COMP 631: Introduction to Information Retrieval

XIA (BEN) HU

CS, Rice University

https://cs.rice.edu/~xh37/index.html

Today's lecture

- Centrality
- Transitivity and Reciprocity
- Balance and Status
- Similarity

Betweenness Centrality

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

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

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

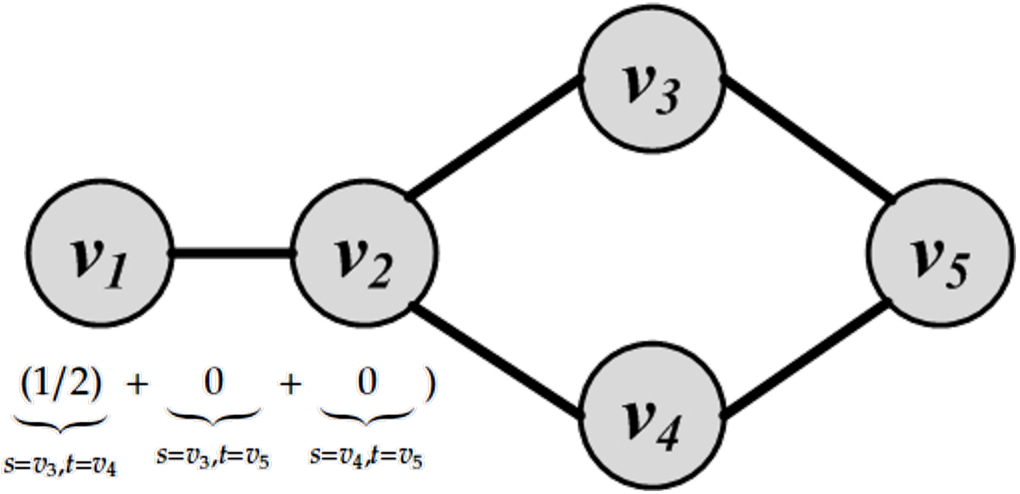
- In the best case, node v_i is on all shortest paths from s to t, hence, $\frac{\sigma_{st}(v_i)}{\sigma_{st}} = 1$

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}} = \sum_{s \neq t \neq v_i} 1.$$

Therefore, the maximum value is $2\binom{n-1}{2} = (n-1)(n-2)$.

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

Betweenness Centrality Example

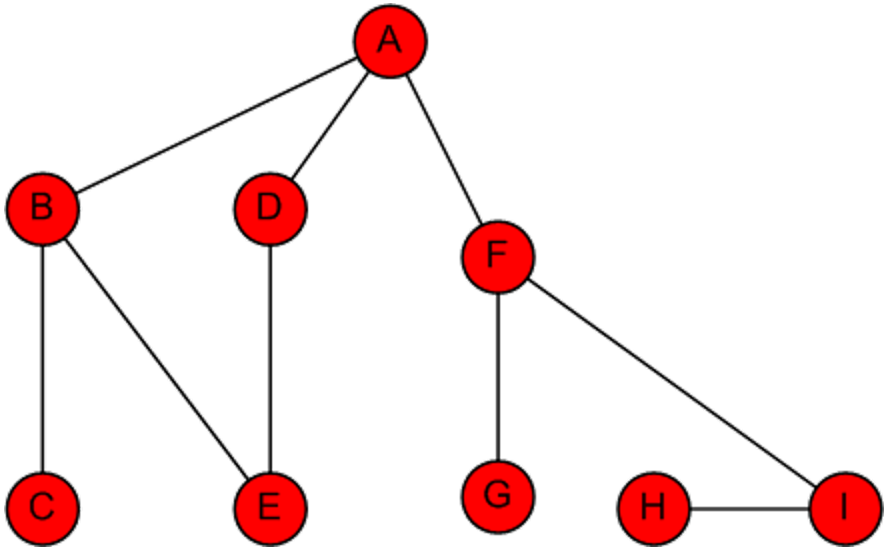


$$C_b(v_2) = 2 \times (\underbrace{(1/1)}_{s=v_1,t=v_3} + \underbrace{(1/1)}_{s=v_1,t=v_4} + \underbrace{(2/2)}_{s=v_1,t=v_5} + \underbrace{(1/2)}_{s=v_3,t=v_4} + \underbrace{0}_{s=v_3,t=v_5} + \underbrace{0}_{s=v_4,t=v_5})$$
$$= 2 \times 3.5 = 7,$$

