^{\text{norm}}(v_i) = \frac{C_b(v_i)}{2\binom{n-1}{2}}$

Betweenness Centrality: Example 1



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

$$= 2 \times 3.5 = 7,$$

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

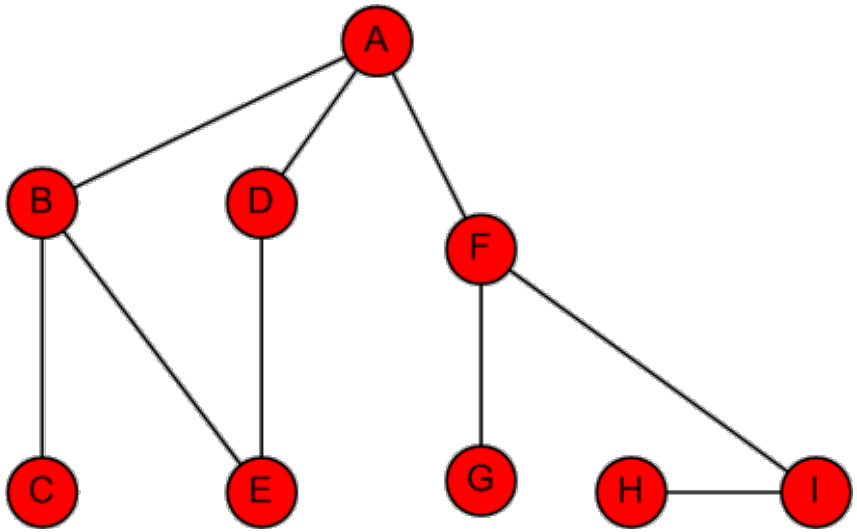
$$= 2 \times 1.0 = 2,$$

$$C_b(v_4) = C_b(v_3) = 2 \times 1.0 = 2,$$

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

$$= 2 \times 0.5 = 1,$$

Betweenness Centrality: Example 2



Node	Betweenness Centrality	Rank
A	16 + 1/2 + 1/2	**1**
B	7+5/2	**3**
C	0	**7**
D	5/2	**5**
E	1/2 + 1/2	**6**
F	15 + 2	**1**
G	0	**7**
H	0	**7**
I	7	**4**

Computing Betweenness

- In betweenness centrality, we compute shortest paths between all pairs of nodes to compute the betweenness value.

- **Trivial Solution**:
  - Use Dijkstra and run it $O(n)$ times
  - We get an $O(n^3)$ solution

- Better Solution:
  - Brandes Algorithm:
    - $O(nm)$ for unweighted graphs
    - $O(nm + n^2 \log n)$ for weighted graphs

Brandes Algorithm [2001]

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}} = \sum_{s \neq v_i} \delta_s(v_i)$$

$$\delta_s(v_i) = \sum_{t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

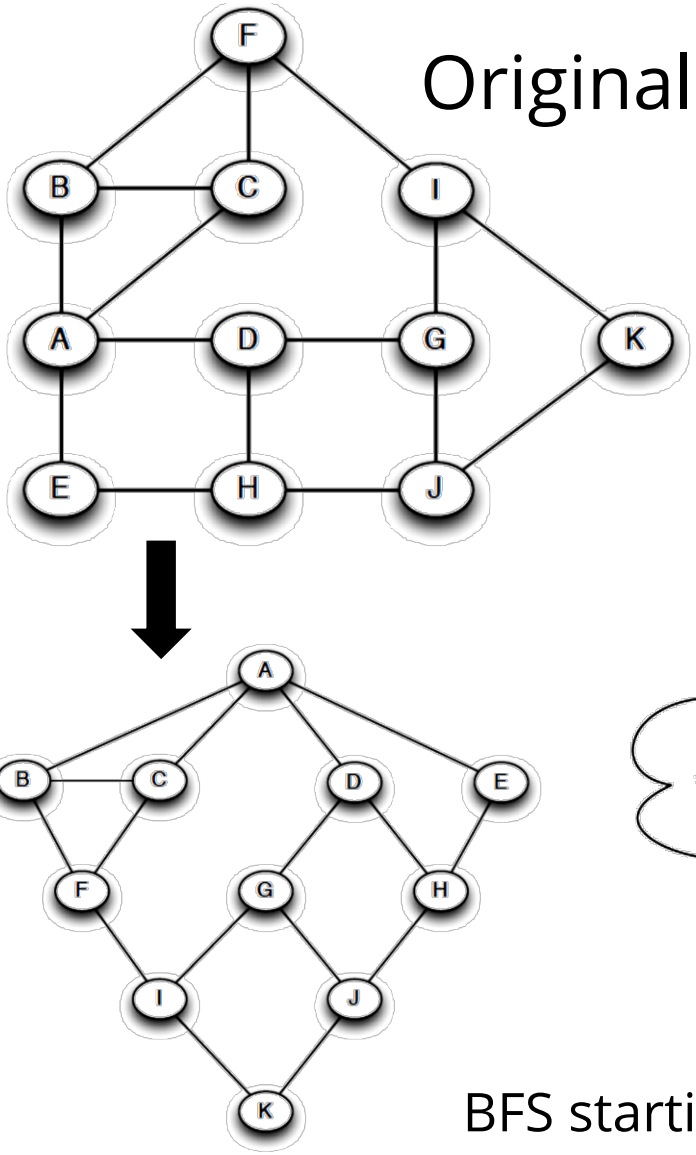
There exists a recurrence equation that can help us determine $\delta_s(v_i)$

