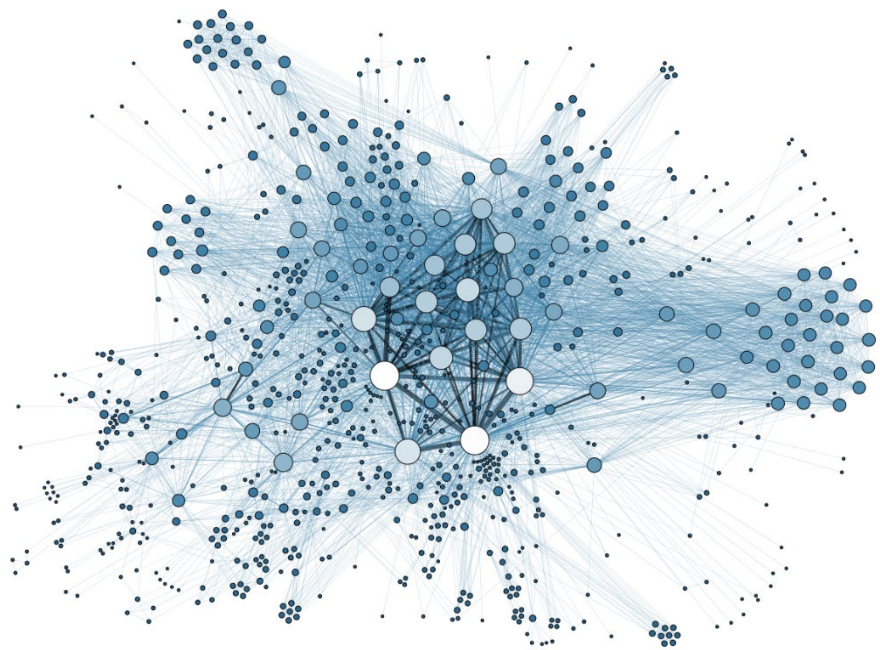
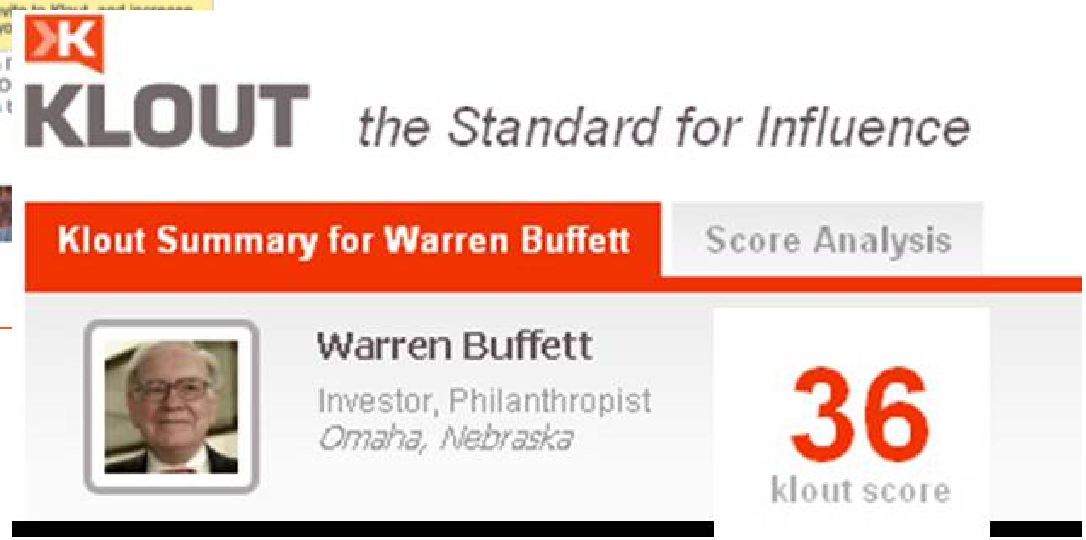
Social Media & Network Analytics

Social Network Analysis



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**
- To answer these and similar questions, one first needs to define measures for quantifying **centrality** and **level of interactions**, among others.

Outline

- Centrality
  - Who you connect with
  - How you connect others
  - How fast you can reach others
- Reciprocity and Transitivity
- Balance and Status