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

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

Betweenness Centrality: Example



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

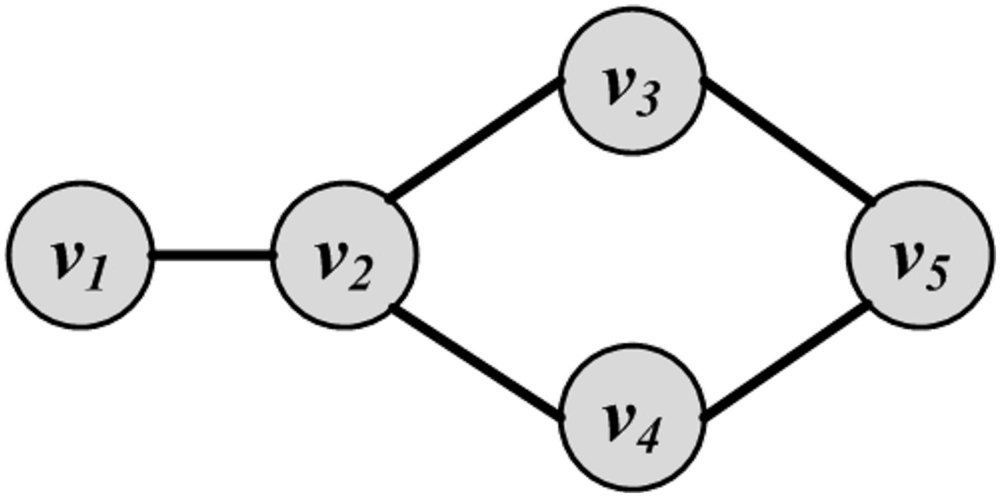
Closeness Centrality

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

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

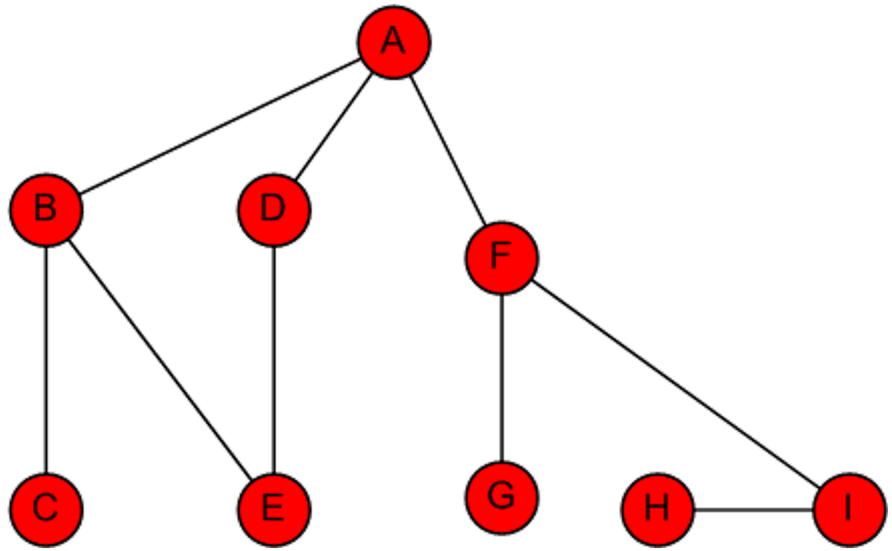
$$\bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

