Network Science (VU) (706.703)
Measuring Network Properties

Denis Helic
ISDS, TU Graz

November 20, 2017
Outline

1. Introduction
2. Centrality
3. Clustering and Reciprocity
4. Similarity
5. Homophily
Once when we know the structure of the network we can calculate many useful quantities.

Such network analysis originates from social network analysis.

Mostly these ideas reflect some sociological concepts, such as influence, status, balance, ...

However, today many of the methods from social network analysis are applied in computer science, physics, biology, and so on.
One of the key topics in network science is *centrality*. What are the most central nodes in a network? What are the most important nodes in a network? What are the most influential nodes in a network?
In different kind of networks different interpretation are possible

E.g. in a social network the most central node might be the most popular person

E.g. on the Web the most central node might be a page with the best quality of content in a specific field

E.g. on the Internet the most central node might be a router with the highest bandwidth

Thus, there are many possible definitions of importance and many possible interpretations and therefore there are many centrality measures
Figure: Sample network
A = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\end{pmatrix} \quad (1)
Degree centrality

- The simplest centrality measure is just the degree of a node
- In directed networks nodes have in- and out-degree and therefore there are two types of degree centrality
- In social networks persons that have high degree centrality might have better prestige, influence, access to information, ...
- In citation networks papers that have high in-degree centrality have a lot of citations
- This is a widely used metric for measuring the scientific impact of a paper
Degree centrality

- However, in many cases simple degree centrality is not enough.
- E.g. a popular actor might have a high in-degree centrality, but is this a good proxy for measuring influence?
- Sometimes not only the number of links counts but also who are the neighbor nodes.
Eigenvector centrality

- A natural extension of the degree centrality
- Degree centrality awards one centrality point for every neighbor a node has
- However, not all neighbors are equally important
- In many cases the importance of the node is increased by having connections to other nodes that are themselves important
Eigenvector centrality

- Basic concept of eigenvector centrality: not only count of neighbors is important but also the importance of the neighbors
- Degree centrality awards nodes with one centrality point for each neighbor
- Eigenvector centrality gives each node a score proportional to the sum of the scores of its neighbors
- Typically, we calculate eigenvector centralities iteratively
Eigenvector centrality

- We make an initial guess about the centrality $x_i$ of each node $i$.
- E.g. we set $x_i^0 = 1$ for all $i$.
- Then we calculate a new iteration $x_i^1$ as the sum of the centralities of $i$’s neighbors.

\begin{equation}
    x_i^1 = \sum_j A_{ij}x_j^0
\end{equation}

\begin{equation}
    x^1 = Ax^0
\end{equation}
Eigenvector centrality

- In matrix form we have

\[ x^1 = Ax^0 \]  

(4)

- After \( t \) steps we have

\[ x^t = A^t x^0 \]  

(5)
Eigenvector centrality

- We can write $x^0$ as a linear combination of the eigenvectors $v_i$ of the adjacency matrix (for appropriate choice of constants $c_i$)

$$x^0 = \sum_i c_i v_i \quad (6)$$

$$x^t = A^t \sum_i c_i v_i = \sum_i c_i A^t v_i = \sum_i c_i \kappa_i^t v_i = \kappa_1^t \sum_i c_i \left[ \frac{\kappa_i}{\kappa_1} \right]^t v_i \quad (7)$$
Eigenvector centrality

\[ x^t = \kappa_1^t \sum_i c_i \left[ \frac{\kappa_i}{\kappa_1} \right]^t v_i \] (8)

- \( \kappa_i \) are eigenvalues, and \( \kappa_1 \) is the largest of themselves
- \( \frac{\kappa_i}{\kappa_1} < 1 \) for all \( i > 1 \)
- When \( t \to \infty \), \( \frac{\kappa_i}{\kappa_1} \to 0 \), for all \( i > 1 \)
- When \( t \to \infty \), \( x^t \to c_1 \kappa_1^t v_1 \)
Eigenvector centrality

- In other words, the limiting vector of centralities is proportional to the leading eigenvector of the adjacency matrix.
- In matrix form the centrality vector $x$ satisfies:

\[
Ax = \kappa_1 x
\]  \hspace{1cm} (9)

\[
x_i = \frac{1}{\kappa_1} \sum_j A_{ij} x_j
\]  \hspace{1cm} (10)
Thus, eigenvector centrality of a node can be large if a node has many neighbors or if it has important neighbors, or both. If a person knows a lot of people (even if they are not important) or if a person knows only a few people but in high places. The eigenvector centralities are all non-negative. If we chose $x^0$ with all non-negative elements, multiplication by $A$ can never introduce negative elements since all elements in $A$ are non-negative.
Eigenvector centrality

