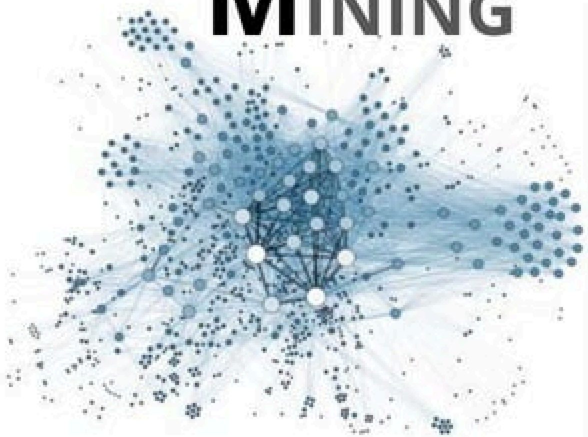
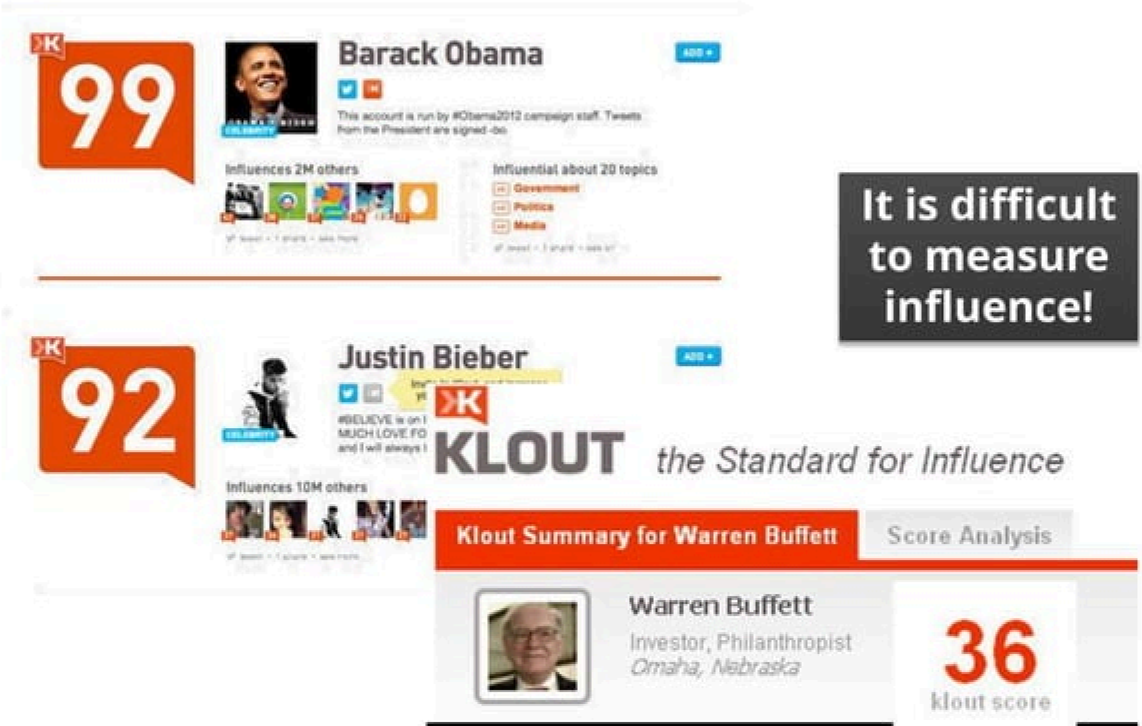
Network Measures

Social Media Mining



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

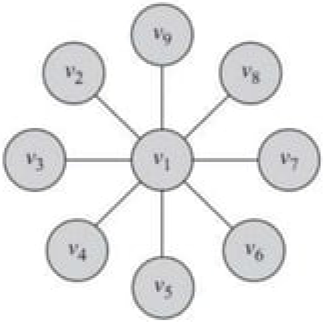
Degree Centrality

- **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.