Compute Closeness Centrality



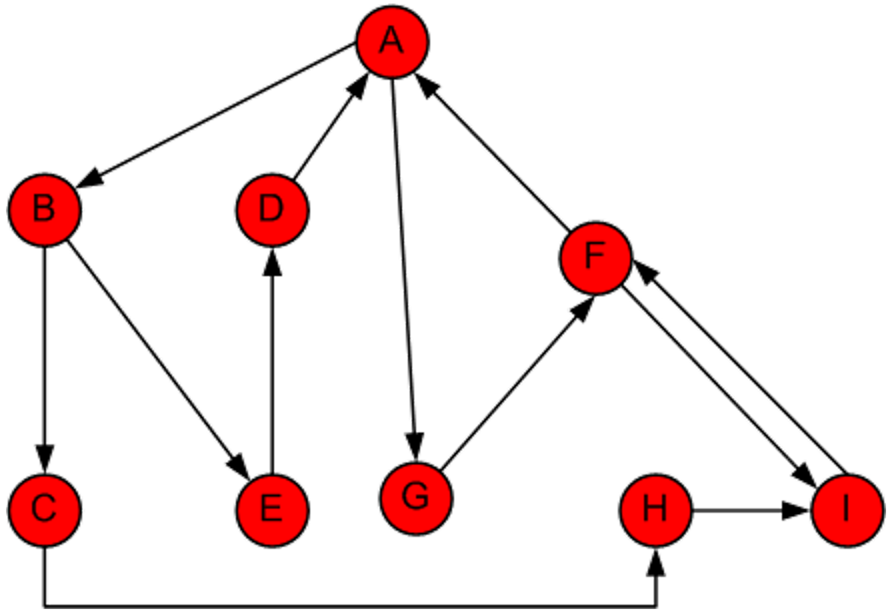
$$
\begin{aligned}
C_c(v_1) &= 1/((1+2+2+3)/4) = 0.5,\\
C_c(v_2) &= 1/((1+1+1+2)/4) = 0.8,\\
C_c(v_3) = C_b(v_4) &= 1/((1+1+2+2)/4) = 0.66,\\
C_c(v_5) &= 1/((1+1+2+3)/4) = 0.57
\end{aligned}
$$

Closeness Centrality (Undirected Graph): Example



Node	A	B	C	D	E	F	G	H	I	D_Avg	Closeness Centaliry	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



Node	A	B	C	D	E	F	G	H	I	D_Avg	Closeness Centaliry	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

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

- Group Degree Centrality

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

  – We can normalize it by dividing it by |V-S|

- Group Betweeness Centrality

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

- We can normalize it by dividing it by $2\binom{|V-S|}{2}$

Group Centrality

- Group Closeness Centrality

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

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

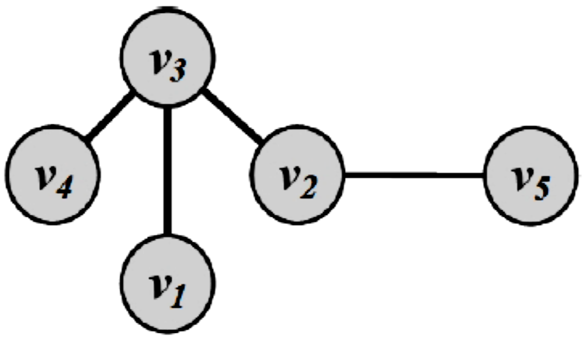
$$\bar{l}_S^{group} = \frac{1}{|V-S|} \sum_{v_j \notin S} l_{S,v_j}.$$

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

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

Group Centrality Example

- Consider S={v2,v3}



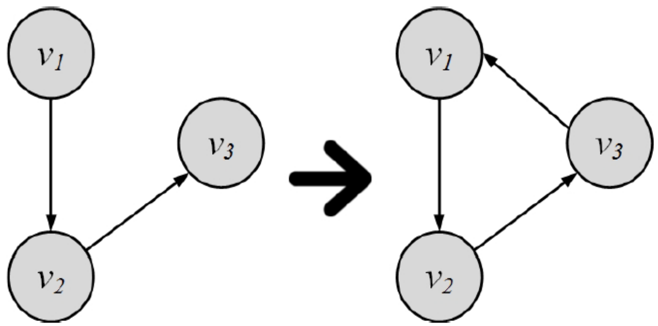
- Group degree centrality=3
- Group betweenness centrality = 6
- Group closeness centrality = 1

Today's lecture

- Centrality
- Transitivity and Reciprocity
- Balance and Status
- Similarity

Transitivity

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



- 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 analyzes transitivity in an undirected graph
  - We measure it by counting paths of length two and check whether the third edge exists

$$C = \frac{|\text{Paths of Length 2 that have the third edge}|}{|\text{Paths of Length 2}|}$$