Centrality

Centrality defines how important a node is within a network

Centrality in terms of those who you are connected to

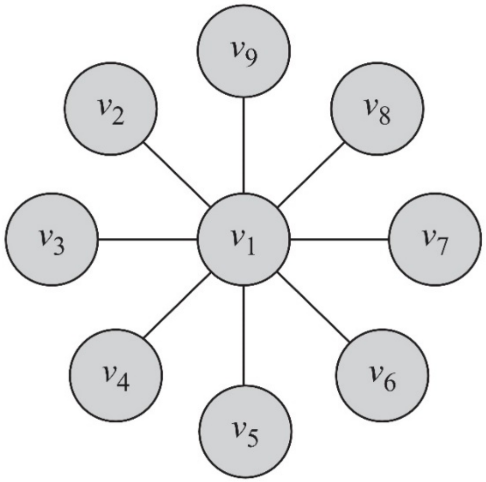
Degree Centrality

- **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

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

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

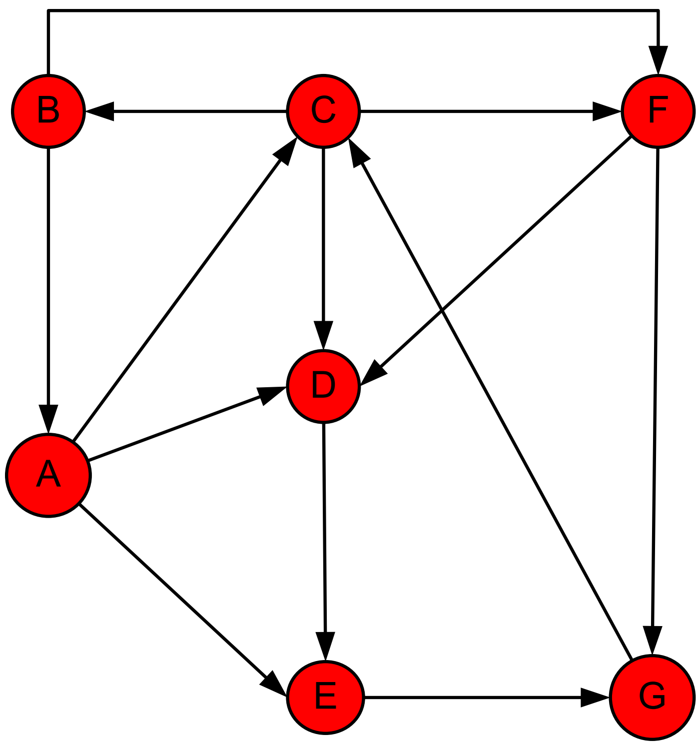
- Normalized by the maximum degree

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

- 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

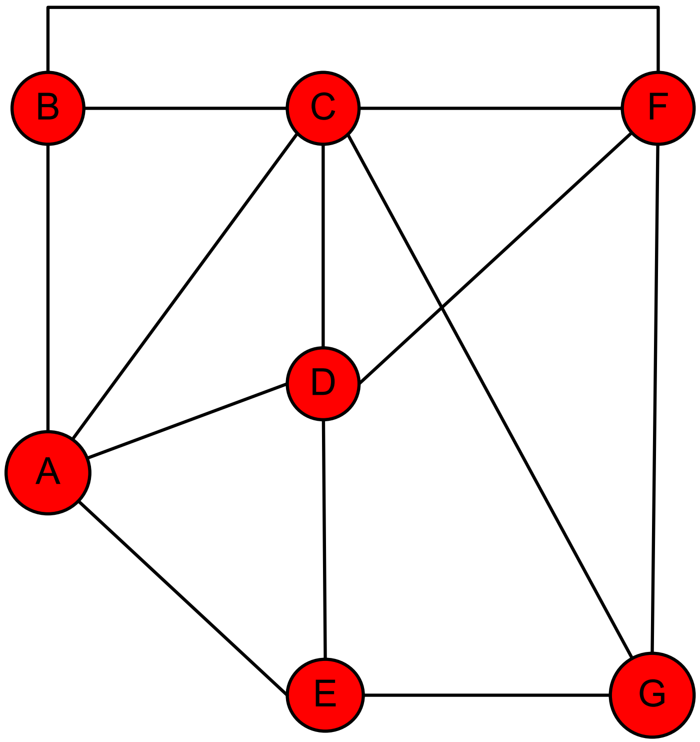


Node	In-Degree	Out-Degree	Centrality (out)	Rank
A	1	3	1/2	**1**
B	1	2	1/3	**3**
C	2	3	1/2	**1**
D	3	1	1/6	**5**
E	2	1	1/6	**5**
F	2	2	1/3	**3**
G	2	1	1/6	**5**

Normalized by the maximum possible degree

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

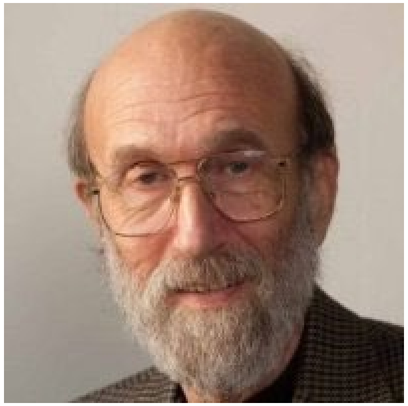
Degree Centrality (undirected Graph) Example