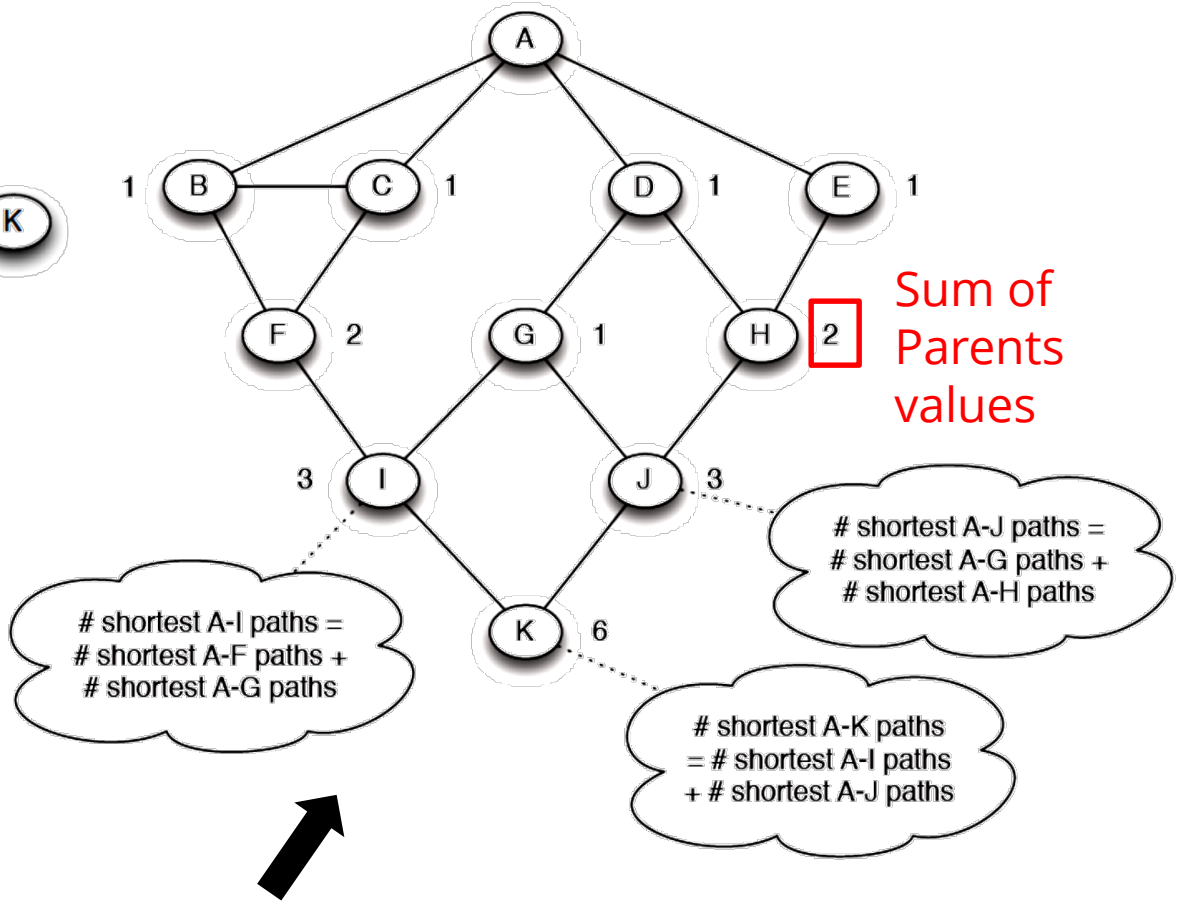
$$\delta_s(v_i) = \sum_{w:\ v_i \in pred(s,w)} \frac{\sigma_{sv_i}}{\sigma_{sw}}(1 + \delta_s(w))$$

$pred(s, w)$ is the set of predecessors of w in the shortest paths from s to w.

- In the most basic scenario, w is the immediate child of v_i

How to compute σ_{st}

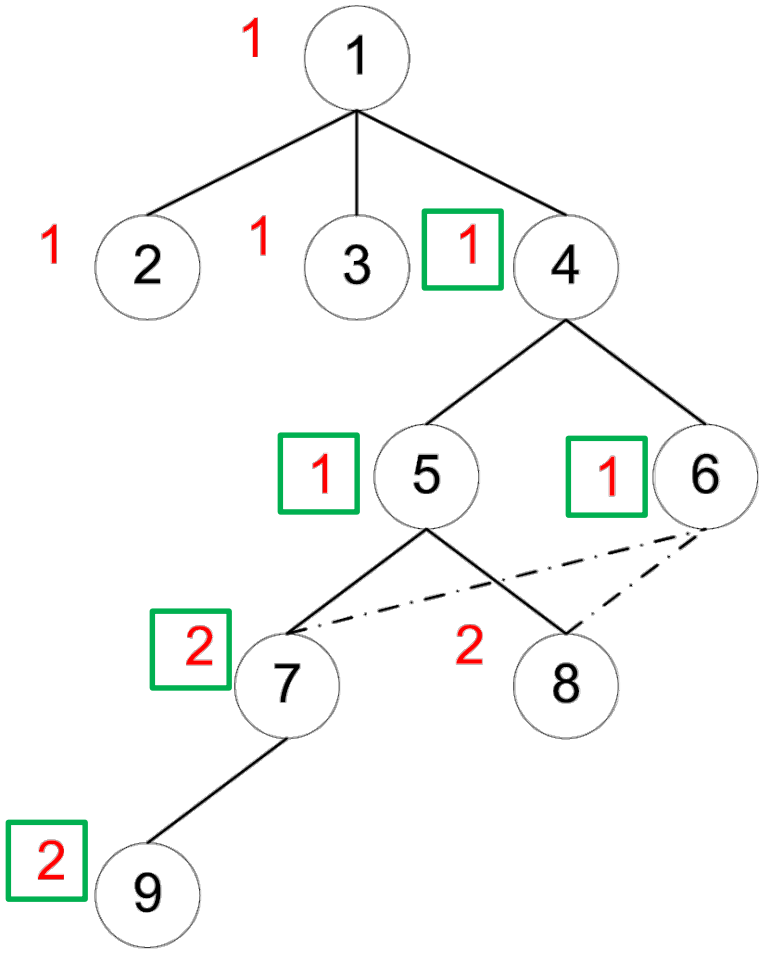


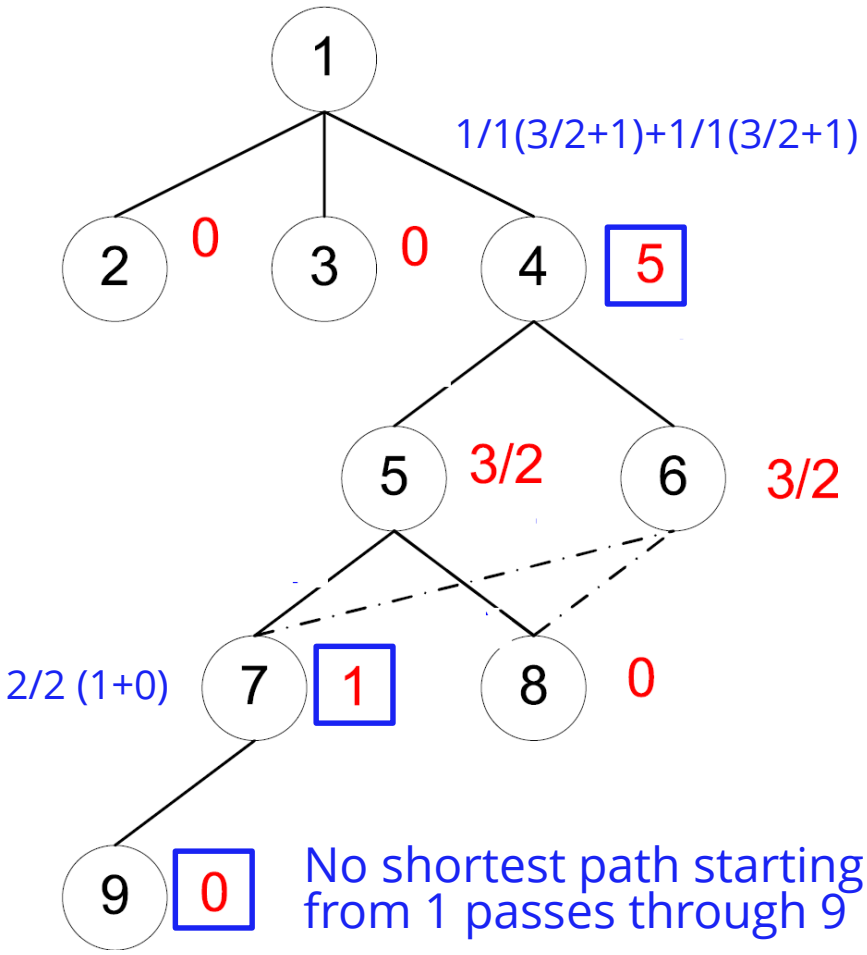


BFS starting at A (i.e., s)

