Big Data on Social Media Mining and Analytics Graph -Network Measures

March 27, 2015

课堂寻人搜索引擎

周可人

Advanced search
Language tools

-将翘课的学生一网打尽
-挖掘有潜力的学生
…………………………

Assignment Show

Mining Social Network: Homework #1

Due on March 22, 2015 at 5:00pm

Professor Hao Wang 712066H

Keren Zhou

201428013229070

1

Keren Zhou   Mining Social Network (Professor Hao Wang 712066H): Homework #1   Problem 1

Problem 1

Proof: In any directed graph, the summation of in-degree is equal to the summation of out-degrees.

$$\sum d^{out} = \sum d^{in}$$

Solution

We prove it by contradiction.

Proof. Suppose that there's one graph that the summation of in-degree is not equal to the summation of out-degrees.
In any directed graph, a edge contributes one unit to all in-degrees and one unit to all out-degrees. Therefore the sum of in-degrees must be equal to out-degrees, which is on the contrary to the supposition. □

2

Keren Zhou   Mining Social Network (Professor Hao Wang 712066H): Homework #1   Problem 2

Problem 2

Algorithm: Please design an algorithm to detect all the bridges in the graph.

Solution

I use the *Tarjan* algorithm to solve this problem.
The key idea inside this algorithm is to attach every node in the dfs-search-tree two *timestamps*, which indicate the search order. In this way, if the sub-tree of a given node cannot reach any node whose tag is less than, we claim that this node is a cutting node. Furthermore, if the sub-tree could only reach the given node, we claim that the ancestor link of the given node is a bridge.

The whole algorithm is as follow:

```
function DFS(u, fa)
    lowu = pre[u] = ++dfs_clock
    child = 0
    for i = 0 to g[u].size() do
        v = g[u][i]
        if !pre[v] then
            ++child
            lowv = dfs(v, u)
            lowu = min(lowu, lowv)
        else if pre[v] < pre[u] and v! = fa then
            lowu = min(lowu, pre[v])
        end if
    end for
    low[u] = lowu
    returnlowu
end function
for i = 0; i < nodes; ++i do
    if i! = rootandpre[i] == low[i] then
        nbridges++
    end if
end for
```

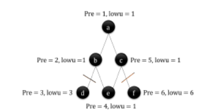


Figure 1: bridge

The figure shows that the red lines cut through the bridges, so that there are two *bridges*. Notice that the root couldn't make any bridge.

3

Keren Zhou   Mining Social Network (Professor Hao Wang 712066H): Homework #1   Problem 3

Problem 3

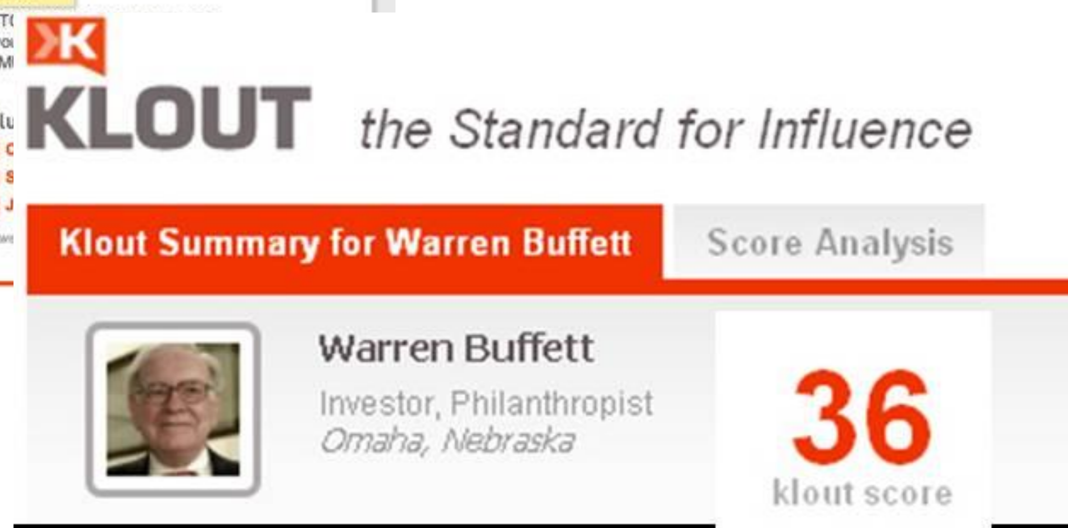
Algorithm: Please present Floyd-Warshall algorithm.

Solution

```
First initialize all the value to Infinite
function FLOYD(n)
    for k = 1 to n do
        for i = 1 to n do
            for j = 1 to n do
                if g[i][k] + g[k][j] < g[i][j] then
                    g[i][j] = g[i][k] + g[k][j]
                end if
            end for
        end for
    end for
end function
```

4

Klout



Klout is a website and mobile app that uses social media analysis to rank its users according to online social influence via the “Klout Score”, which is a numerical value between 1 and 100.

必应影响力

必应影响力分数是根据多个社交网络、搜索引擎和媒体网站的数据，以科学的方式计算产生的。一个人在社交网站上的粉丝互动量、在搜索引擎上的被搜索量和在媒体网站上的浏览量都是其影响力分数的重要构成因素。

Why Do We Need Measures?