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

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

Normalized Degree Centrality

- Normalized by the maximum possible degree

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

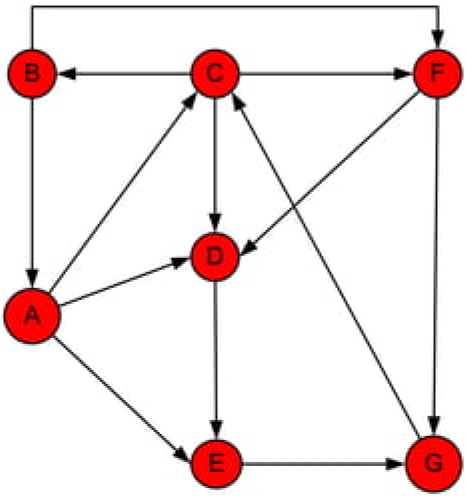
- Normalized by the maximum degree

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

- 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

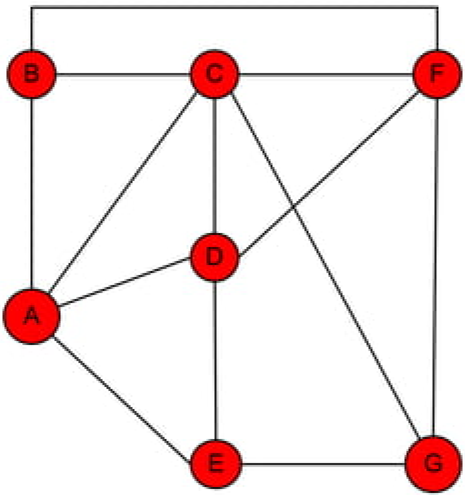


Node	In-Degree	Out-Degree	Centrality	Rank
A	1	3	1/2	**1**
B	1	2	1/3	**3**
C	2	3	1/2	**1**
D	3	1	1/6	**5**
E	2	1	1/6	**5**
F	2	2	1/3	**3**
G	2	1	1/6	**5**

Normalized by the maximum possible degree

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

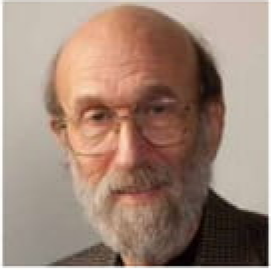
Degree Centrality (undirected Graph) Example



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

λ_1 is the largest eigenvalue

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

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

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

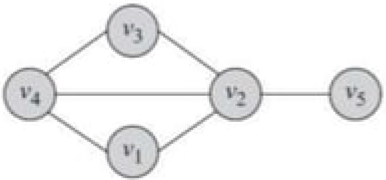
Largest Eigenvalue

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

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

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

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

Elihu Katz

Eigenvector Centrality

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

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.

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

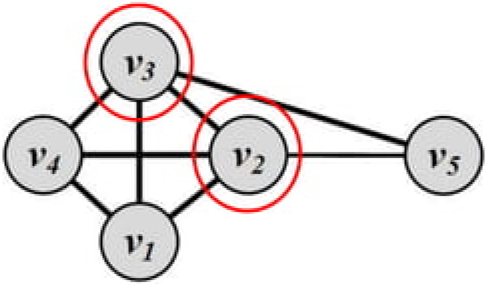
- When α=0, the eigenvector centrality is removed and all nodes get the same centrality value β
  - As α gets larger the effect of β is reduced

- 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 $\alpha < 1/\lambda$, where λ is the largest eigenvalue of 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, $\boxed{3.32}$
- We assume $\alpha=0.25 < \boxed{1/3.32}$ 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?

$$\begin{cases} d_j^{out} > 0 \\ D = diag(d_1^{\text{out}}, d_2^{\text{out}}, \ldots, d_n^{\text{out}}) \end{cases}$$

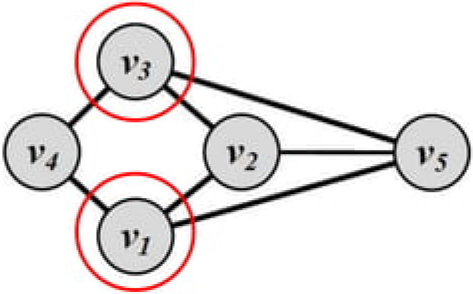
$$\mathbf{C}_p = \alpha A^T D^{-1} \mathbf{C}_p + \beta \mathbf{1}$$

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

Similar to Katz Centrality, in practice, $\boldsymbol{\alpha} < \mathbf{1}/\boldsymbol{\lambda}$, 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, $\boldsymbol{\alpha} < \mathbf{1}$.