Source: Networks, Crowds, and Markets: Reasoning about a Highly Connected World. By David Easley and Jon Kleinberg

How do you compute $\delta_s(v_i)$



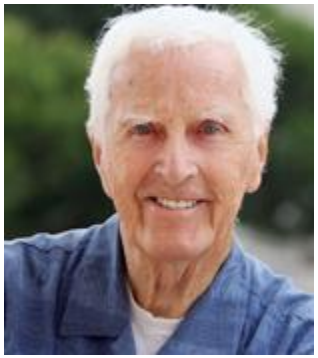


$$\delta_s(v_i) = \sum_{w:\ v_i \in pred(s,w)} \frac{\sigma_{sv_i}}{\sigma_{sw}} (1 + \delta_s(w))$$

Centrality in terms of how fast you can reach others

Closeness Centrality

- The intuition is that influential/central nodes can quickly reach other nodes
- These nodes should have a smaller average shortest path length to others

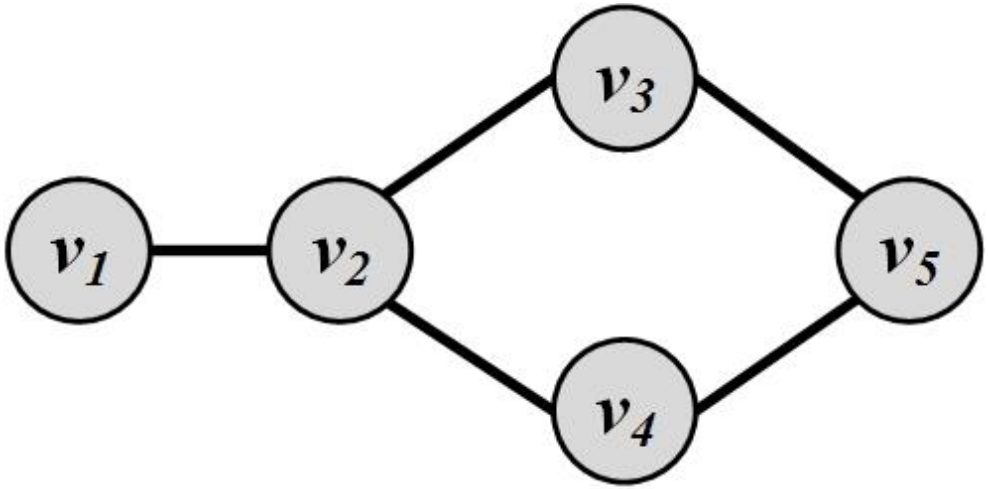


Linton Freeman

Closeness centrality: $C_c(v_i) = \frac{1}{\bar{l}_{v_i}}$

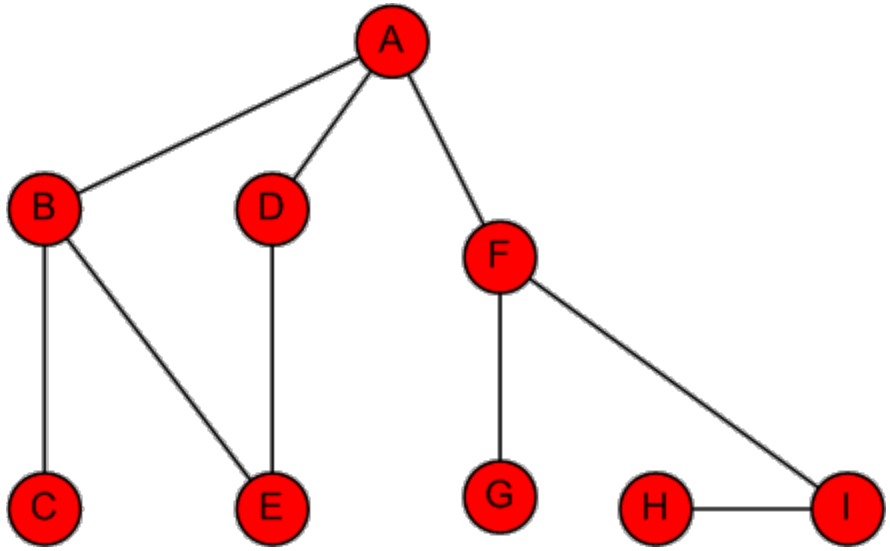
$$\bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

Closeness Centrality: Example 1



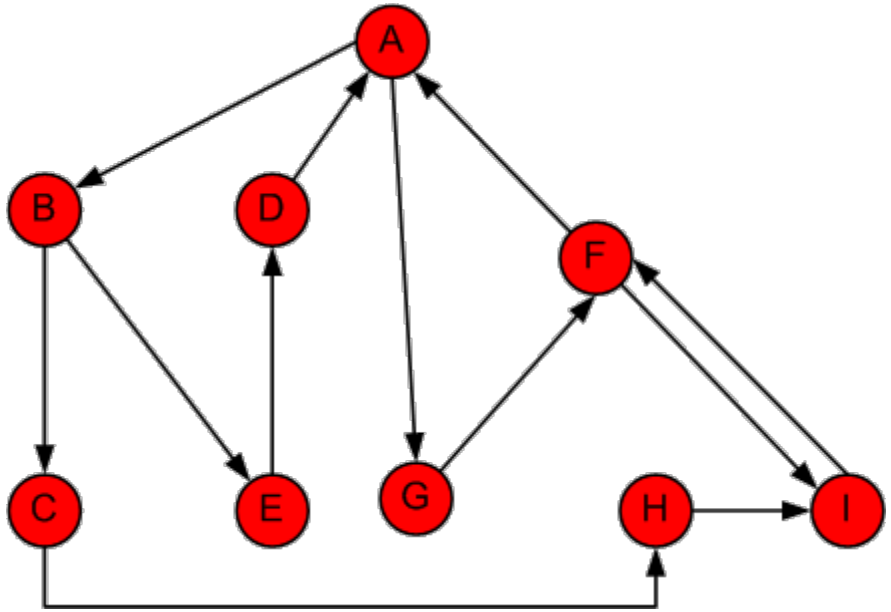
$$
\begin{aligned}
C_c(v_3) = \quad & C_c(v_1) = 1 \,/\, (\,(1+2+2+3)/4\,) = 0.5, \\
& C_c(v_2) = 1 \,/\, (\,(1+1+1+2)/4\,) = 0.8, \\
& C_b(v_4) = 1 \,/\, (\,(1+1+2+2)/4\,) = 0.66, \\
& C_c(v_5) = 1 \,/\, (\,(1+1+2+3)/4\,) = 0.57.
\end{aligned}
$$

Closeness Centrality: Example 2 (Undirected)