- Centralities are not normalized
- Typically, we are interested only in relative centralities of nodes
- We want to know which are important nodes and how their importance compares to others
- Absolute values are not needed
Eigenvalue centrality

- For directed networks some complications arise
- Directed networks have an asymmetric adjacency matrix
- Asymmetric matrices have two sets of eigenvectors: the left and the right eigenvectors
- The right eigenvectors sum over in-coming links
- The left eigenvectors sum over out-going links
Eigenvector centrality

- We assume that the importance is given through links pointing to a node.
- E.g. in citation networks citations of a paper.
- E.g. on the Web links from other Web pages on a particular page.
- The right eigenvectors sum over incoming links.
- Thus, the correct definition is same as for undirected case.

\[ x_i = \frac{1}{\kappa_1} \sum_j A_{ij} x_j \]  

(11)
Figure: Sample network
Eigenvector centrality

- Nodes 7 and 1 have no incoming links
- Such nodes will always have eigenvector centrality zero
- This might be ok since they do not have incoming links
- However, node 8 has one incoming link but still centrality zero
- The link pointing to node 8 originates at node 7, and hence node 8 “inherits” centrality of node 7 which is zero
- This can propagate through many levels
Eigenvector centrality

- Only nodes that are in a strongly connected component or in its out-component can have eigenvector centralities larger than zero.
- However, even nodes in an in-component might have many incoming links and be therefore important.
- Acyclic networks have no strongly connected components and therefore all nodes have eigenvector centrality zero.
- E.g. citation networks.
Katz centrality

- One simple solution: we give each node a small amount of centrality

\[ x_i = \alpha \sum_j A_{ij} x_j + \beta \]  \hspace{1cm} (12)

- \( \alpha \) and \( \beta \) are positive constants
- The first term is the normal eigenvector centrality and the second term is the “free” centrality
Katz centrality

- Even nodes with zero in-degree still get $\beta$ centrality and can pass on this amount of centrality
- A node that has a high in-degree will always have a high centrality
- Also, nodes pointed by few other nodes with high centrality will also have a high centrality
Katz centrality

In matrix form

\[ x = \alpha Ax + \beta 1 \]  \hspace{1cm} (13)

\[ x = \beta (I - \alpha A)^{-1} 1 \]  \hspace{1cm} (14)
Typically, we do not care about absolute values, thus $\beta$ is unimportant. We set $\beta = 1$.

\[ x = (I - \alpha A)^{-1}1 \]  

(15)
Katz centrality

- The difference from standard eigenvector centrality is the free parameter $\alpha$.
- It weights the eigenvector term and the constant term.
- Before we calculate the Katz centrality we have to choose a value for $\alpha$.
- If $\alpha \to 0$ then the eigenvector term disappears and only the constant term $\beta$ remains.
- By increasing $\alpha$ the centralities increase and there is a point at which they diverge.
Katz centrality

- This happens when \((I - \alpha A)^{-1}\), i.e. when \((I - \alpha A)\) does not have an inverse
- i.e., \(det(I - \alpha A) = 0\)

\[det(A - \alpha^{-1}I) = 0\]  \(\text{(16)}\)
Katz centrality

- But this is the characteristic equation with roots $\alpha^{-1}$ and these are eigenvalues of the adjacency matrix.
- Thus, as $\alpha$ increases the determinant first becomes zero when $\alpha^{-1} = \kappa_1$.
- After that the centralities diverge, i.e. whenever the determinant becomes zero again.
- Thus, we should chose $\alpha$ less than $\frac{1}{\kappa_1}$ for the centralities to converge.
- No further suggestions on choosing a value for $\alpha$, i.e. chose it empirically.
Note on the largest eigenvalue

- Symmetric matrix has all real eigenvalues
- Eigenvectors are orthogonal and form $\mathbb{R}^n$ vector basis
- According to Perron-Frobenius theorem an irreducible non-negative matrix has a real largest eigenvalue
- Other eigenvalues might be complex but come always in complex-conjugate form
- In both cases the leading eigenvector has all non-negative values
Figure: Sample network
Katz centrality

- Inverting a matrix has $n^3$ time complexity
- For large networks this is extremely slow
- Repeating the process many times $x$ converges to a value close to the correct centrality
Katz centrality