- Who are the central figures (influential individuals) in the network?
- What interaction patterns are common in friends?
- Who are the *like-minded* users and how can we find these similar individuals?

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

Centrality

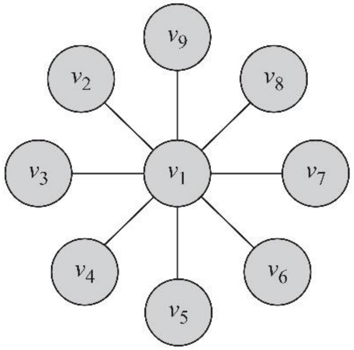
Centrality defines how important a node is within a network.

Degree Centrality

In real-world interactions, we often consider people with many connections to be important. Degree centrality transfers the same idea into a measure.

$$C_d(v_i) = d_i$$

d_i is the degree (number of adjacent edges) for vertex v_i



Undirected graph

In this graph degree centrality for vertex 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:

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

When using in-degrees, degree centrality measures how popular a node is and its value shows *prominence* or *prestige*.

When using out-degrees, it measures the *gregariousness* of a node.

Normalized Degree Centrality

The degree centrality measure does not allow for centrality values to be compared across networks (e.g., Facebook and Twitter).

To overcome this problem, we can normalize the degree centrality values.

- Normalized by the *maximum possible degree*

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

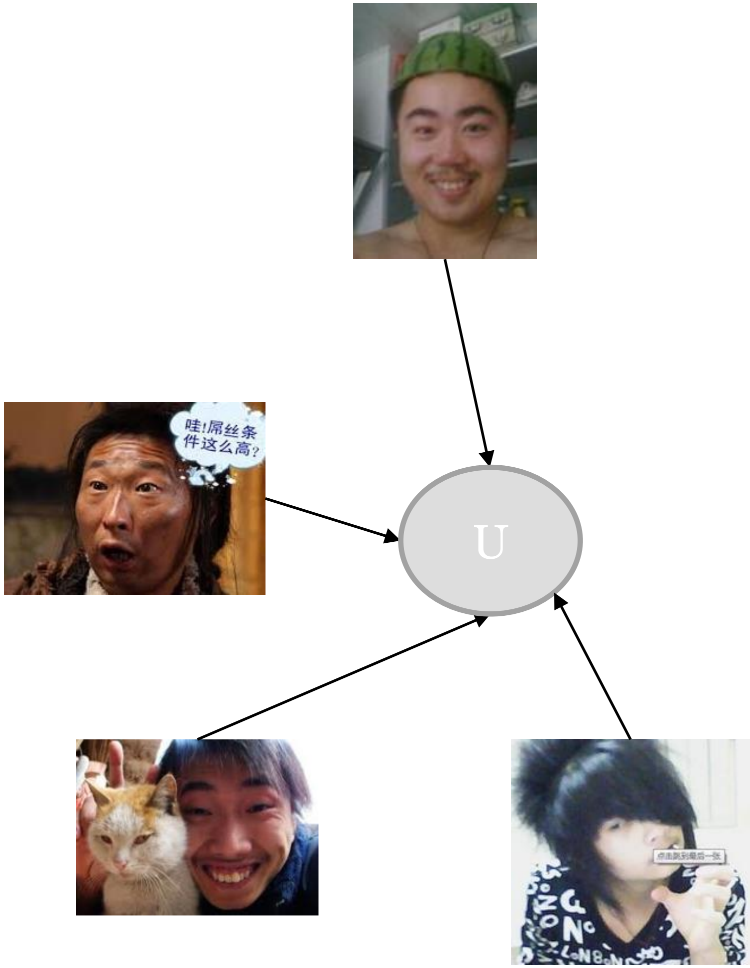
- Normalized by the *maximum degree*

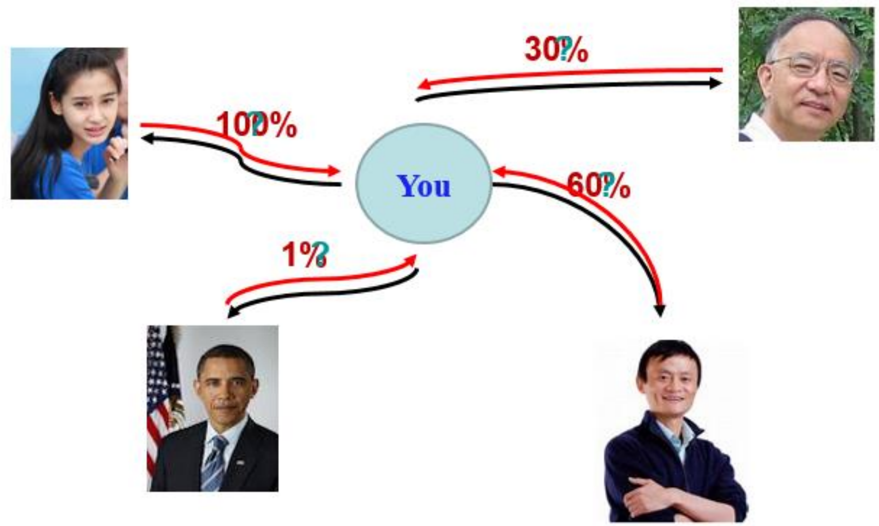
$$C_d^{max}(v_i) = \frac{d_i}{\max_j d_j}$$

- Normalized by the *degree sum*

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

Problem





Eigenvector Centrality

- Having more friends does not by itself guarantee that someone is more important, but *having more **important friends*** provides a stronger signal.
- Eigenvector centrality tries to generalize degree centrality by incorporating the importance of the neighbors (or incoming neighbors in directed graphs).
- To keep track of neighbors, we can use the adjacency matrix A of a graph.