Node	A	B	C	D	E	F	G	H	I	D_Avg	Closeness Centrality	Rank
A	0	1	2	1	2	1	2	3	2	1.750	0.571	**1**
B	1	0	1	2	1	2	3	4	3	2.125	0.471	**3**
C	2	1	0	3	2	3	4	5	4	3.000	0.333	**8**
D	1	2	3	0	1	2	3	4	3	2.375	0.421	**4**
E	2	1	2	1	0	3	4	5	4	2.750	0.364	**7**
F	1	2	3	2	3	0	1	2	1	1.875	0.533	**2**
G	2	3	4	3	4	1	0	3	2	2.750	0.364	**7**
H	3	4	5	4	5	2	3	0	1	3.375	0.296	**9**
I	2	3	4	3	4	1	2	1	0	2.500	0.400	**5**

Closeness Centrality: Example 3 (Directed)



Node	A	B	C	D	E	F	G	H	I	D_Avg	Closeness Centrality	Rank
A	0	1	2	3	2	2	1	3	3	2.125	0.471	**1**
B	3	0	1	2	1	4	4	2	3	2.500	0.400	**2**
C	4	5	0	7	6	3	5	1	2	4.125	0.242	**9**
D	1	2	3	0	3	3	2	4	5	2.875	0.348	**3**
E	2	3	4	1	0	4	3	5	5	3.375	0.296	**6**
F	1	2	3	4	3	0	2	4	4	2.875	0.348	**4**
G	2	3	4	5	4	1	0	5	2	3.250	0.308	**5**
H	4	4	5	6	5	2	4	0	1	3.875	0.258	**8**
I	2	3	4	5	4	1	4	5	0	3.500	0.286	**7**

An Interesting Comparison!

Comparing three centrality values

- Generally, the 3 centrality types will be positively correlated
- When they are not (or low correlation), it usually reveals interesting information

	Low Degree	Low Closeness	Low Betweenness
High Degree		*Node is embedded in a community that is far from the rest of the network*	*Ego's connections are redundant - communication bypasses the node*
High Closeness	*Key node connected to important/active alters*		*Probably multiple paths in the network, ego is near many people, but so are many others*
High Betweenness	*Ego's few ties are crucial for network flow*	*Very rare! Ego monopolizes the ties from a small number of people to many others.*	

This slide is modified from a slide developed by James Moody

Centrality for a group of nodes

Group Centrality

- All centrality measures defined so far measure centrality for a single node. These measures can be generalized for a group of nodes.

- A simple approach is to replace all nodes in a group with a super node
  - The group structure is disregarded.

- Let S denote the set of nodes in the group and $V - S$ the set of outsiders

Group Centrality

I. Group Degree Centrality

$$C_d^{\text{group}}(S) = |\{v_i \in V - S | v_i \text{ is connected to } v_j \in S\}|$$

– **Normalization**: divide by $|V - S|$

II. Group Betweenness Centrality

$$C_b^{\text{group}}(S) = \sum_{s \neq t, s \notin S, t \notin S} \frac{\sigma_{st}(S)}{\sigma_{st}}$$

– **Normalization**: divide by $2\binom{|V-S|}{2}$

III. Group Closeness Centrality

$$C_c^{\text{group}}(S) = \frac{1}{\bar{l}_S^{\text{group}}}$$

- It is the average distance from non-members to the group

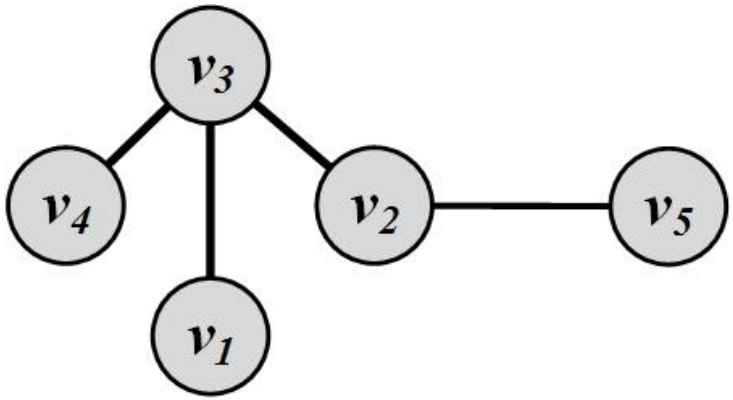
$$\bar{l}_S^{\text{group}} = \frac{1}{|V-S|} \sum_{v_j \notin S} l_{S,v_j}$$

$$l_{S,v_j} = \min_{v_i \in S} l_{v_i,v_j}$$

- One can also utilize the *maximum distance* or the *average distance*

Group Centrality Example

- Consider $S = \{v_2, v_3\}$



- Group degree centrality = **3**
- Group betweenness centrality = **3**
- Group closeness centrality = **1**

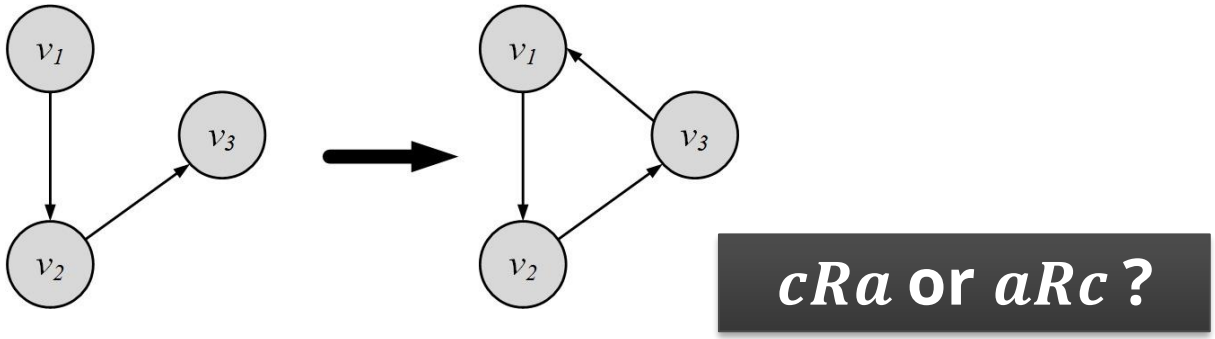
Friendship Patterns

- **Transitivity/Reciprocity**
- **Status/Balance**

I. Transitivity and Reciprocity

Transitivity

- Mathematic representation:
  - For a transitive relation R: $aRb \wedge bRc \rightarrow aRc$



- In a social network:
  - ***Transitivity is when a friend of my friend is my friend***
  - Transitivity in a social network leads to a denser graph, which in turn is closer to a complete graph
  - We can determine how close graphs are to the complete graph by measuring transitivity

[Global] Clustering Coefficient

- **Clustering coefficient** measures transitivity in undirected graphs
  - Count paths of length two and check whether the third edge exists

$$C = \frac{|\text{Closed Paths of Length 2}|}{|\text{Paths of Length 2}|}$$

When counting triangles, since every triangle has 6 closed paths of length 2

$$C = \frac{(\text{Number of Triangles}) \times 6}{|\text{Paths of Length 2}|}$$

Clustering Coefficient and Triples

Or we can rewrite it as

$$C = \frac{(\text{Number of Triangles}) \times 3}{\text{Number of Connected Triples of Nodes}}$$