- In each iteration step we have \( m \) multiplications as \( \mathbf{A} \) has \( m \) non-zero elements.
- Thus, the total time complexity is \( rm \), where \( r \) is the number of iterations.
- \( r \) depends on the network and \( \alpha \) and no general guidelines exist.
- Observe \( x_i \), apply thresholds, etc.
- However, for large networks iteration is much faster than inverting the matrix.
One problem with the Katz centrality

If a node with high centrality points to many others then all of these nodes get also high centrality

However, in many cases it means less if a node gets a link if it is only one of many

E.g. Yahoo has many links but not all of the Web pages included in the directory are as important as Yahoo

Better solution would be that a high centrality node passes only a fraction of its centrality to the neighbors
We can define a variation in which the centrality derived from neighbors is proportional to their centrality divided by their out-degree:

\[ x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_{out}^j} + \beta \]  

(17)

If \( k_{out}^i = 0 \) we set \( k_{out}^i = 1 \), since \( A_{ij} = 0 \) for all \( i \) and the contribution of a node without outgoing links remains zero.
PageRank

- In matrix form

\[ x = \alpha AD^{-1}x + \beta 1 \]  

- \( D \) id the diagonal matrix with elements \( D_{ii} = \max(k_i^{out}, 1) \)
PageRank

- Again, we do not care about absolute values, thus $\beta$ is unimportant
- We set $\beta = 1$

$$x = \alpha AD^{-1}x + 1$$ (19)
PageRank

Solving for $x$

$$x = (I - \alpha AD^{-1})^{-1}1 = D(D - \alpha A)^{-1}1$$  \hspace{1cm} (20)
Centrality

PageRank

- Again, PageRank has the free parameter $\alpha$
- Before we calculate PageRank we have to chose a value for $\alpha$
- By analogy with the Katz centrality, $\alpha$ should be less then inverse of the largest eigenvalue of $AD^{-1}$
- For undirected networks the largest eigenvalue is 1, thus $\alpha$ should be less than 1
- For directed networks the largest eigenvalue can be different than 1, but it is roughly of order 1
- Google uses $\alpha = 0.85$ (empirical choice)
Figure: Sample network
## Centrality

<table>
<thead>
<tr>
<th>Divide by out-degree</th>
<th>with constant term</th>
<th>without constant term</th>
</tr>
</thead>
<tbody>
<tr>
<td>x = D(D - αA)^{-1} 1</td>
<td>PageRank</td>
<td>x = AD^{-1}x</td>
</tr>
<tr>
<td>x = (I - αA)^{-1} 1</td>
<td>Katz centrality</td>
<td>x = \kappa_1^{-1}Ax</td>
</tr>
</tbody>
</table>

**Table:** Comparison of centrality measures
Centralities Notebook

- Jupyter Notebook example
- [http://kti.tugraz.at/staff/denis/courses/netsci/cent.ipynb](http://kti.tugraz.at/staff/denis/courses/netsci/cent.ipynb)
Closeness centrality

- A completely different measure is the *closeness centrality*
- It measures the average distance of a node to other nodes
- I.e. it measures the average shortest path length of a node to other nodes
- Let $d_{ij}$ be the shortest path length between nodes $i$ and $j$

$$\ell_i = \frac{1}{n} \sum_j d_{ij} \quad (21)$$
Closeness centrality

- That quantity is the average shortest path length of node $i$
- It is low for nodes that are separated by short distances from other nodes in the network
- E.g. such nodes might have better access to information, or more influence on the others in a social network
- This is however not a centrality measure since it gives low values to central nodes

$$C_i = \frac{1}{\ell_i} = \frac{n}{\sum_j d_{ij}}$$ (22)
Closeness centrality

- It is often used in network analysis, however it has some problems
- What is the dynamic range of the shortest path length in empirical networks
- Lower bound on $d_{ij}$ is 1
- Upper bound is typically $\log n$, which is e.g. 5, 6, or similar
- Thus, the range is small
Closeness centrality

- In practice the values of closeness centrality are very close to each other with differences in the trailing digits.
- Very often you have huge number of nodes with the exact same closeness centrality.
- The values are also very unstable.
- Small changes in the network structure tend to have huge impact on the closeness centralities.
Centrality