PageRank Example

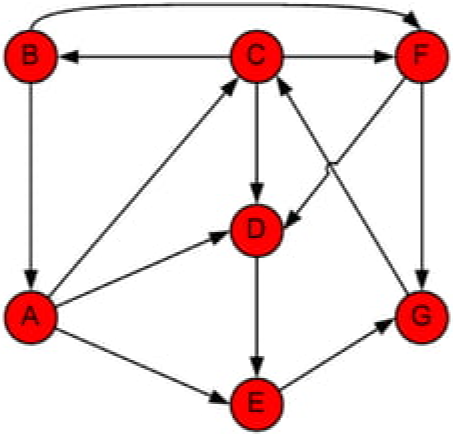
- We assume α=0.95 < 1 and and $\beta = 0.1$



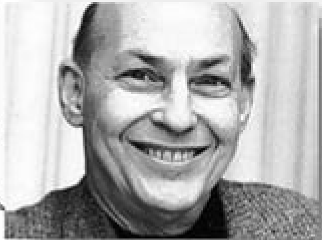
$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

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

PageRank Example – Alternative Approach [Markov Chains]



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



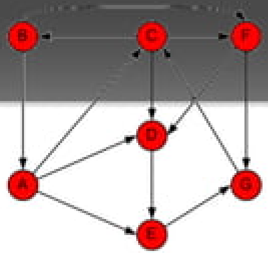
Marvin Minsky (1927-2016)

Using Power Method

α=1 and β =0?

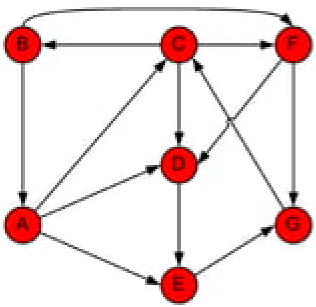
Step	A	B	C	D	E	F	G
0	1/7	1/7	1/7	1/7	1/7	1/7	1/7
1	B/2	C/3	A/3 + G	A/3 + C/3 + F/2	A/3 + D	C/3 + B/2	F/2 + E
	0.071	0.048	0.190	0.167	0.190	0.119	0.214

PageRank: Example

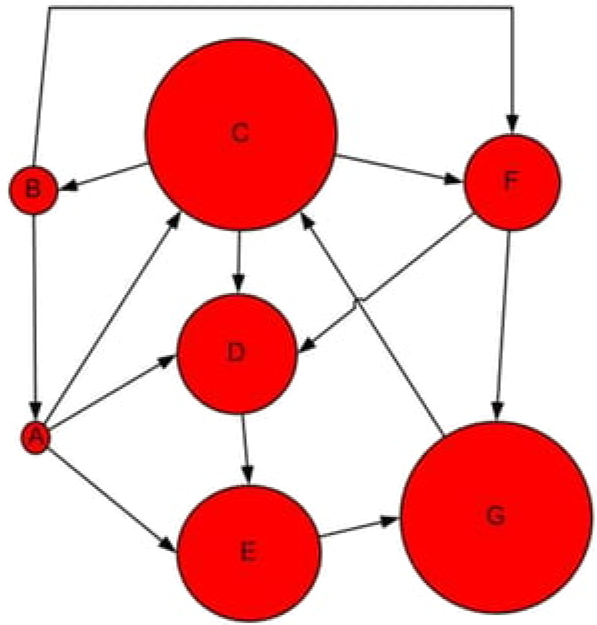


Step	A	B	C	D	E	F	G	Sum
1	0.143	0.143	0.143	0.143	0.143	0.143	0.143	1.000
2	0.071	0.048	0.190	0.167	0.190	0.119	0.214	1.000
3	0.024	0.063	0.238	0.147	0.190	0.087	0.250	1.000
4	0.032	0.079	0.258	0.131	0.155	0.111	0.234	1.000
5	0.040	0.086	0.245	0.152	0.142	0.126	0.210	1.000
6	0.043	0.082	0.224	0.158	0.165	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



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

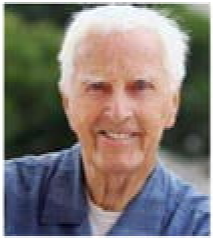


Centrality in terms of how you connect others

(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

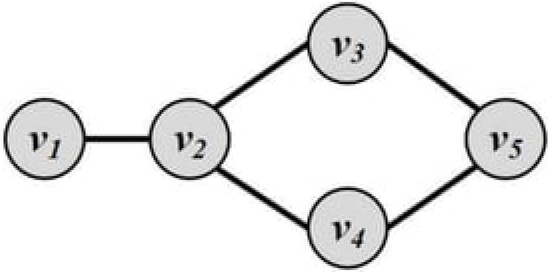
- In the best case, node v_i is on all shortest paths from s to t, hence, $\frac{\sigma_{st}(v_i)}{\sigma_{st}} = 1$

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

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



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