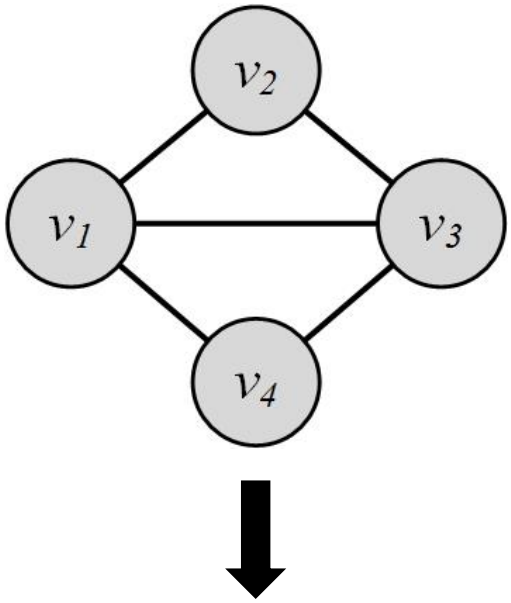
- **Triple**: an ordered set of three nodes,
  - connected by two (open triple) edges or
  - three edges (closed triple)

- A triangle can miss any of its three edges
  - A triangle has **3 Triples**

$v_i v_j v_k$ and $v_j v_k v_i$ are different triples
- The same members
- First missing edge $e(v_k, v_i)$ and second missing $e(v_i, v_j)$

$v_i v_j v_k$ and $v_k v_j v_i$ are the same triple

[Global] Clustering Coefficient: Example



$$C = \frac{\text{(Number of Triangles)} \times 3}{\text{Number of Connected Triples of Nodes}}$$

$$= \frac{2 \times 3}{2 \times 3 + \underbrace{2}_{v_2v_1v_4, v_2v_3v_4}} = 0.75.$$

Local Clustering Coefficient

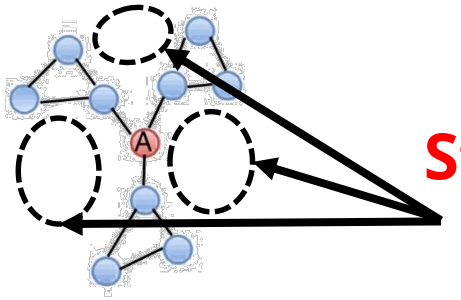
- Local clustering coefficient measures transitivity at the node level
  - Commonly employed for undirected graphs
  - Computes how strongly neighbors of a node v (nodes adjacent to v) are themselves connected

$$C(v_i) = \frac{\text{Number of Pairs of Neighbors of } v_i \text{ That Are Connected}}{\text{Number of Pairs of Neighbors of } v_i}.$$

In an undirected graph, the denominator can be rewritten as:

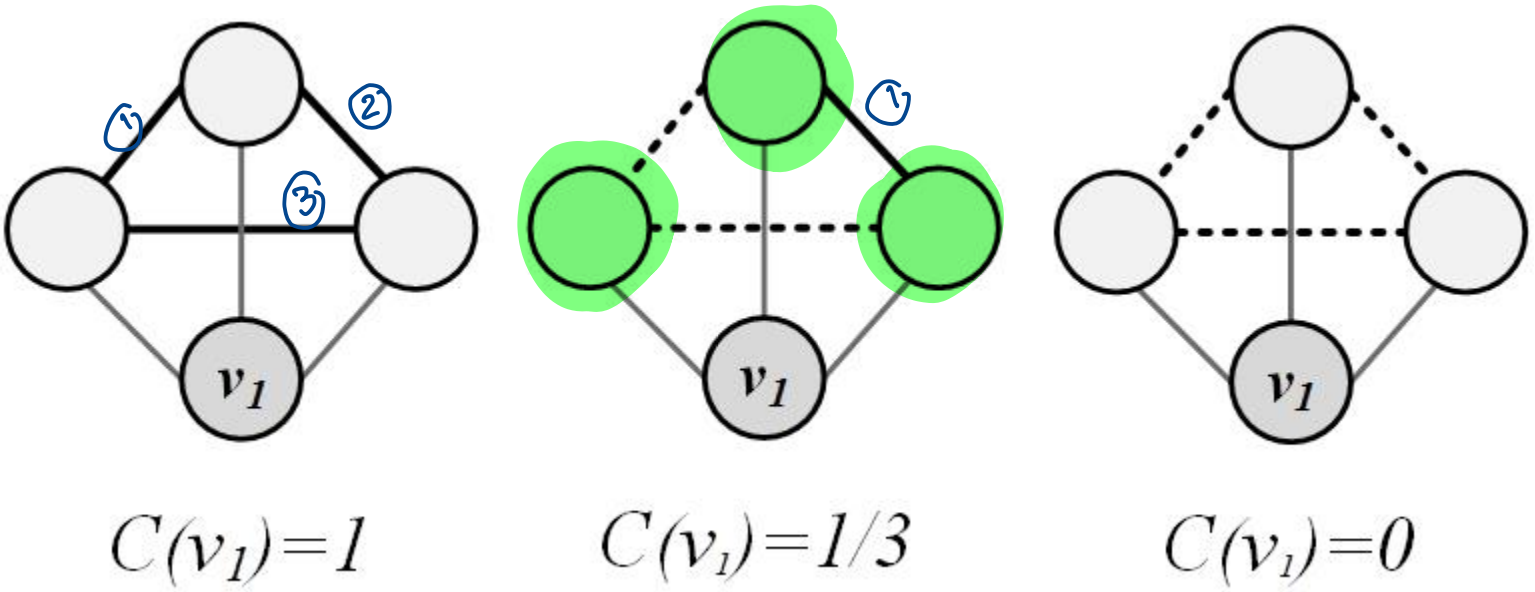
$$\binom{d_i}{2} = d_i(d_i - 1)/2$$

Provides a way to determine **structural holes**



Structural Holes

Local Clustering Coefficient: Example

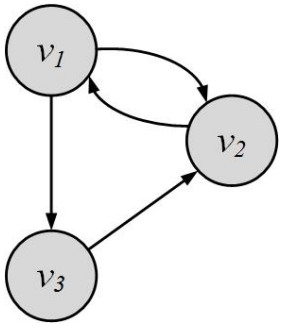


$C(v_1)=1$ $C(v_1)=1/3$ $C(v_1)=0$

- Thin lines depict connections to neighbors
- Dashed lines are the missing link among neighbors
- Solid lines indicate connected neighbors
  - When none of neighbors are connected $C = 0$
  - When all neighbors are connected $C = 1$