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

$C_e(v_i)$: the eigenvector centrality of node v_i

λ: some fixed constant

Eigenvector Centrality, cont.

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

  ➔ $\lambda \mathbf{C}_e = A^T \mathbf{C}_e.$

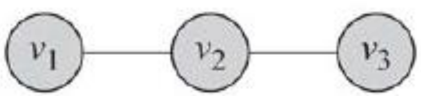
- This means that C_e is an eigenvector of adjacency matrix A^T and λ is the corresponding eigenvalue

- Which *eigenvalue-eigenvector pair* should we choose?

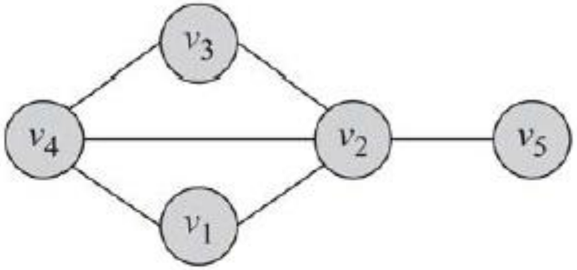
Eigenvector Centrality, cont.

Theorem 3.1 (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 $\forall v_i > 0$.*

Therefore, to have positive centrality values, we can compute the eigenvalues of A and then select the largest eigenvalue. The corresponding eigenvector is C_e.



(a) A three node graph



(b) A five node graph

Eigenvector Centrality Example

Example 3.2. *For the graph shown in Figure 3.2(a), the adjacency matrix is*

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

(a) A three node graph

Based on Equation $\lambda \mathbf{C}_e = A^T \mathbf{C}_e$. *solve* $\lambda \mathbf{C}_e = A\mathbf{C}_e$, *or*

$$(A - \lambda I)\mathbf{C}_e = 0.$$

Assuming $\mathbf{C}_e = [u_1\ u_2\ u_3]^T$,

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

Since $\mathbf{C}_e \neq [0\ 0\ 0]^T$, *the characteristic equation is*

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

or equivalently,

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

Eigenvector Centrality Example

So the eigenvalues are $(-\sqrt{2}, 0, +\sqrt{2})$. *We select the largest eigenvalue:* $\sqrt{2}$.
We compute the corresponding eigenvector:

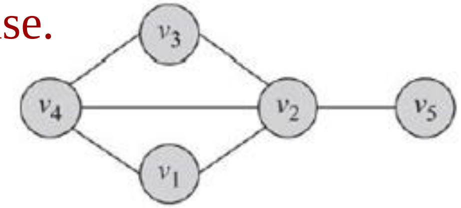
$$\begin{bmatrix} 0-\sqrt{2} & 1 & 0 \\ 1 & 0-\sqrt{2} & 1 \\ 0 & 1 & 0-\sqrt{2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Assuming $\mathbf{C}_e$ *vector has norm 1, its solution is*

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

which denotes that node v_2 *is the most central node and nodes* v_1 *and* v_3 *have equal centrality values.*

Left as an exercise.



(b) A five node graph

Katz Centrality

- A major problem with eigenvector centrality arises when it deals with directed graphs

Come up with an example of a directed connected graph in which eigenvector centrality becomes zero for some nodes. Describe when this happens.

- Centrality only passes over *outgoing* edges and in special cases such as when a node is in a weakly connected component centrality becomes zero even though 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

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

Katz Centrality, cont.

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

Controlling term

Bias term

Rewriting equation in a vector form

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

vector of all 1's

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

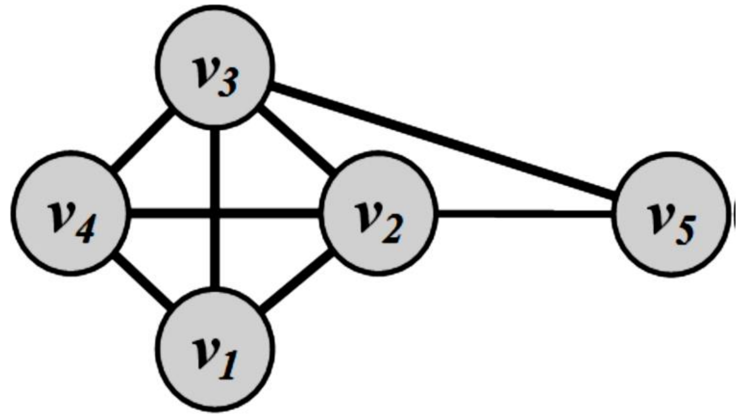
Katz Centrality, cont.