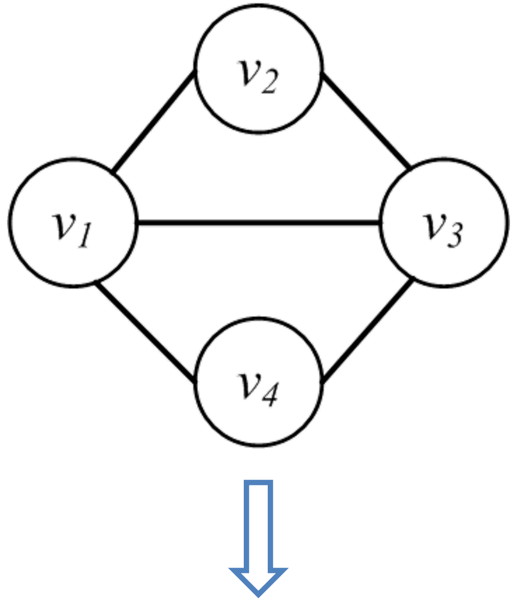
When counting triangles, since every triangle has 6 closed paths of length 2:

$$C = \frac{\text{number of triangles} \times 6}{|\text{paths of length 2}|}$$

In undirected networks:

$$C = \frac{(\text{number of triangles}) \times 3}{\text{number of connected 3 nodes}}$$

[Global] Clustering Coefficient: Example



$$C = \frac{(\textit{Number of Triangles}) \times 3}{\textit{Number of Connected Triples of Nodes}} = \frac{2 \times 3}{2 \times 3 + \underbrace{2}_{v_2v_1v_4,\ \ v_2v_3v_4}} = 0.75.$$

Local Clustering Coefficient

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

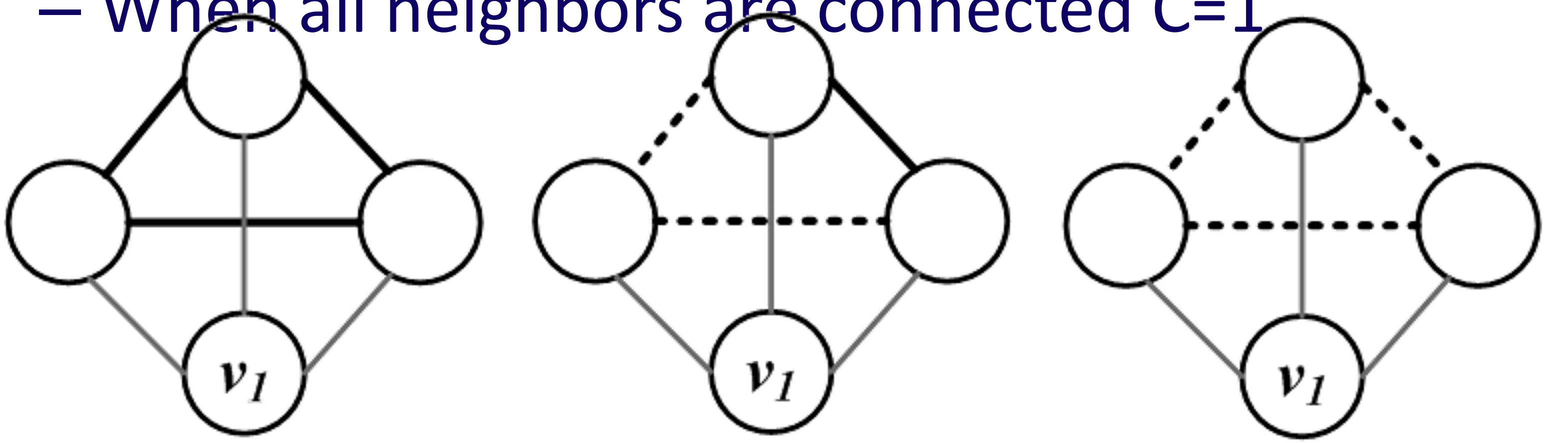
$$C(v_i) = \frac{\text{number of pairs of neighbors of } v_i \text{ that are connected}}{\text{number of pairs of neighbors of } v_i}.$$