Closeness centrality

- There is another problem
- $\ell_i$ is infinite for all $i$ in a network with two or more components
- This can be solved by defining closeness centrality as the harmonic mean of distances between nodes

$$C_i = \frac{1}{n-1} \sum_{j(\neq i)} \frac{1}{d_{ij}}$$

(23)

- We have to exclude $d_{ii}$ since this is zero
Betweenness centrality

- Betweenness centrality addresses dynamic processes that can take place on a network.
- For example, suppose we have a network with something flowing around.
- E.g. messages, news, information, data packets.
- A simple assumption is that objects will flow using shortest paths.
- Then, the total number of messages that crosses a node is proportional to the number of shortest paths that each node lies on.
- This is the betweenness centrality where more central nodes are more important for the communication processes that take place on the network.
Betweenness centrality

- Let $n^i_{st}$ be 1 if node $i$ lies on a shortest path between $s$ and $t$ and zero if it does not, or there is no such path.

$$x_i = \sum_{st} n^i_{st} \quad (24)$$

- This is the case where there is only one shortest path between $s$ and $t$.
- However, if we have more than one, e.g. $g_{st}$.

$$x_i = \sum_{st} \frac{n^i_{st}}{g_{st}} \quad (25)$$
Betweenness centrality differs from the other centrality measures because it does not measure how well connected is a node. Rather it measures its position in the network, and how much a node lies “between” other nodes.

Can you imagine a node that has high betweenness centrality but all other centralities are low? I.e. it has low degree, it is on the periphery, etc.
Betweenness centrality

Figure: Node with high betweenness centrality
Betweenness centrality

- The range for betweenness centralities is rather large
- The maximum possible value for the betweenness centrality of a node is when the node lies on the shortest path between all pairs of nodes
- This occurs for the central node in a star network
Betweenness centrality

Figure: Star network
Betweenness centrality

- It lies on $n^2$ shortest paths between node pairs except for $n - 1$ paths from the peripheral nodes to themselves.
- Thus, the betweenness centrality of the central node is $n^2 - n + 1$.
- The smallest possible value of the betweenness centrality in a connected network is when a node lies only on shortest paths to or from itself.
- $n - 1$ from such a node, $n - 1$ to such a node and 1 to itself.
- Thus, the minimum is $2n - 1$. 

Denis Helic  (ISDS, TU Graz)
Betweenness centrality

- The ratio is \( \frac{n^2 - n + 1}{2n - 1} \)
- Theoretically, it is approx. \( \frac{1}{2}n \)
- Moreover, it increases with \( n \)
- Also, differences between centralities of different nodes are larger and therefore the relative order of nodes is quite stable
Centrality: Project suggestions

- Correlations between various centralities
- Time evolution of centralities
- Rank-correlations
- Comparison of ranks with null models
- Configuration model: keep the degrees but create links at random
Clustering

- In social network a very important property is *transitivity*
- If "connected by a link" is transitive that would mean that if \((u, v)\) and \((v, w)\) then \((u, w)\)
- The friend of a friend is also my friend
- Total transitivity occurs in a *clique*, i.e. in a fully connected network
More interesting is *partial transitivity*

Social networks exhibit a high degree of partial transitivity

$(u, v)$ and $(v, w)$ does not guarantee $(u, w)$

But, it makes it much more likely
Clustering

- We can quantify it in the following way
- If \((u, v)\) and \((v, w)\) then we have a path of length two: \(uvw\)
- If also \((u, w)\) then the path is closed forming a triangle
- In social networks literature this is called *closed triad*
Clustering

Figure: A closed triad
Clustering coefficient

- **Clustering coefficient** is the fraction of closed triads in the network, i.e. the fraction of paths of length two that are closed.

\[
C = \frac{\text{(number of closed paths of length two)}}{\text{(number of paths of length two)}} \quad (26)
\]

\[
C = \frac{\text{(number of triangles)} \times 6}{\text{(number of paths of length two)}} \quad (27)
\]

\[
C = \frac{\text{(number of triangles)} \times 3}{\text{(number of connected triples)}} \quad (28)
\]
Clustering

- Social networks tend to have quite high values of the clustering coefficient
- E.g. actor collaborations $C = 0.2$, e-mail communication $C = 0.16$, etc.
- Technological and biological networks have smaller values of the clustering coefficient
- E.g. the Internet $C = 0.01$
Clustering

- What does it mean that these values for social networks are high?
- Clustering coefficient is the probability that two of my friends are also friends
- E.g. in a random network with the average degree $c$ and $n$ nodes, this probability is $\frac{c}{n}$
- E.g. for film actors this is 0.0003, for e-mail communication 0.00002
- Thus, the measured clustering coefficient is much larger than the estimation based on random network connections
Local clustering