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

- When $\alpha=0$, 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
  - $\det(I- \alpha A^T)$!=0
  - By rearranging we get $\det(A^T- \alpha^{-1}I)=0$
  - This is basically the characteristic equation, which first becomes zero when the largest eigenvalue equals α^{-1} or equivalently $\alpha =1/ \lambda$.
- 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, 0.35, 3.32
- We assume α=0.25 < 1/3.32 β=0.2

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

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
- To mitigate this problem we can divide the value of passed centrality by the number of outgoing links, i.e., out-degree of that node such that each connected neighbor gets a fraction of the source node's centrality

PageRank, cont.

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

$$\Downarrow$$

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

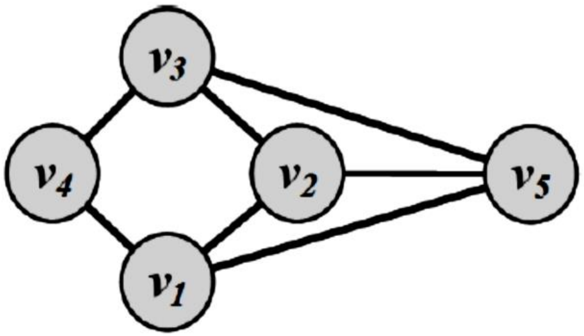
$$\Downarrow$$

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

When dj_{out} = 0, we know that since the out-degree is zero, Vi, $A_{j,i}$ = 0. This makes the term inside the summation 0/0. We can fix this problem by setting dj_{out} = 1 since the node will not contribute any centrality to any other nodes.

PageRank Example

- We assume α=0.95 and β=0.1



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

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

Betweenness Centrality

Another way of looking at centrality is by considering how important nodes are in connecting other nodes. One approach, for a node v_i, is to compute the number of shortest paths between other nodes that pass through v_i,

Brandes' algorithm(shortest paths)

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

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

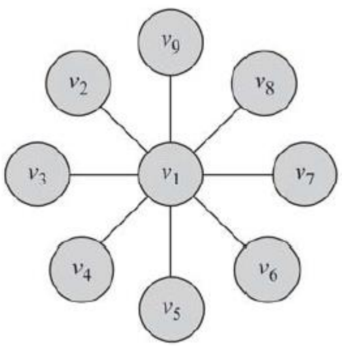
$\sigma_{st}(v_i)$ the number of shortest paths from s to t that pass through v_i

Normalizing Betweenness Centrality

In the best case, node v_i is on all shortest paths from s to t, hence,

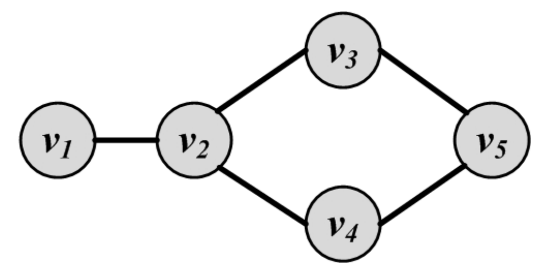
$$\frac{\sigma_{st}(v_i)}{\sigma_{st}} = 1$$

Betweenness centrality needs to be normalized to be comparable *across networks*.



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

Therefore, the maximum value is $2\binom{n-1}{2}$



The betweenness can be divided by its maximum value to obtain the normalized betweenness.

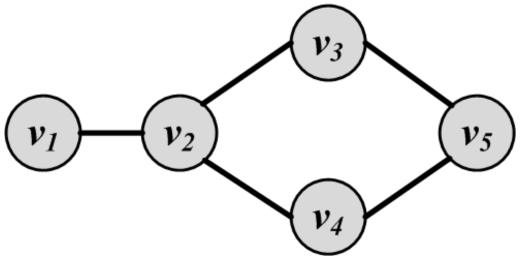
$$C_b^{\text{norm}}(v_i) = \frac{C_b(v_i)}{2\binom{n-1}{2}}.$$

Betweenness Centrality Example

Example *Figure 3.5 depicts a sample graph. In this graph, the betweenness centrality for node v_1 is 0, since no shortest path passes through it. For other nodes, we have*

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

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



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

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

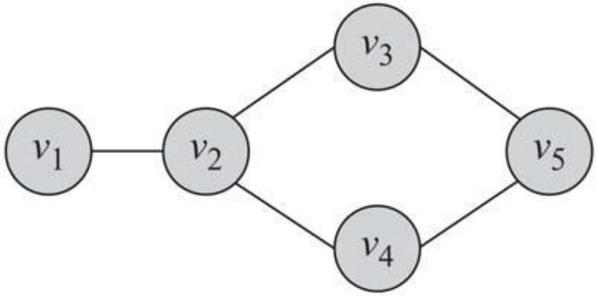
where centralities are multiplied by 2 because in an undirected graph $\sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}} = 2 \sum_{s \neq t \neq v_i, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$.