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

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

Intuition

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

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

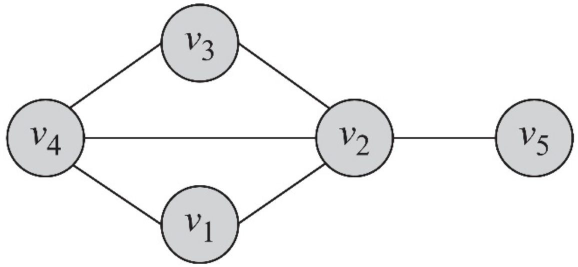
- We could iterate recursive equation till convergence
- Can do better!
  - By definition, $\boldsymbol{C_e}$ is an eigenvector of adjacency matrix A^T (or A when undirected) and λ is the corresponding eigenvalue
- Which eigenvalue-eigenvector pair should we choose?
  - We want C_e to be non-negative
  - If we start with an all positive centrality, $C_e(v_i) > 0\ \forall i$, by Perron-Frobenius Theorem the eigenvector C_e of the largest eigenvalue $\lambda_{largest}$ satisfies this (it is the only eigenvector guaranteed to)

Eigenvector Centrality, cont.

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



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

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

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

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

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

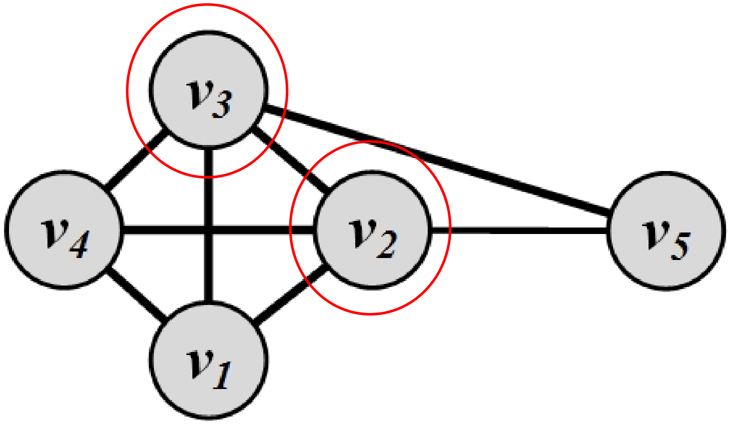
$$\mathbf{C}_{\text{Katz}} = \alpha A^T \mathbf{C}_{\text{Katz}} + \beta \mathbf{1}$$

vector of all 1's

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

Usually choose α to be between 0 and $1/\lambda_{largest}$

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 α=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?

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

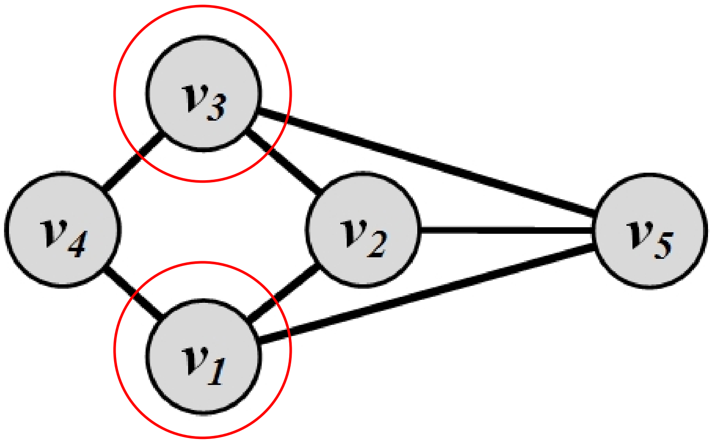
$$\mathbf{C}_p = \alpha A^T D^{-1} \mathbf{C}_p + \beta \mathbf{1}$$

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

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

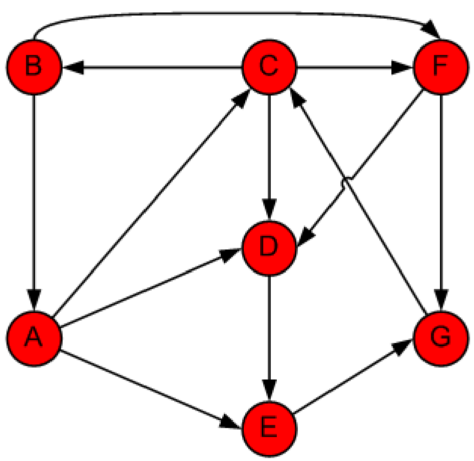
- We assume α=0.95 < 1 and and $\beta = 0.1$