- Clustering for a single node

\[ C_i = \frac{\text{(number of connected pairs of neighbors of } i)}{\text{(number of pairs of neighbors of } i)} \]  

(29)
Local clustering

- The number of pairs of neighbors of $i$ equals $\frac{1}{2}k_i(k_i - 1)$
- It is the average probability that a pair of $i$’s friends are friends with each other
- Local clustering depends on degree
- In most networks nodes with higher degrees have lower values of the local clustering coefficient
Local clustering

- The local clustering coefficient measures the existence of *structural holes*
- A lower value of the local clustering coefficient means that a lot of expected links between $i$’s friends is actually missing
- Such structural holes improve the importance of the central node $i$
- E.g. the communication between friends can be controlled by $i$
Local clustering

- It is a local version of the betweenness centrality.
- It measures the importance of the node for local communication whereas the betweenness centrality measures this at the global level.
- Also, more central nodes have lower values for the local clustering coefficient.
- In practice, betweenness and local clustering are strongly correlated (Burt, *Structural Holes: The Social Structure of Competition*).
- However, local clustering is faster to calculate.
Global clustering

\[ C_{WS} = \frac{1}{n} \sum_{i=1}^{n} C_i \]  

- This is an alternative equation for calculating global clustering coefficient
- It gives typically different results as the previous equation
- It is dominated by the nodes of lower degrees
Reciprocity

- The clustering coefficient measures the frequency with which the loops of length three appear in a network.
- In directed networks we can also concentrate on loops of length two.
- A pair of nodes with links between them running in both directions.
- The frequency of such loops is measured by *reciprocity*.
- Reciprocity tells us how likely, on average, is that a node that you point to, points back to you.
- Back link from a Web page, or a person you follow on Twitter follows you.
Reciprocity

- The reciprocity $r$ is defined as the fraction of links that are reciprocated.
- $A_{ij}A_{ji} = 1$, if and only if we have a link from $i$ to $j$ and from $j$ to $i$.
- $A_{ij}A_{ji} = 0$, otherwise.

$$r = \frac{1}{m} \sum_{ij} A_{ij}A_{ji} = \frac{1}{m} Tr A^2$$  \hspace{1cm} (31)

- $r \approx 0.57$ on the Web.
How nodes can be similar to each other and how can we quantify this similarity

Similarity can be calculated in many different ways even without the networks

E.g. content-based similarity of text documents

E.g. User similarities based on their profiles

Here we are interested in measuring similarities based on the network properties such as links, degrees, etc.
There are two approaches to measuring similarity of nodes in a network:

- **Structural similarity** is based on the number of the common neighbors.
- **Regular similarity** is based on the similarity of their respective (not necessarily common) neighbors.

The distinction is similar to centrality measures:

- Degree (neighbors count) vs. eigenvector centralities (recursion over the neighbors).
The most obvious measure of structural similarity is the number of common neighbors of two nodes

\[ n_{ij} = \sum_k A_{ik}A_{kj} \]  

This is the \( ij \)th element of \( A^2 \)

It is "cocitation" for undirected networks

This is also the number of paths of length 2 between \( i \) and \( j \)
Structural similarity

- A simple count of common neighbors on its own is not a good measure of node similarity
- E.g. what does it mean that two nodes have 3 or 1000 common neighbors
- We need normalization
- We could normalize by dividing by the maximal number of common neighbors \((n - 2)\)
- However, this penalizes the nodes with low degrees, e.g. two nodes with degrees 3 and all three common neighbors will get a small amount of similarity in a large network
Cosine similarity

- The scalar product of two vectors $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{x} \cdot \mathbf{y} = |\mathbf{x}| |\mathbf{y}| \cos \theta$$  \hspace{1cm} (33)

- where $|\mathbf{x}|$ and $|\mathbf{y}|$ and $\theta$ is the angle between the two vectors

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| |\mathbf{y}|}$$  \hspace{1cm} (34)
Cosine similarity

- We can regard $i$th and $j$th rows and columns of the adjacency matrix as vectors

$$\sigma_{ij} = \cos \theta = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2} \sqrt{\sum_k A_{jk}^2}}$$  \hspace{1cm} (35)
Cosine similarity

- For simple networks: $A_{ij}^2 = A_{ij}$
- Then: $\sum_k A_{ik}^2 = \sum_k A_{ik} = k_i$

**Definition**

Cosine similarity

$$\sigma_{ij} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$  \hspace{5cm} (36)

- If any of degrees equals to zero then we set $\sigma_{ij} = 0$
- It always lies in the range 0 to 1
Another normalization possibility is to compare the number of common neighbors to the number of common neighbors if nodes select their neighbors randomly. I.e. we compare the actual network structure with a random network structure. We obtain in this way *Pearson correlation coefficient*. 