Closeness Centrality

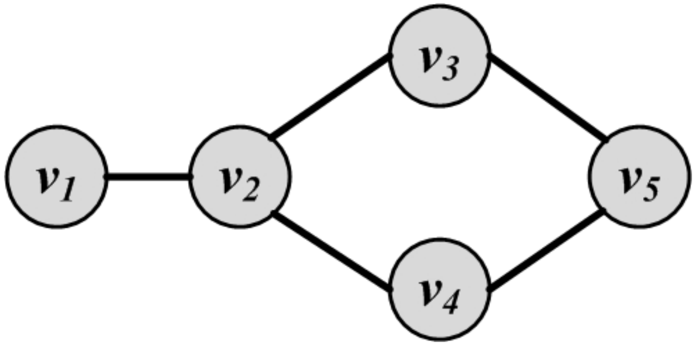
- The intuition is that influential and central nodes can quickly reach other nodes
- These nodes should have a smaller average shortest path length to other nodes

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

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



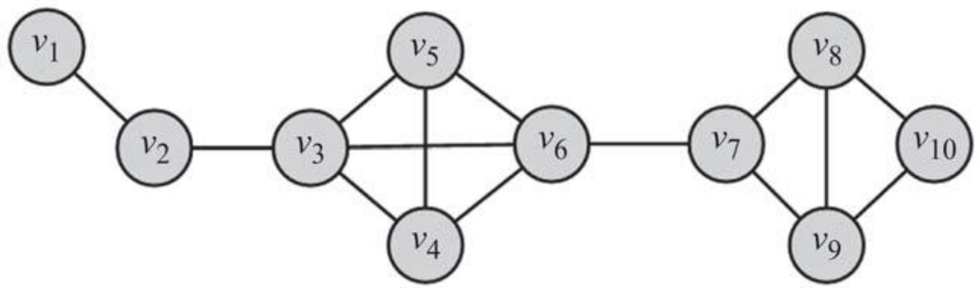
Compute Closeness Centrality



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

Hence, node v_2 has the highest closeness centrality.

Homework



Compute the top three central nodes based on *degree*, *eigenvector*, *Katz*(Alpha = Beta = 0.3), *PageRank*, *betweenness*, and *closeness* centrality methods.

	First Node	Second Node	Third Node
Degree Centrality			
Eigenvector Centrality			
Katz Centrality: $\alpha = \beta = 0.3$			
PageRank: $\alpha = \beta = 0.3$			
Betweenness Centrality			
Closeness Centrality			

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

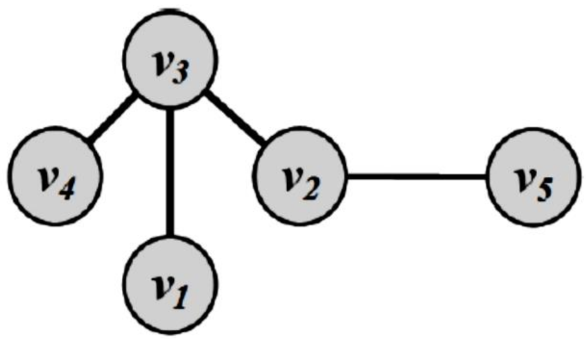
$$\bar{l}_S^{group} = \frac{1}{|V-S|} \sum_{v_j \notin S} l_{S,v_j}.$$

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

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

Group Centrality Example

- Consider S={v2,v3}

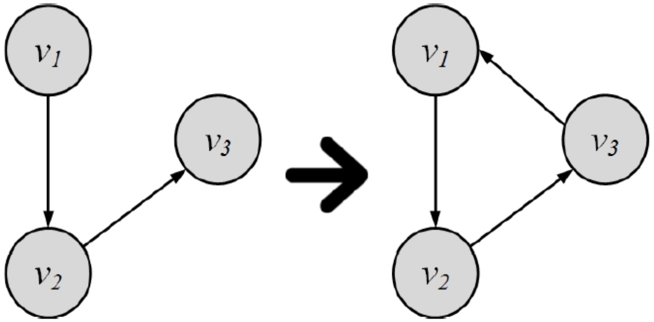


- Group degree centrality=3
- Group betweenness centrality = 3
- Group closeness centrality = 1

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

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

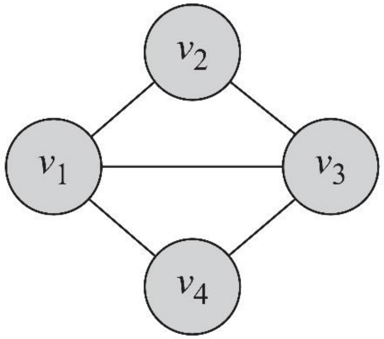
When counting triangles, since every triangle has 6 closed paths of length 2:

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

In undirected networks:

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

[Global] Clustering Coefficient: Example



$$C = \frac{(Number\ of\ Triangles) \times 3}{Number\ of\ Connected\ Triples\ of\ Nodes} = \frac{2 \times 3}{2 \times 3 + \underbrace{2}_{v_2 v_1 v_4,\ \ v_2 v_3 v_4}} = 0.75.$$