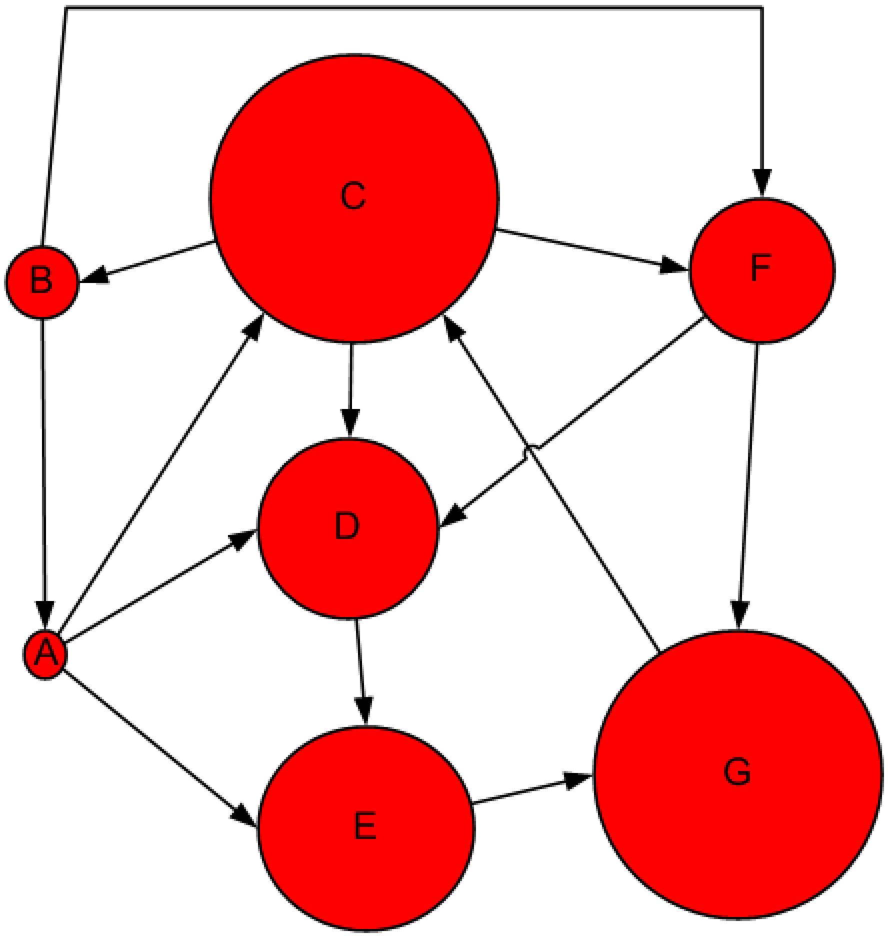
$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

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

Effect of PageRank



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



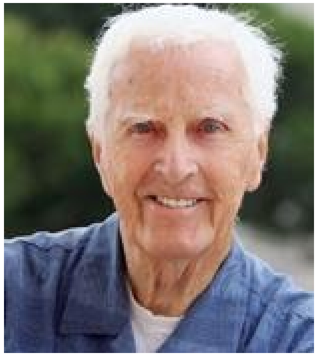
Break time

- **Trivia**: What is Dunbar's number?

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

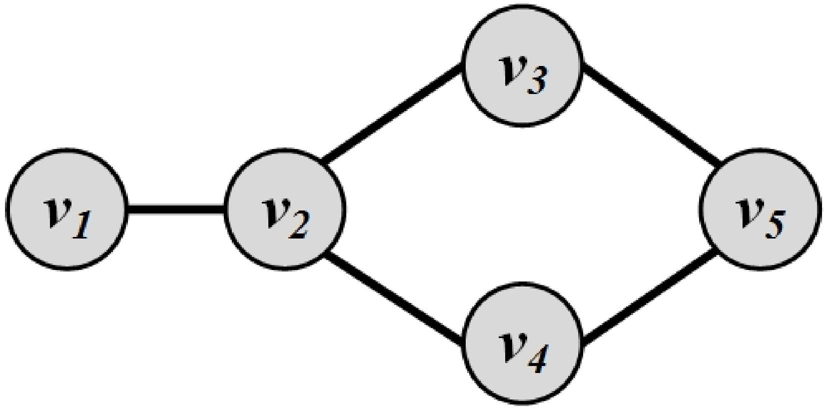
- In the best case, node v_i is on all shortest paths from s to t, hence, $\frac{\sigma_{st}(v_i)}{\sigma_{st}} = 1$

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

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

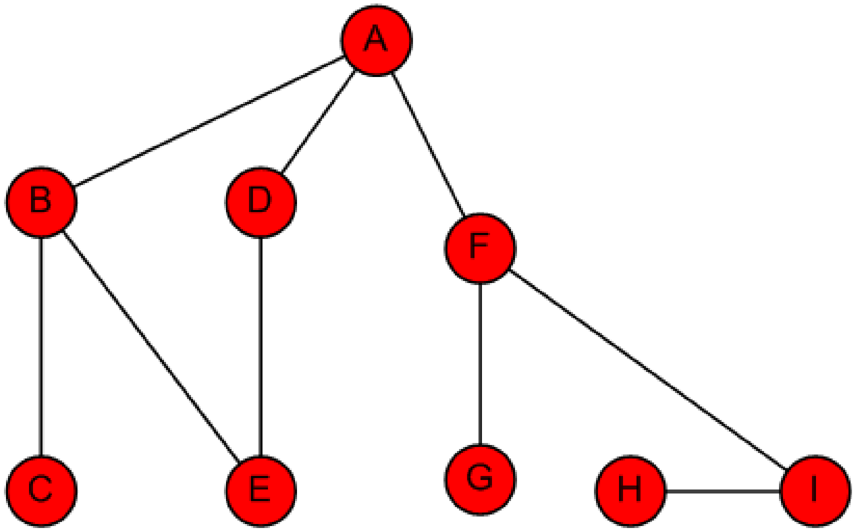
Betweenness centrality: $C_b^{\text{norm}}(v_i) = \frac{C_b(v_i)}{2\binom{n-1}{2}}$