Let nodes $i$ and $j$ have degrees $k_i$ and $k_j$ respectively.

Now we first let $i$ and then $j$ to choose their neighbors randomly.

The probability that $j$ selects one node that $i$ already has chosen is equal to $\frac{k_i}{n}$.

In total: the expected number of common neighbors is $\frac{k_i k_j}{n}$.

We define now similarity as the difference between the actual number of common neighbors and the expected number if they chose their neighbors randomly.
Pearson coefficients

\[ \sum_k A_{ik} A_{kj} - \frac{k_i k_j}{n} = \sum_k A_{ik} A_{kj} - \frac{1}{n} \sum_k A_{ik} \sum_l A_{jl} \]

\[ = \sum_k A_{ik} A_{kj} - n \bar{A}_i \bar{A}_j \]

\[ = \ldots \]

\[ = \sum_k (A_{ik} - \bar{A}_i) (A_{kj} - \bar{A}_j) \]  (37)
Pearson coefficients

- The last equation is \( n \) times covariance \( \text{cov}(A_i, A_j) \) of the \( i \)th and \( j \)th row of the adjacency matrix.
- It is positive if \( i \) and \( j \) have more common neighbors than what would be expected by chance.
- It is negative if \( i \) and \( j \) have less common neighbors than what would be expected by chance.
- It is zero if \( i \) and \( j \) have exactly as what would be expected by chance.
Similarity

Pearson coefficients

- We can normalize by the maximal value of the covariance which occurs when two set of quantities are the same
- Then covariance equals to variance of either sets $\sigma_i^2$, or $\sigma_j^2$, or $\sigma_i \sigma_j$
- Normalizing by this quantity we obtain the standard Pearson correlation coefficient
- $-1 \leq r_{ij} \leq 1$
Pearson coefficients

Definition

Pearson correlation coefficient

\[ r_{ij} = \frac{\text{cov}(A_i, A_j)}{\sigma_i \sigma_j} = \frac{\sum_k (A_{ik} - \overline{A_i})(A_{kj} - \overline{A_j})}{\sqrt{\sum_k (A_{ik} - \overline{A_i})^2 \sqrt{\sum_k (A_{kj} - \overline{A_j})^2}} \]  

(38)
Regular similarity

- The structural similarity measures the extent to which two nodes share the same neighbors.
- Regularly similar nodes are those that have neighbors that are similar.
- These neighbors must not be shared.
- The basic idea is to define a similarity score \( \sigma_{ij} \) such that \( i \) and \( j \) have high similarity if they have neighbors \( k \) and \( l \) that themselves have high similarity.
Regular similarity

\[ \sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl} \]  \hspace{1cm} (39)

\[ \sigma = \alpha A \sigma A \]  \hspace{1cm} (40)
Regular similarity

- The formula does not give a high value for “self-similarity” $\sigma_{ii}$
- As a consequence this does not give a high similarity score to nodes that share neighbors
- If self-similarity is high this would also give a high similarity score to nodes with many common neighbors

\[
\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl} + \delta_{ij} \tag{41}
\]

\[
\sigma = \alpha A \sigma A + I \tag{42}
\]
Regular similarity

- What happens if we evaluate the formula iteratively with $\sigma^0 = 0$

\[
\begin{align*}
\sigma^1 &= I \\
\sigma^2 &= \alpha A^2 + I \\
\sigma^3 &= \alpha^2 A^4 + \alpha A^2 + I
\end{align*}
\]
Regular similarity

- The pattern is clear
- In the limit of many iterations we get a sum over even powers of the adjacency matrix
- The elements of the $r$th power of $A$ count the number of paths of length $r$ between nodes
- Why should we count only paths of even length?
Regular similarity