Local Clustering Coefficient

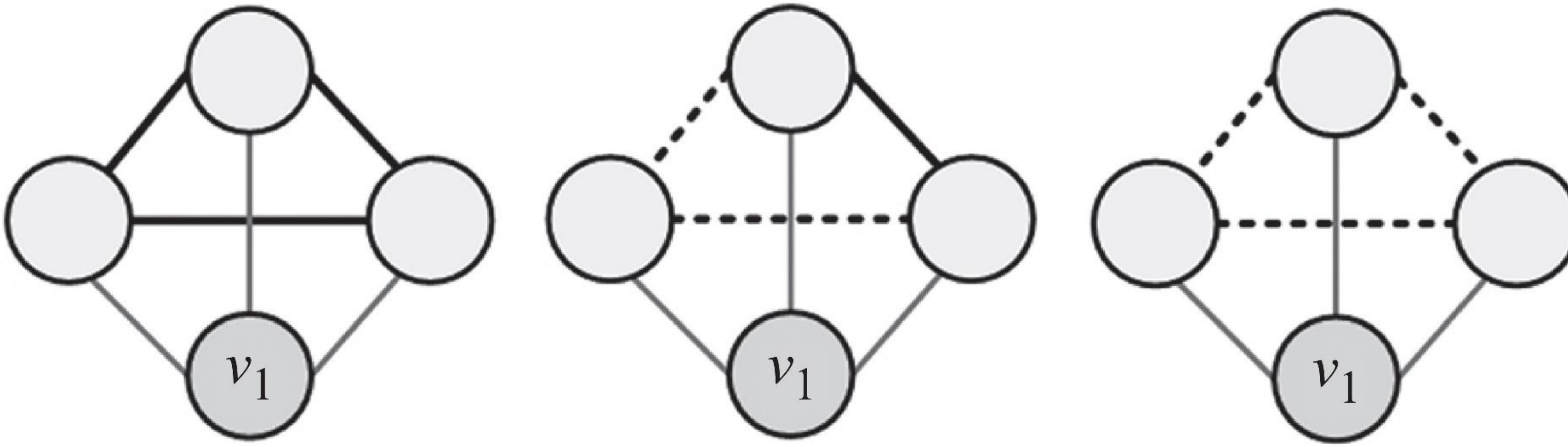
- Local clustering coefficient measures transitivity at the node level
- Commonly employed for undirected graphs, it computes how strongly neighbors of a node v (nodes adjacent to v) are themselves connected

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

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

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

Local Clustering Coefficient: Example



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

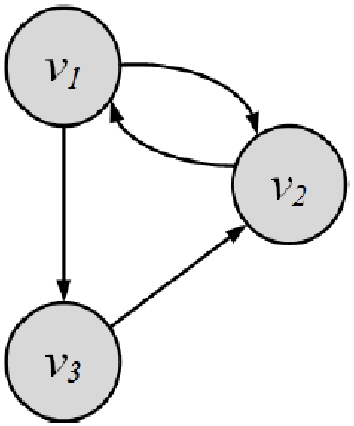
- Thin lines depict connections to neighbors
- Dashed lines are the missing connections among neighbors
- Solid lines indicate connected neighbors
  - When none of neighbors are connected C=0
  - When all neighbors are connected C=1

Reciprocity

If you become my friend, I'll be yours

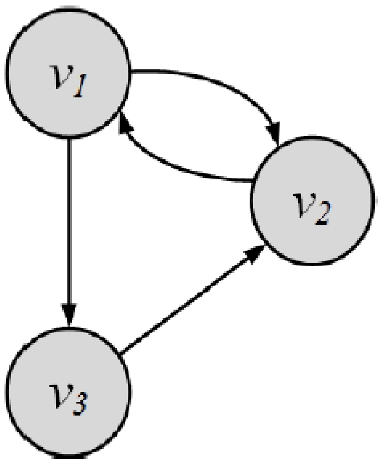
- Reciprocity is a more simplified version of transitivity as it considers closed loops of length 2
- If node v is connected to node u, u by connecting to v, exhibits reciprocity

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



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

Reciprocity: Example



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

Reciprocal nodes: v_1, v_2

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

Balance and Status

- Assume we observe a signed graph that represents friends/foes or social status.
- Can we measure the consistency of attitudes that individual have toward one another?