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
  - Floyd-Warshall's algorithm, also $O(n^3)$

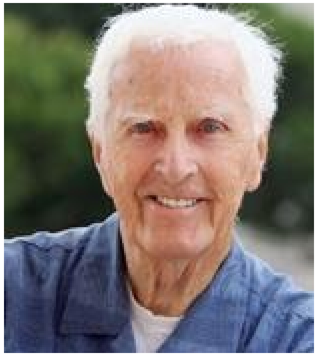
- Better Solution:
  - Brandes Algorithm:
    - $O(nm)$ for unweighted graphs
    - $O(nm + n^2 \log n)$ for weighted graphs

https://people.csail.mit.edu/jshun/6886-s18/papers/BrandesBC.pdf

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

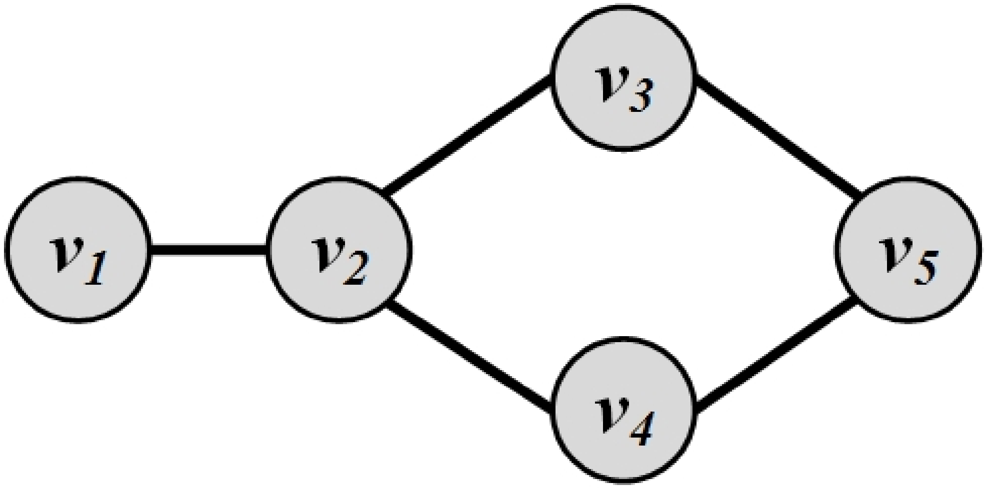


Linton Freeman

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

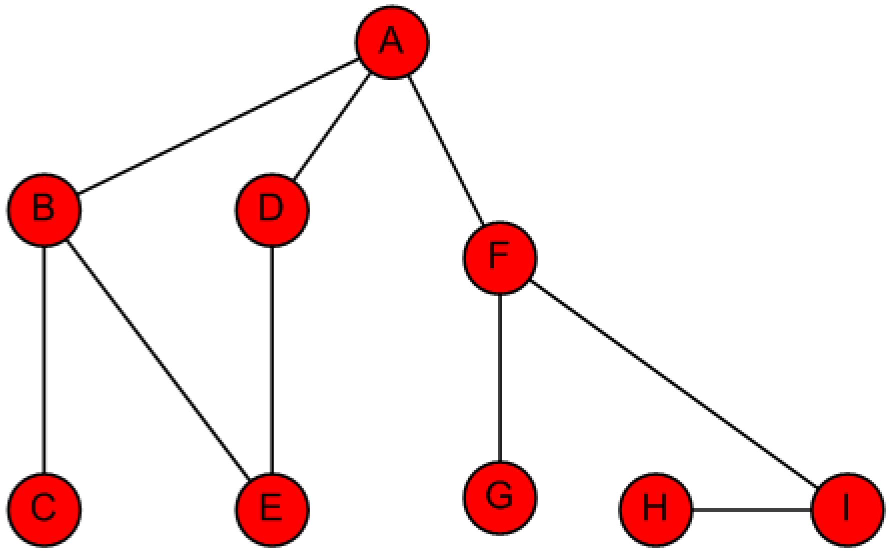
$$\bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