- This leads to a better definition of regular similarity
- Nodes $i$ and $j$ are similar if $i$ has a neighbor $k$ that is itself similar to $j$
- Again we assume that nodes are similar to themselves

\[
\sigma_{ij} = \alpha \sum_k A_{ik} \sigma_{kj} + \delta_{ij} \tag{46}
\]

\[
\sigma = \alpha A \sigma + I \tag{47}
\]
Regular similarity

- Evaluating the new formula iteratively with $\sigma^0 = 0$

\[
\begin{align*}
\sigma^1 &= I \\ 
\sigma^2 &= \alpha A + I \\ 
\sigma^3 &= \alpha^2 A^2 + \alpha A + I
\end{align*}
\]
Regular similarity

- In the limit of a large number of iterations:

\[ \sigma = \sum_{m=0}^{\infty} (\alpha A)^m \]  

(51)

- And also, by rearranging:

\[ \sigma = (I - \alpha A)^{-1} \]  

(52)

\[ \sum_{m=0}^{\infty} (\alpha A)^m = (I - \alpha A)^{-1} \]  

(53)
Regular similarity

- This similarity measure includes counts of paths of all lengths.
- A weighted count of all the paths between nodes $i$ and $j$ with paths of length $r$ getting weight $\alpha^r$.
- As long as $\alpha < 1$ longer paths get less weight than shorter ones.
- In effect we say that two nodes are similar either if they are connected by few short paths or by many long paths.
The matrix \((I - \alpha A)\) does not have inverse when \(det(I - \alpha A) = 0\)

\[
det(I - \alpha A) = det(-\alpha(A - \frac{1}{\alpha}I))
\]

\[
= (-\alpha)^n det(A - \frac{1}{\alpha}I)
\]

Since \(\alpha \neq 0\), there is no inverse when \(det(A - \frac{1}{\alpha}I) = 0\)

This is characteristic polynomial and the solutions \(\frac{1}{\alpha} = \kappa\) are eigenvalues
Thus, the matrix does not have inverse (divergence) whenever \( \alpha = \frac{1}{\kappa} \)

If we start with small \( \alpha \) values and increase it the first time we hit the divergence is when \( \alpha = \frac{1}{\kappa_1} \)

After that it happens always when we hit another eigenvalue

Thus, if we pick \( \alpha < \frac{1}{\kappa_1} \) we guarantee convergence
“Katz similarity”

It is a generalization of the structural similarity

With structural similarity we count common neighbors

The number of common neighbors is the number of paths of length two

Our “Katz similarity” counts paths of all lengths and weight them differently
Similarly to discussion of PageRank and Katz centrality we can remove the effect of forwarding to much similarity to neighbors by dividing with node degree

$$\sigma_{ij} = \frac{\alpha}{k_i} \sum_k A_{ik} \sigma_{kj} + \delta_{ij}$$  \hspace{1cm} (57)$$

$$\sigma = (D - \alpha A)^{-1} D$$  \hspace{1cm} (58)$$
Figure: Friendship network at a US high school
Homophily and assortative mixing

- Division of the network into two groups
- Along lines of a membership in a class, e.g. race
- In social networks this phenomenon has been long observed
- Sociologists have observed such division along side many different dimensions
Homophily and assortative mixing

- People tend to make friends based on all sorts of characteristics, e.g. age, nationality, language, income, etc
- People tend to associate with other who are similar to them
- This tendency is called *homophily* or *assortative mixing*
- Sometimes, *disassortative mixing* is also observed
- This is tendency for people to associate with others who are unlike them, e.g. gender in romantic partnerships
Homophily

Assortative mixing

- Assortative mixing by enumerative characteristics
- E.g. belonging to a certain class such as nationality, race, or gender
- These are discrete values and in most cases binary values
- Assortative mixing by scalar characteristics
- E.g. age, income, degree, etc.
Homophily

Assortative mixing by enumerative characteristics

- We have a network in which the nodes are classified according to some characteristic that has a finite set of possible values
- For instance, nodes are people classified by nationality, or gender
- Nodes are Web pages classified by language
- Nodes are Wikipedia pages classified by topic
Homophily

Assortative mixing by enumerative characteristics

- The network is assortative if a significant fraction of links run between nodes of the same type.
- An elegant way to measure the assortiveness is to find the fraction of links that run between nodes of the same type and then subtract the fraction of such links that we would expect in a random network.
- If the fraction of links between nodes of the same type equals the expected number then our measure gives 0.
- Only if the fraction of links between nodes of the same type is significantly higher than the expected number we will have a positive difference.
In mathematical terms, let us denote by $c_i$ the class of node $i$.