Reciprocity

If you become my friend, I'll be yours



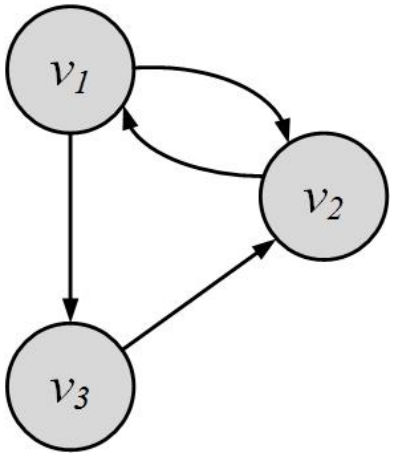
- Reciprocity is simplified version of transitivity
  - It considers closed loops of length 2
- If node v is connected to node u,
  - u by connecting to v, exhibits reciprocity

$$
\begin{aligned}
R &= \frac{\sum_{i,j,i<j} A_{i,j} A_{j,i}}{|E|/2}, \\
&= \frac{2}{|E|} \sum_{i,j,i<j} A_{i,j} A_{j,i}, \\
&= \frac{2}{|E|} \times \frac{1}{2} \mathrm{Tr}(A^2), \\
&= \frac{1}{|E|} \mathrm{Tr}(A^2), \\
&= \frac{1}{m} \mathrm{Tr}(A^2).
\end{aligned}
$$

What about $i = j$?

$$\mathrm{Tr}(A) = A_{1,1} + A_{2,2} + \cdots + A_{n,n} = \sum_{i=1}^{n} A_{i,i}$$

Reciprocity: Example



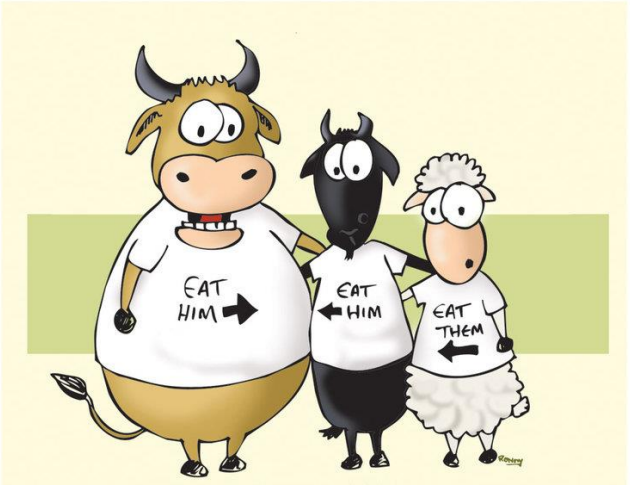
$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Reciprocal nodes: v_1, v_2

$$R = \frac{1}{m}\mathrm{Tr}(A^2) = \frac{1}{4}\mathrm{Tr}\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}\right) = \frac{2}{4} = \frac{1}{2}.$$

II. Balance and Status

- Measuring consistency in friendships



Social Balance Theory

Social balance theory

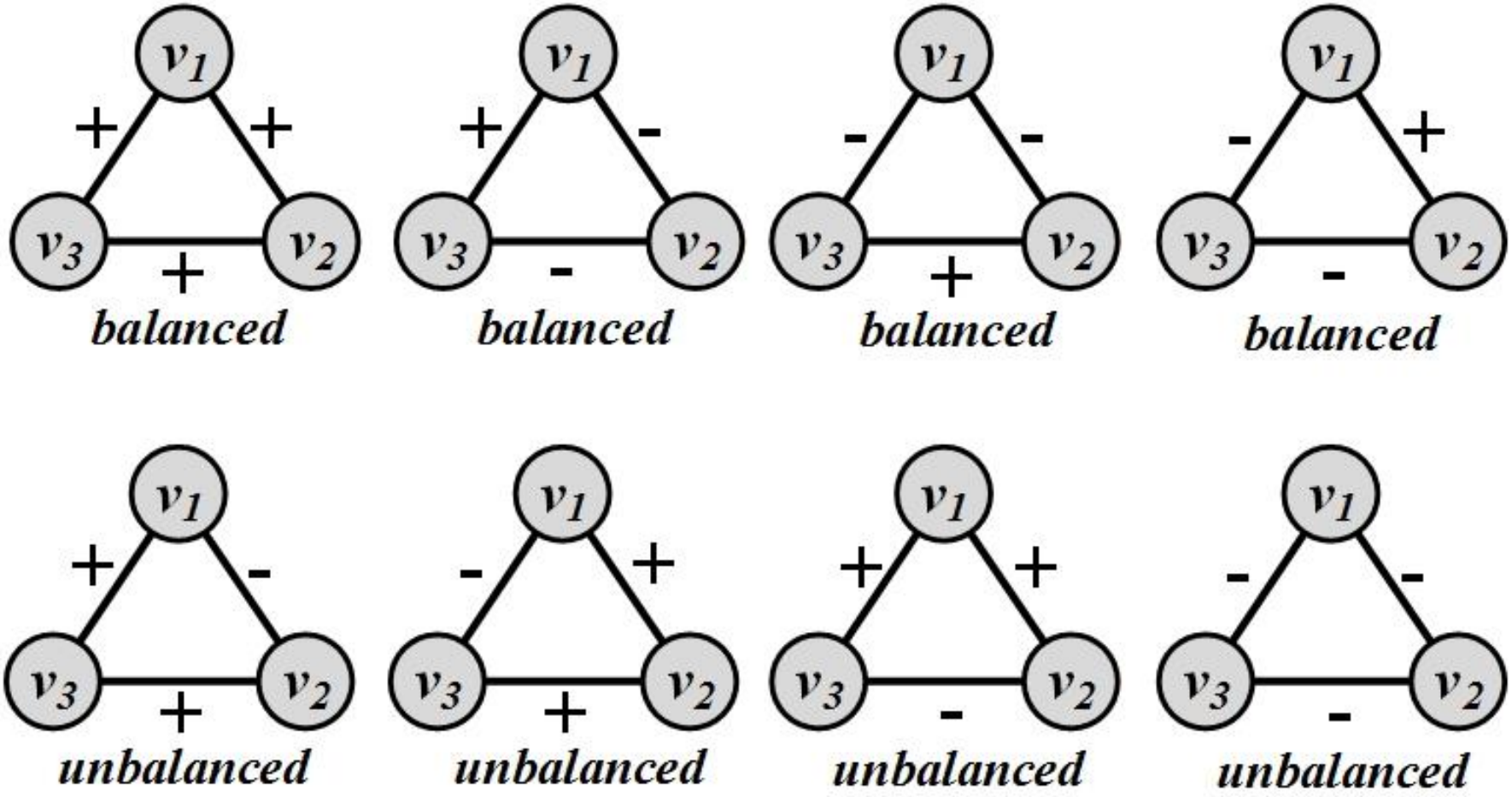
- Consistency in friend/foe relationships among individuals
- Informally, friend/foe relationships are consistent when

$$\begin{aligned}&\textit{The friend of my friend is my friend,}\\&\textit{The friend of my enemy is my enemy,}\\&\textit{The enemy of my enemy is my friend,}\\&\textit{The enemy of my friend is my enemy.}\end{aligned}$$