Closeness Centrality: Example 1



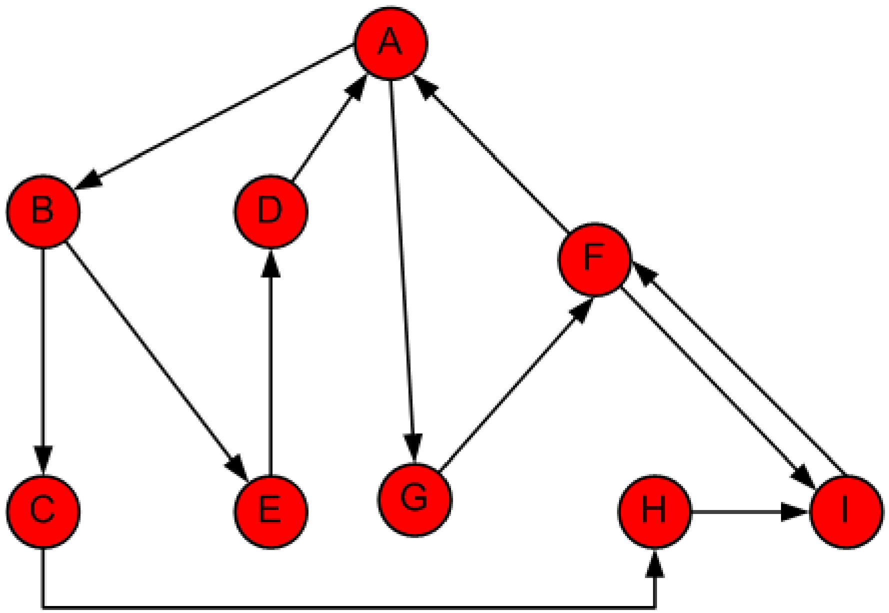
$$
\begin{aligned}
C_c(v_1) &= 1\ /\ (\ (1+2+2+3)/4\ ) = 0.5, \\
C_c(v_2) &= 1\ /\ (\ (1+1+1+2)/4\ ) = 0.8, \\
C_c(v_3) = \quad C_b(v_4) &= 1\ /\ (\ (1+1+2+2)/4\ ) = 0.66, \\
C_c(v_5) &= 1\ /\ (\ (1+1+2+3)/4\ ) = 0.57.
\end{aligned}
$$

Closeness Centrality: Example 2 (Undirected)



Node	A	B	C	D	E	F	G	H	I	l_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	l_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

Outline

- Centrality
  - Who you connect with
  - How you connect others
  - How fast you can reach others
- Reciprocity and Transitivity
- Balance and Status

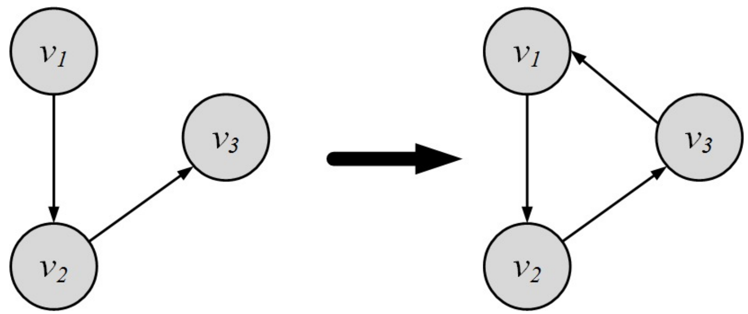
Friendship Patterns

- Transitivity/Reciprocity
- Status/Balance

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 or Transitivity

- **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

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

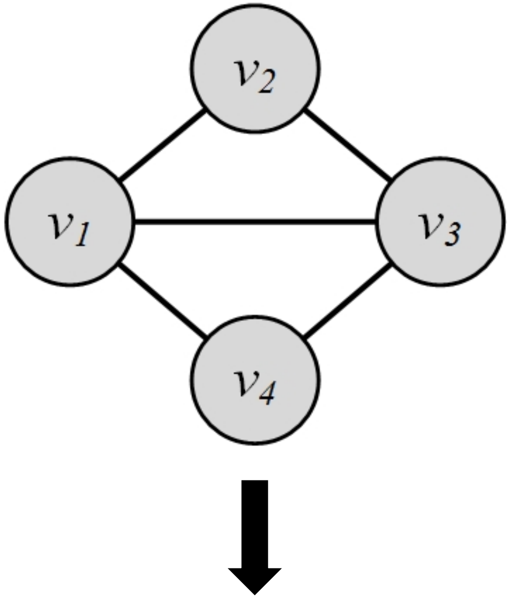
Clustering Coefficient and Triples

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 (v_i, v_j, v_k), that at least has two edges:
  - connected by two edges (open triple) with (i,k) edge absent or
  - three edges (closed triple)
- A triangle has **3 Triples**

[Global] Clustering Coefficient: Example



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

$$= \frac{2 \times 3}{2 \times 3 + \underbrace{2}_{v_2 v_1 v_4, v_2 v_3 v_4}} = 0.75.$$

Local Clustering Coefficient

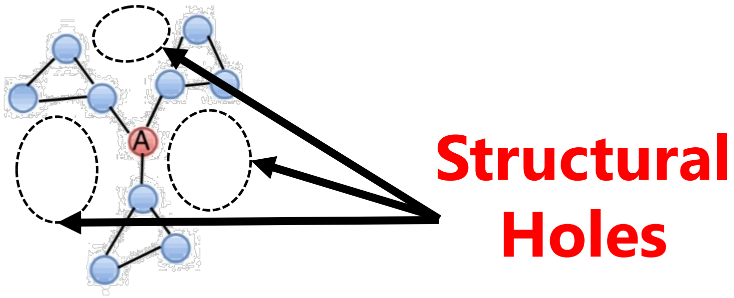
- Local clustering coefficient measures transitivity at the node level
  - Commonly employed for undirected graphs
  - Computes how strongly neighbors of a node v (nodes adjacent to v) are themselves connected