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

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

Local Clustering Coefficient: Example

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

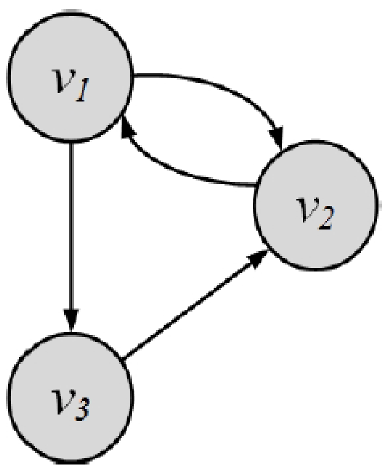


$C(v_1)=1$ $C(v_1)=1/3$ $C(v_1)=0$

Reciprocity

If you become my friend, I'll be yours

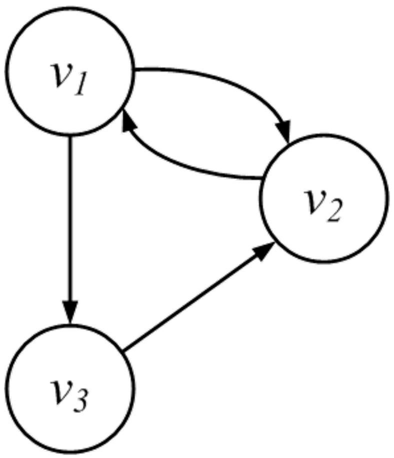
- Reciprocity is a more simplified version of transitivity as it considers closed loops of length 2
- If node v is connected to node u, u by connecting to v, exhibits reciprocity



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

$$Trace(A) = A_{1,1} + A_{2,2} + \ldots + A_{n,n} = \sum_{i=1}^{n} A_{i,i}$$

Reciprocity: Example



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

Reciprocal nodes: v_1, v_2

$$\Downarrow$$

$$R = \frac{1}{m} Trace(A^2) = \frac{2}{4} = \frac{1}{2}$$

Today’s lecture

- Centrality
- Transitivity and Reciprocity
- Balance and Status
- Similarity

Balance and Status

- Measuring stability based on an observed network

Social Balance Theory

- Social balance theory discusses consistency in friend/foe relationships among individuals. Informally, social balance theory says friend/foe relationships are consistent when