- In the network
  - Positive edges demonstrate friendships ($w_{ij} = 1$)
  - Negative edges demonstrate being enemies ($w_{ij} = -1$)
- Triangle of nodes i, j, and k, is balanced, if and only if
  - w_{ij} denotes the value of the edge between nodes i and j

$$w_{ij}w_{jk}w_{ki} \geq 0.$$

Social Balance Theory: Possible Combinations



For any cycle, if the multiplication of edge values become positive, then the cycle is socially balanced

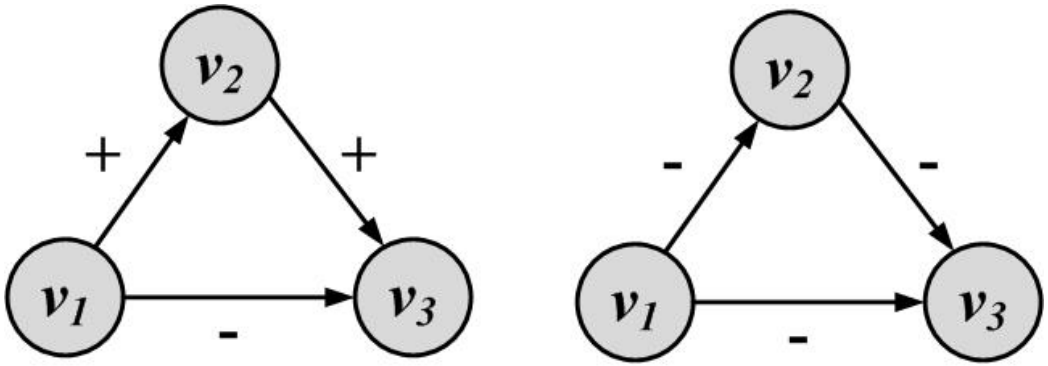
Social Status Theory

- **Status:** how prestigious an individual is ranked within a society

- **Social status theory:**
  - How consistent individuals are in assigning status to their neighbors
  - Informally,

If X has a higher status than Y and Y has a higher status than Z, then X should have a higher status than Z.

Social Status Theory: Example



Unstable configuration Stable configuration

- A directed **'+'** edge from node X to node Y shows that Y has a higher status than X and a **'-'** one shows vice versa

Similarity

How similar are two nodes in a network?

- **Structural Equivalence**
- **Regular Equivalence**

Structural Equivalence

- **Structural Equivalence:**
  - We look at the neighborhood shared by two nodes;
  - The size of this shared neighborhood defines how similar two nodes are.

- ***Example:***
  - *Two brothers have in common*
    - *sisters, mother, father, grandparents, etc.*
  - *This shows that they are similar,*
  - *Two random male or female individuals do not have much in common and are dissimilar.*

Structural Equivalence: Definitions

- **Vertex similarity**: $\sigma(v_i, v_j) = |N(v_i) \cap N(v_j)|$

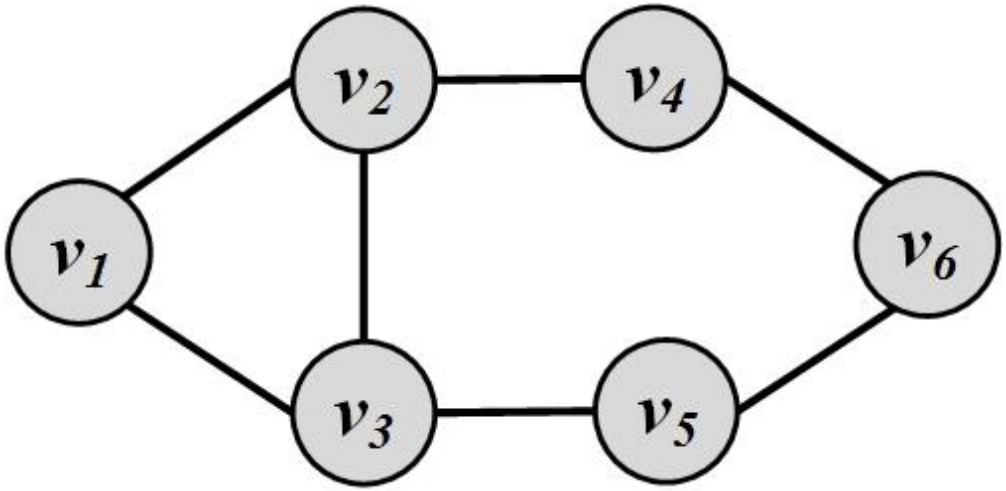
Normalize?

Jaccard Similarity: $\sigma_{Jaccard}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$

Cosine Similarity: $\sigma_{Cosine}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{\sqrt{|N(v_i)||N(v_j)|}}$

- The neighborhood $N(v)$ often excludes the node itself v.
  - **What can go wrong?**
    - Connected nodes not sharing a neighbor will be assigned zero similarity
  - **Solution:**
    - We can assume nodes are included in their neighborhoods

Similarity: Example



$$\sigma_{\text{Jaccard}}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cap \{v_3, v_6\}|}{|\{v_1, v_3, v_4, v_6\}|} = 0.25 \quad \frac{1}{4}$$

$$\sigma_{\text{Cosine}}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cap \{v_3, v_6\}|}{\sqrt{|\{v_1, v_3, v_4\}||\{v_3, v_6\}|}} = 0.40. \quad \frac{1}{\sqrt{6}}$$

Similarity Significance

Measuring Similarity Significance: compare the calculated similarity value with its expected value where vertices pick their neighbors at random

- For vertices v_i and v_j with degrees d_i and d_j this expectation is $d_i d_j / n$
  - There is a d_i/n chance of becoming v_i's neighbor
  - v_j selects d_j neighbors

- We can rewrite neighborhood overlap as

$$\sigma(v_i, v_j) = |N(v_i) \cap N(v_j)| = \sum_k A_{i,k} A_{j,k}$$

Normalized Similarity, cont.

$$\sigma_{\text{significance}}(v_i, v_j) = \sum_k A_{i,k} A_{j,k} - \frac{d_i d_j}{n} \qquad \bar{A}_i = \frac{1}{n} \sum_k A_{i,k}$$

$$= \sum_k A_{i,k} A_{j,k} - n \frac{1}{n} \sum_k A_{i,k} \frac{1}{n} \sum_k A_{j,k}$$

$$= \sum_k A_{i,k} A_{j,k} - n \bar{A}_i \bar{A}_j$$

$$= \sum_k (A_{i,k} A_{j,k} - \bar{A}_i \bar{A}_j)$$

$$= \sum_k (A_{i,k} A_{j,k} - \bar{A}_i \bar{A}_j - \bar{A}_i \bar{A}_j + \bar{A}_i \bar{A}_j)$$

$$= \sum_k (A_{i,k} A_{j,k} - A_{i,k} \bar{A}_j - \bar{A}_i A_{j,k} + \bar{A}_i \bar{A}_j)$$

$$= \boxed{\sum_k (A_{i,k} - \bar{A}_i)(A_{j,k} - \bar{A}_j)}$$

What is this?