Social Balance Theory （structural balance theory）

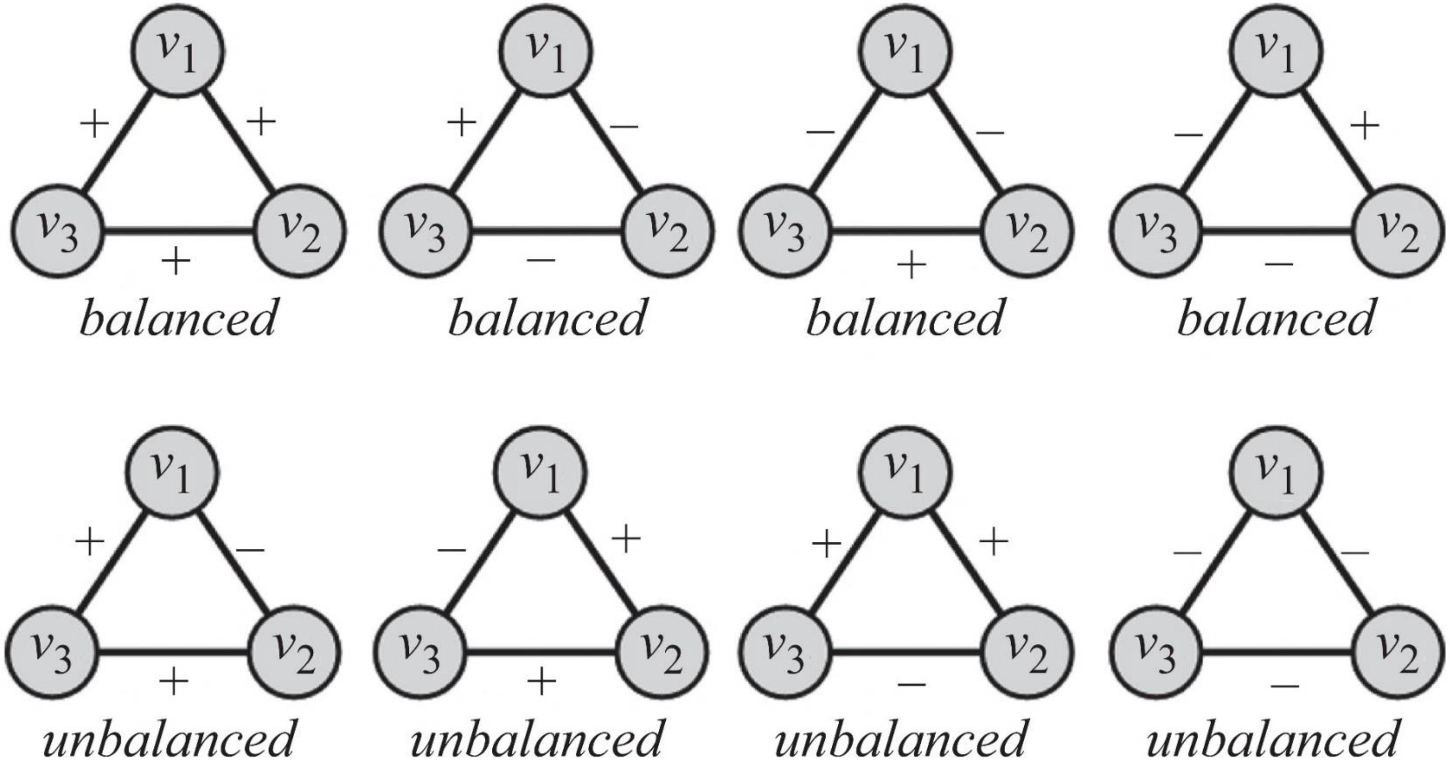
- Social balance theory discusses consistency in friend/foe relationships among individuals. Informally, social balance theory says friend/foe relationships are consistent when

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



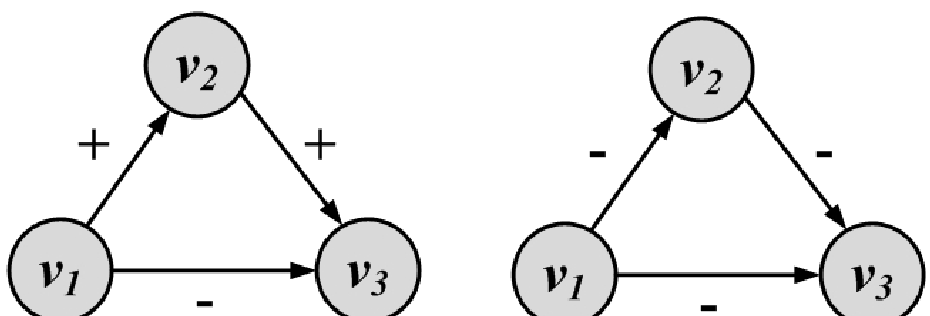
Sample Graphs for Social Balance Theory. In balanced triangles, there are an even number of negative edges.

- For any cycle if the multiplication (product) of edge values become positive, then the cycle is socially balanced.
- Social balance can also be generalized to subgraphs other than triangles.

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.



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

network similarity content similarity

- In social media, these nodes can represent individuals in a friendship network or products that are related.
- How similar are two nodes in a network?

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

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

Normalization Procedure

Jaccard Similarity:

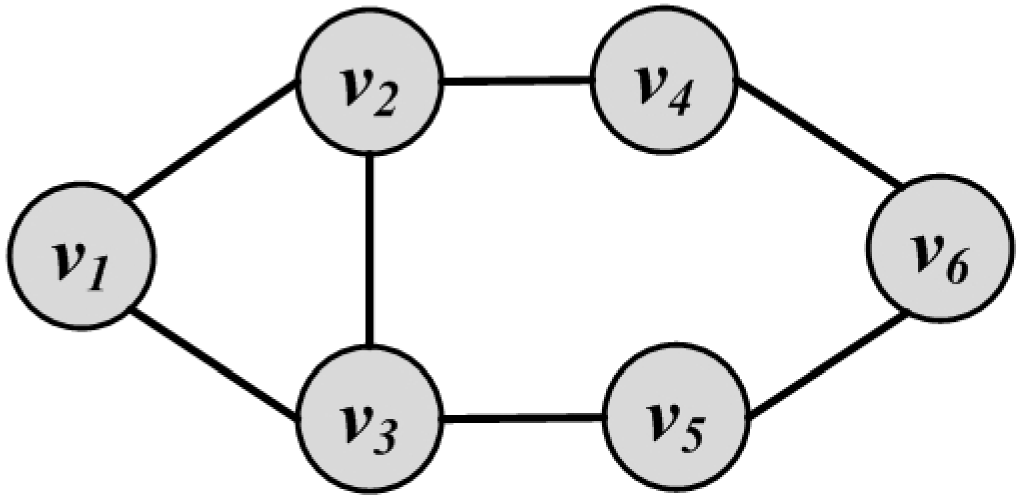
$$\sigma_{Jaccard}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|},$$