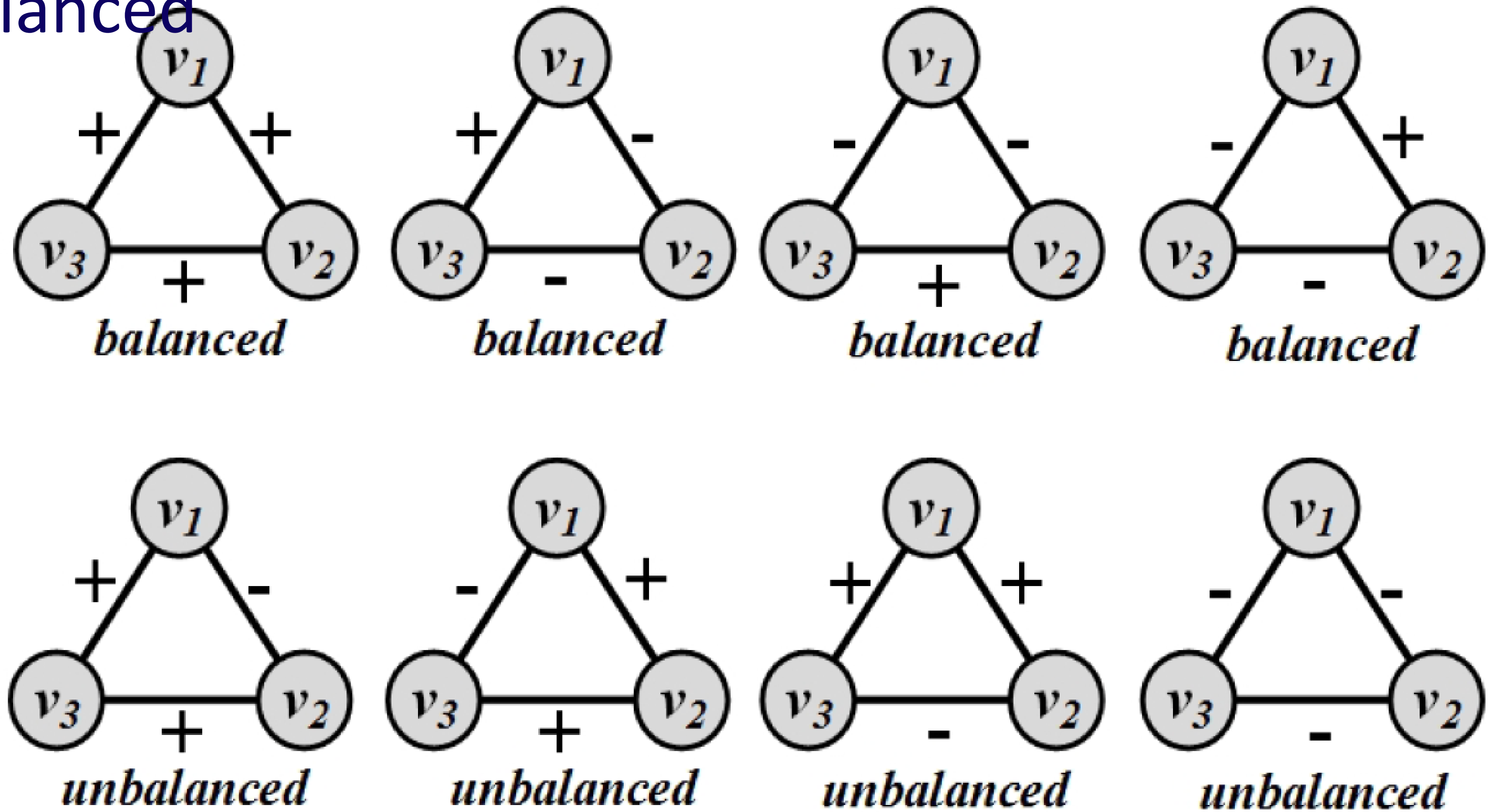
  The friend of my friend is my friend,
  The friend of my enemy is my enemy,
  The enemy of my enemy is my friend,
  The enemy of my friend is my enemy.

- 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



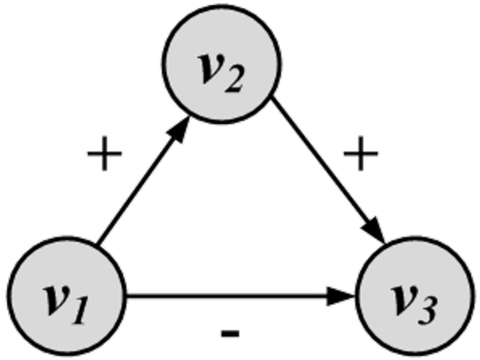
Social Status Theory

- Status defines how prestigious an individual is ranked within a society
- Social status theory measures how consistent individuals are in assigning status to their neighbors

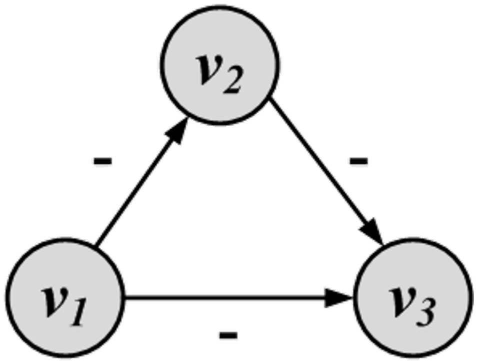
If X has a higher status than Y and Y has a higher status than Z, then X should have a higher status than Z.

Social Status Theory: Example

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



Unstable configuration



Stable configuration

Today's lecture

- Centrality
- Transitivity and Reciprocity
- Balance and Status
- Similarity

Structural Equivalence

- In structural equivalence, we look at the neighborhood shared by two nodes; the size of this neighborhood defines how similar two nodes are.