Normalized Similarity, cont.

n times the Covariance between A_i and A_j

$$\frac{1}{n}\sum_k (A_{i,k} - \bar{A}_i)(A_{j,k} - \bar{A}_j)$$

Normalize covariance by the multiplication of Variances.

$$\sqrt{\frac{1}{n}\sum_k (A_{i,k} - \bar{A}_i)^2}\sqrt{\frac{1}{n}\sum_k (A_{j,k} - \bar{A}_j)^2}$$

We get **Pearson correlation coefficient**

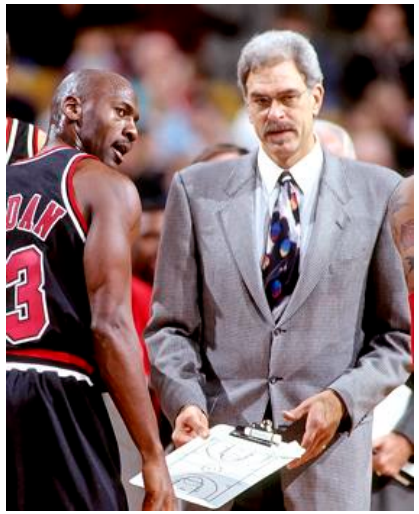
$$\begin{aligned}\sigma_{\text{pearson}}(v_i, v_j) &= \frac{\sigma_{\text{significance}}(v_i, v_j)}{\sqrt{\sum_k (A_{i,k} - \bar{A}_i)^2}\sqrt{\sum_k (A_{j,k} - \bar{A}_j)^2}} \\ &= \frac{\sum_k (A_{i,k} - \bar{A}_i)(A_{j,k} - \bar{A}_j),}{\sqrt{\sum_k (A_{i,k} - \bar{A}_i)^2}\sqrt{\sum_k (A_{j,k} - \bar{A}_j)^2}}\end{aligned}$$

(range of $\sigma \in [-1,1]$)

Regular Equivalence

- In regular equivalence,
  - We **do not** look at neighborhoods shared between individuals, but
  - **How neighborhoods themselves are similar**

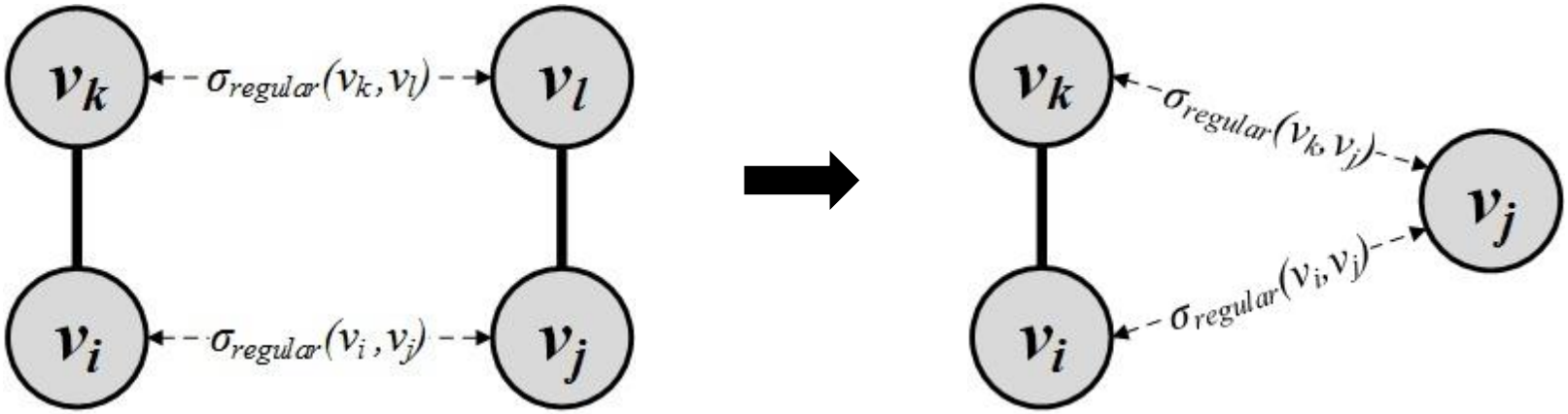
- ***Example*:**
  - ***Athletes are similar not because they know each other in person, but since they know similar individuals, such as coaches, trainers, other players, etc.***



Regular Equivalence

- v_i, v_j are similar when their neighbors v_k and v_l are similar

$$\sigma_{\text{regular}}(v_i, v_j) = \alpha \sum_{k,l} A_{i,k} A_{j,l} \sigma_{\text{regular}}(v_k, v_l)$$



- The equation (left figure) is hard to solve since it is self referential so we relax our definition using the right figure

Regular Equivalence

- v_i and v_j are similar when v_j is similar to v_i's neighbors v_k

$$\sigma_{regular}(v_i, v_j) = \alpha \sum_k A_{i,k} \sigma_{Regular}(v_k, v_j)$$

- In vector format

$$\sigma_{regular} = \alpha A \sigma_{Regular}$$

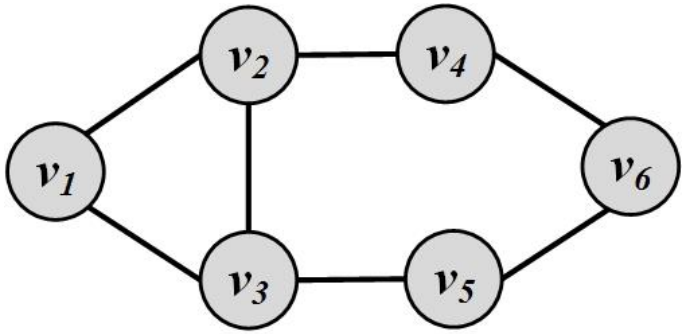
A vertex is highly similar to itself, we guarantee this by adding an identity matrix to the equation

$$\sigma_{regular} = \alpha A \sigma_{Regular} + \mathbf{I}$$

$$\sigma_{regular} = (\mathbf{I} - \alpha A)^{-1}$$

When $\alpha < 1/\lambda_{max}$ the matrix is invertible

Regular Equivalence: Example



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

The largest eigenvalue of A is 2.43

Set $\alpha = 0.3 < 1/2.43$

$$\sigma_{\text{regular}} = (I - 0.3A)^{-1} = \begin{bmatrix} 1.43 & 0.73 & 0.73 & 0.26 & 0.26 & 0.16 \\ 0.73 & 1.63 & 0.80 & 0.56 & 0.32 & 0.26 \\ 0.73 & 0.80 & 1.63 & 0.32 & 0.56 & 0.26 \\ 0.26 & 0.56 & 0.32 & 1.31 & 0.23 & 0.46 \\ 0.26 & 0.32 & 0.56 & 0.23 & 1.31 & 0.46 \\ 0.16 & 0.26 & 0.26 & 0.46 & 0.46 & 1.27 \end{bmatrix}$$

- Any row/column of this matrix shows the similarity to other vertices
- Vertex 1 is most similar (other than itself) to vertices 2 and 3
- Nodes 2 and 3 have the highest similarity (**regular equivalence**)