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

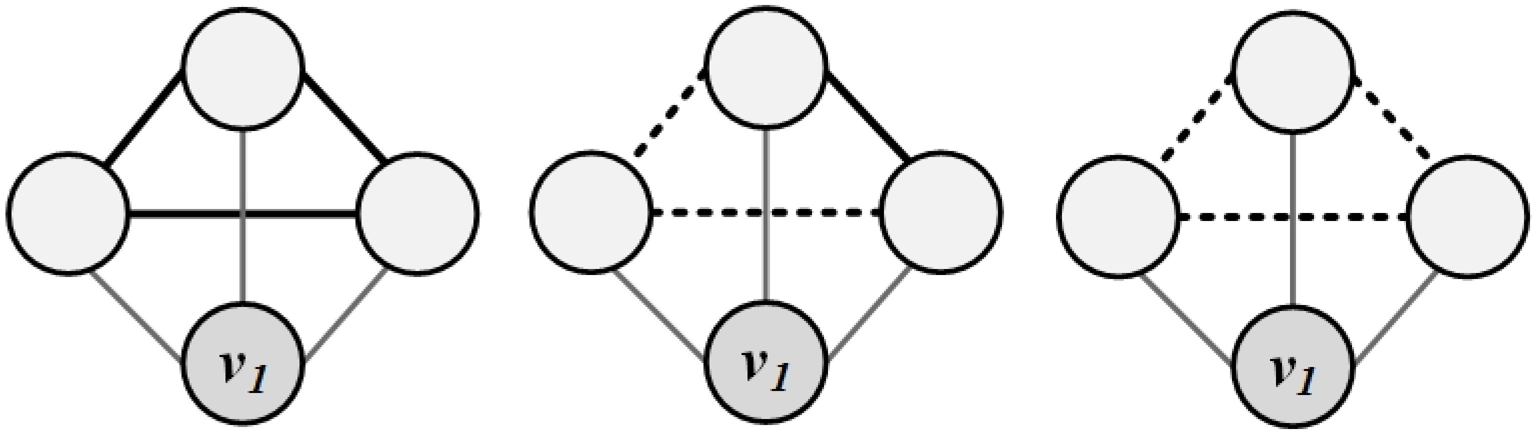
In an undirected graph, the denominator can be rewritten as:

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

Provides a way to determine **structural holes**



Local Clustering Coefficient: Example



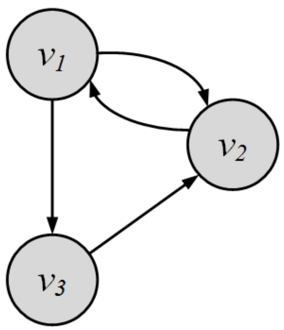
$C(v_1)=1$ $C(v_1)=1/3$ $C(v_1)=0$

- 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 (simplification of transitivity)
  - It considers closed loops of length 2 on directed graphs

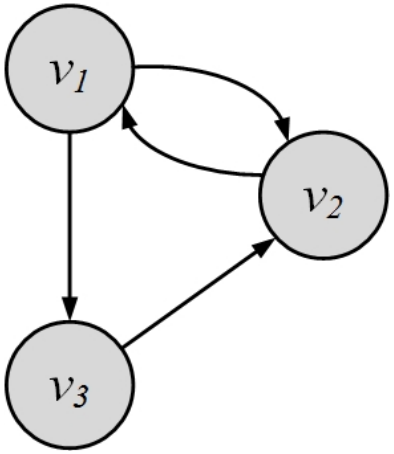


- $R = \dfrac{|reciprocal\ edges|}{|E|}$

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

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

Reciprocity: Example



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

Reciprocal nodes: v_1, v_2

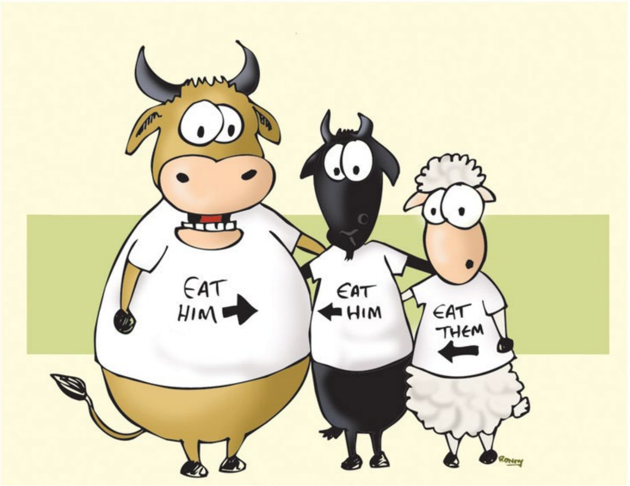
$$R = \frac{1}{m}\mathrm{Tr}(A^2) = \frac{1}{4}\mathrm{Tr}\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}\right) = \frac{2}{4} = \frac{1}{2}.$$