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



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

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

Structural Equivalence

A more interesting way of measuring the similarity between v_i and v_j is to compare $\sigma(v_i, v_j)$ with the expected value of $\sigma(v_i, v_j)$ when nodes pick their neighbors at random. The more distant these two values are, the more significant the similarity observed between v_i and v_j ($\sigma(v_i, v_j)$) is. For nodes v_i and v_j with degrees d_i and d_j, this expectation is $\frac{d_i d_j}{n}$, where n is the number of nodes. This is because there is a $\frac{d_i}{n}$ chance of becoming v_i's neighbor and, since v_j selects d_j neighbors, the expected overlap is $\frac{d_i d_j}{n}$. We can rewrite $\sigma(v_i, v_j)$ as

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

Structural Equivalence

Hence, a similarity measure can be defined by subtracting the random expectation $\frac{d_i d_j}{n}$ from Equation :

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

Structural Equivalence

where $\bar{A}_i = \frac{1}{n}\sum_k A_{i,k}$. The term $\sum_k (A_{i,k} - \bar{A}_i)(A_{j,k} - \bar{A}_j)$ is basically the covariance between A_i and A_j. The covariance can be normalized by the multiplication of variances,

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

which is called the *Pearson correlation coefficient*. Its value, unlike the other two measures, is in the range $[-1, 1]$. A positive correlation value denotes that when v_i befriends an individual v_k, v_j is also likely to befriend v_k. A negative value denotes the opposite (i.e., when v_i befriends v_k, it is unlikely for v_j to befriend v_k). A zero value denotes that there is no linear relationship between the befriending behavior of v_i and v_j.

Regular Equivalence

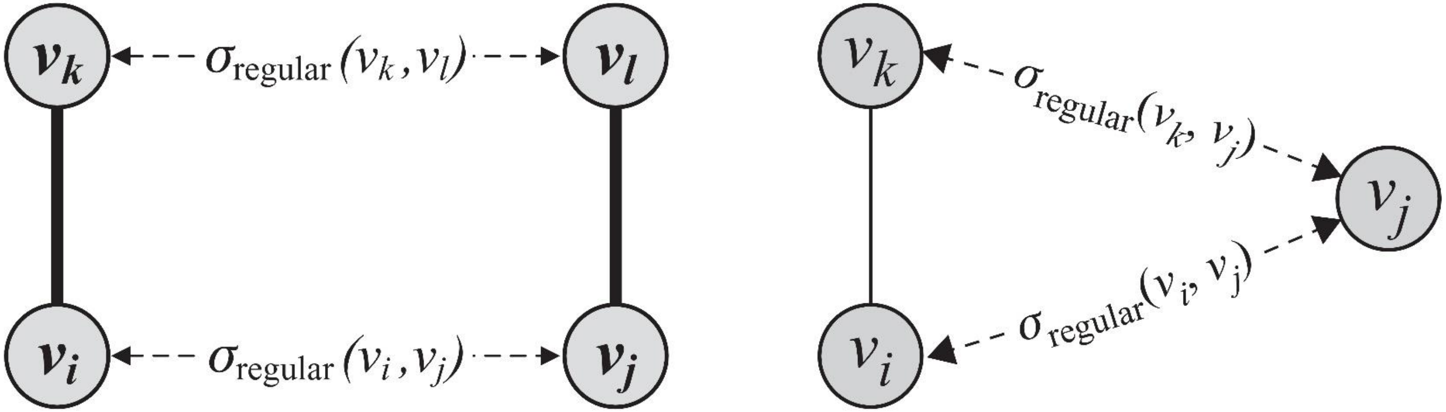
- In regular equivalence, we do not look at neighborhoods shared between individuals, but how neighborhoods themselves are similar

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

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



(a) Original Formulation (b) Relaxed Formulation

- 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

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

- 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

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

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

Regular Equivalence

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

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

- In vector format

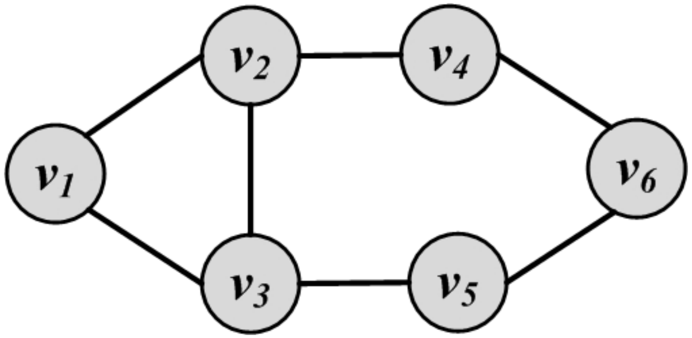
$$\sigma_{regular} = \alpha A \sigma_{Regular}$$

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

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

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

Regular Equivalence: Example



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

The largest eigenvalue of A is 2.43

Set $\alpha = 0.4 < 1/2.43$

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

- 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

Summary

We discussed measures for a social media network.

Centrality measures attempt to find the most central node within a graph.

Linking between nodes (e.g., befriending in social media) is the most commonly observed phenomenon in social media. (transitivity and its reciprocity)

To analyze if relationships are consistent in social media, we used various social theories (social balance and social status) to validate outcomes.

Finally, we analyzed node similarity measures (structural equivalence and regular equivalence).