For instance, two brothers have in common sisters, mother, father, grandparents, etc. This shows that they are similar, whereas two random male or female individuals do not have much in common and are not similar.

Structural Equivalence: Definitions

- Vertex similarity

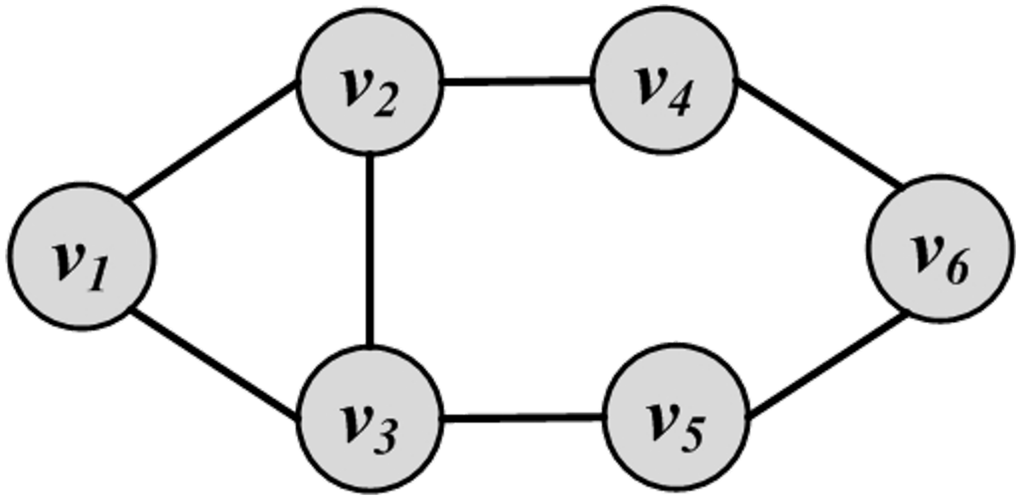
$$\sigma(v_i, v_j) = |N(v_i) \cap N(v_j)|.$$

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

- In general, the definition of neighborhood N(v) excludes the node itself v.
  - Nodes that are connected and do not share a neighbor will be assigned zero similarity
  - This can be rectified by assuming nodes to be included in their neighborhoods

Similarity: Example



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

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

Normalized Similarity

Comparing the calculated similarity value with its expected value where vertices pick their neighbors at random

- For vertices i and j with degrees d_i and d_j this expectation is $d_i d_j / n$

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

Normalized Similarity, cont.

$$\bar{A}_i = \frac{1}{n}\sum_k A_{i,k}$$

$$\begin{aligned}
\sigma_{expected}(v_i, v_j) &= \sum_k A_{i,k}A_{j,k} - \frac{d_i d_j}{n} \\
&= \sum_k A_{i,k}A_{j,k} - n\frac{1}{n}\sum_k A_{i,k}\frac{1}{n}\sum_k A_{j,k} \\
&= \sum_k A_{i,k}A_{j,k} - n\bar{A}_i\bar{A}_j \\
&= \sum_k (A_{i,k}A_{j,k} - \bar{A}_i\bar{A}_j) \\
&= \sum_k (A_{i,k}A_{j,k} - \bar{A}_i\bar{A}_j - \bar{A}_i\bar{A}_j + \bar{A}_i\bar{A}_j) \\
&= \sum_k (A_{i,k}A_{j,k} - A_{i,k}\bar{A}_j - \bar{A}_i A_{j,k} + \bar{A}_i\bar{A}_j) \\
&= \sum_k (A_{i,k} - \bar{A}_i)(A_{j,k} - \bar{A}_j),
\end{aligned}$$

Normalized Similarity, cont.