Outline

- Centrality
  - Who you connect with
  - How you connect others
  - How fast you can reach others
- Reciprocity and Transitivity
- Balance and Status

Balance and Status

- Measuring consistency in friendships



Social Balance Theory

Social balance theory

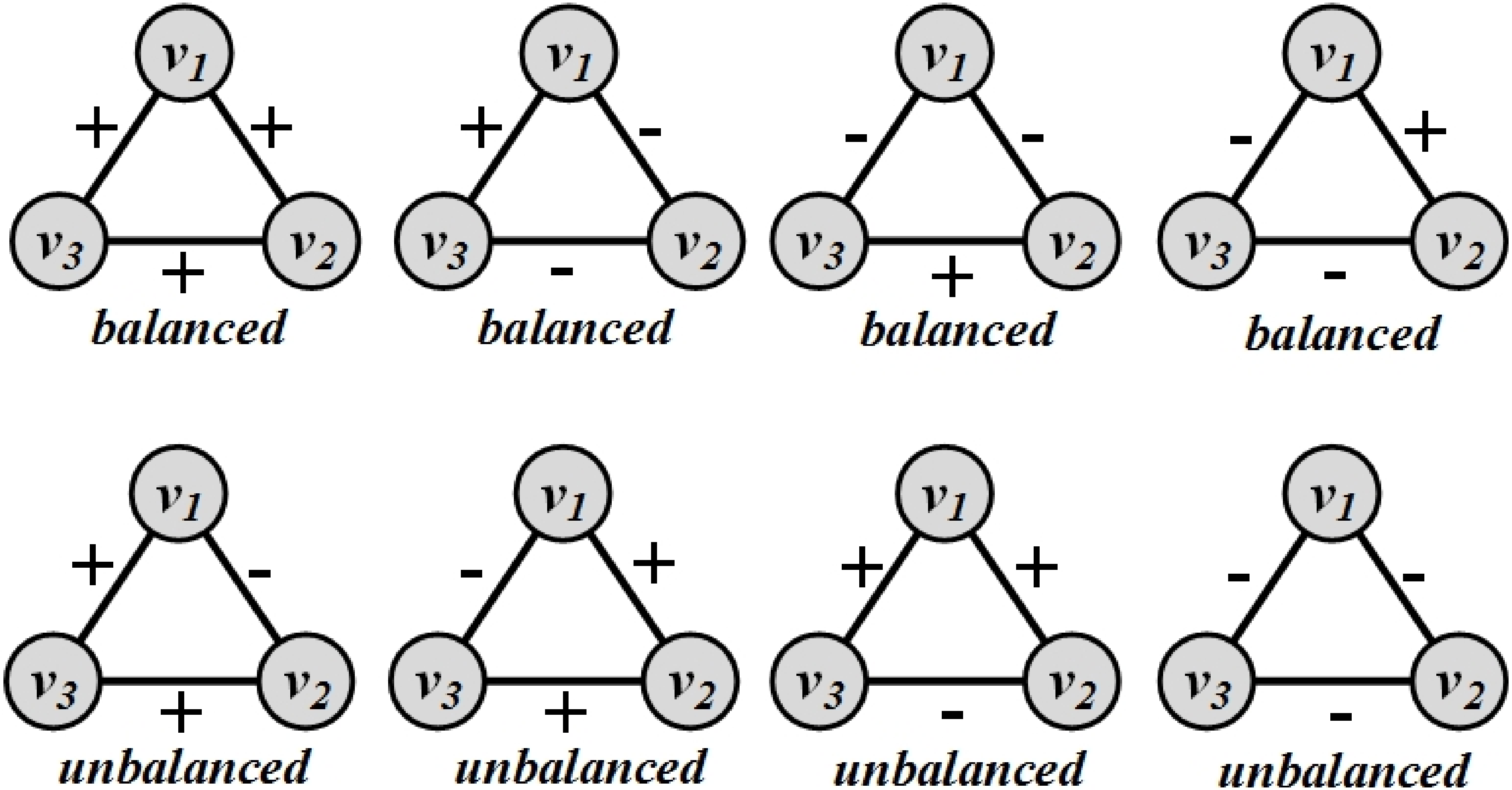
- Consistency in friend/foe relationships among individuals
- Informally, friend/foe relationships are consistent when

The friend of my friend is my friend,
The friend of my enemy is my enemy,
The enemy of my enemy is my friend,
The enemy of my friend is my enemy.

- 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

Summary

- Centrality
  - Neighbourhood
    - Degree
    - Eigenvector
      - Katz
    - Pagerank
  - Information broker
    - Betweeneness
  - Closeness
- Friendship
  - Transitivity and Reciprocity
  - Social Balance and Status Theories