Covariance between A_i and A_j

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

Covariance can be normalized by the multiplication of Variances:

Pearson correlation coefficient:

$\sigma \in [-1,1]$

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

Regular Equivalence

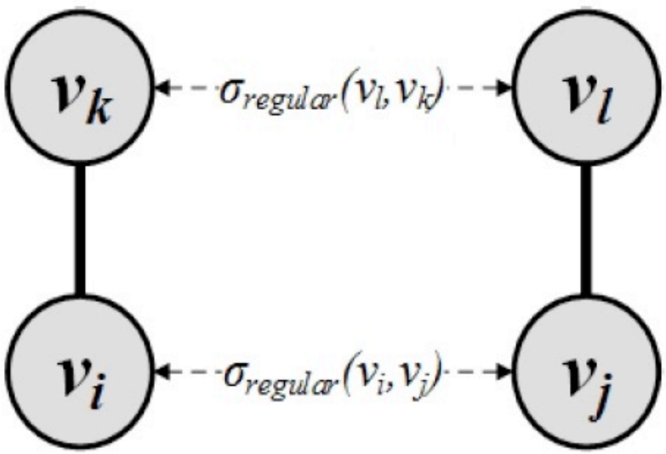
- In regular equivalence, we do not look at neighborhoods shared between individuals, but how neighborhoods themselves are similar

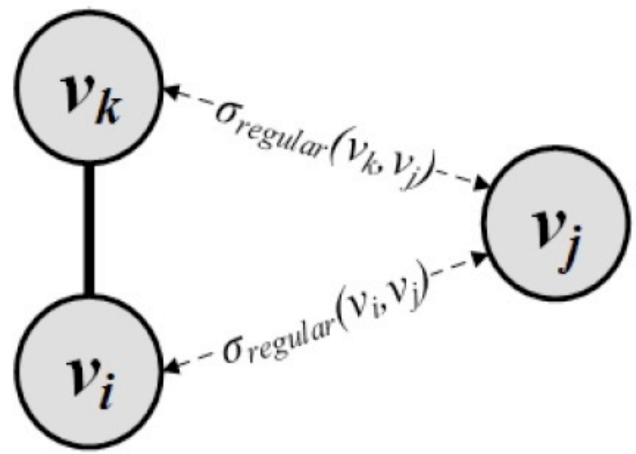
For instance, athletes are similar not because they know each other in person, but since they know similar individuals, such as coaches, trainers, other players, etc.

Regular Equivalence

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

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





- 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, v_j are similar when v_j is similar to ~~v_i’s~~ neighbors v_k

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

$$\Downarrow$$

- 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 $\Rightarrow$

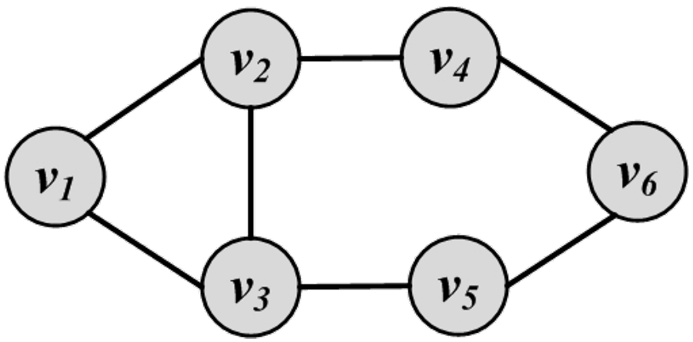
$$\sigma_{regular} = \alpha A \sigma_{Regular} + \mathbf{I}$$

$$\Downarrow$$

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

Regular Equivalence: Example

- Any row/column of this matrix shows the similarity to other vertices
- We can see that vertex 1 is most similar (other than itself) to vertices 2 and 3
- Nodes 2 and 3 have the highest similarity



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

The largest eigenvalue of A is 2.43

Set $\alpha = 0.4 < 1/2.43$

$$\sigma_{regular} = (I - 0.4A)^{-1} = \begin{bmatrix} 1.43 & 0.73 & 0.73 & 0.26 & 0.26 & 0.16 \\ 0.73 & 1.63 & 0.80 & 0.56 & 0.32 & 0.26 \\ 0.73 & 0.80 & 1.63 & 0.32 & 0.56 & 0.26 \\ 0.26 & 0.56 & 0.32 & 1.31 & 0.23 & 0.46 \\ 0.26 & 0.32 & 0.56 & 0.23 & 1.31 & 0.46 \\ 0.16 & 0.26 & 0.26 & 0.46 & 0.46 & 1.27 \end{bmatrix}$$