Then the total number of links that run between nodes of the same type is:

$$\sum_{\text{links}(i,j)} \delta(c_i, c_j) = \frac{1}{2} \sum_{ij} A_{ij} \delta(c_i, c_j)$$  (59)
What is the expected number of links between nodes $i$ and $j$?

Let nodes $i$ and $j$ have degrees $k_i$ and $k_j$ respectively.

Now we let $j$ attach the second end of its single link to a random node.

The probability that $j$ selects node $i$ is equal to $\frac{k_i}{2m}$.

In total: the expected number of links between $i$ and $j$ is $\frac{k_j k_i}{2m}$, and the expected number of links between nodes of the same type:

$$\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(c_i, c_j)$$

(60)
Taking the difference we get:

$$\frac{1}{2} \sum_{ij} A_{ij} \delta(c_i, c_j) - \frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(c_i, c_j) = \frac{1}{2} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$$  \hspace{1cm} (61)

Typically, we will calculate the fraction of such links:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$$  \hspace{1cm} (62)
Homophily

Assortative mixing by enumerative characteristics

- The quantity $Q$ is called the modularity and is a measure of the extent to which like is connected to like in a network.
- It is strictly less than 1.
- It takes positive values if there are more links between nodes of the same type than what we would expect by chance.
- It takes negative values otherwise.
- We can also define $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$ as an element of matrix $B$.
- Modularity matrix.
Homophily

Assortative mixing by scalar characteristics

- Scalar characteristics allows us to say that two nodes are approximately the same
- E.g. two people are approximately of the same age
- In fact, people tend to associate with others on the basis of such approximate ages
- Thus, if nodes tend to be connected more often with other nodes having a similar characteristic then we say that the network is assortitavely mixed by that characteristic
Assortative mixing by scalar characteristics

**Figure:** Assortative mixing by age (from Mixing patterns in networks by Newman)
Assortative mixing by scalar characteristics

- How to measure the magnitude of the assortative mixing
- Let us use again covariance of a scalar quantity over all links
- We have the pairs of values \((x_i, x_j)\) for nodes linked by link \((i, j)\)

**Definition**

Average \(\mu\) of the value \(x_i\) at the end of a link

\[
\mu = \frac{\sum_{ij} A_{ij} x_i}{\sum_{ij} A_{ij}} = \frac{\sum_i k_i x_i}{\sum_i k_i} = \frac{1}{2m} \sum_i k_i x_i
\]  

(63)
Homophily

Assortative mixing by scalar characteristics

**Definition**

Covariance of $x_i$ and $x_j$ over links

\[
\text{cov}(x_i, x_j) = \frac{\sum_{ij} A_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{ij} A_{ij}}
\]

\[
= \ldots
\]

\[
= \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j
\]  

(64)
Homophily

Assortative mixing by scalar characteristics

- The covariance is positive if values $x_i$ and $x_j$ et both ends of a link tend to be either both small or both large
- It will be negative if they vary in opposite directions
- Thus, if we have assortative mixing the covariance is positive
- If we have disassortative mixing the covariance is negative
Homophily

Assortative mixing by scalar characteristics

- We can normalize to obtain 1 for a perfect mixed network
- In a perfectly mixed network $x_i$ and $x_j$ at both ends of a link are always equal
- We put $x_j = x_i$ in the previous equation and obtain the maximal covariance as our normalization constant
- In fact, it is the variance in this case
Homophily

Assortative mixing by scalar characteristics

**Definition**

Assortativity coefficient

\[
r = \frac{\sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) x_i x_j}{\sum_{ij} (k_i \delta_{ij} - \frac{k_i k_j}{2m}) x_i x_j}
\]  

(65)

- \(-1 \leq r \leq 1\)
- \(r = 0\) if the values on both ends of links are uncorrelated
- For the data from the previous figure \(r = 0.574\)
Assortative mixing by degree

- It is of special interest because degree is a network property.
- E.g., if we have assortative mixing by degree, high-degree nodes tend to connect to other high-degree nodes.
- Low-degree nodes tend to connect to other low-degree nodes.
- Typically, we obtain a network structure with a core of high-degree nodes and a periphery of low-degree nodes.
- In a disassortative mixing network, we obtain a star-like structure where high-degree nodes connect to low-degree nodes.
Homophily

Assortative mixing by degree

Definition
Covariance of $x_i$ and $x_j$ over links

$$\text{cov}(k_i, k_j) = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) k_i k_j$$  \hspace{1cm} (66)

Definition
Assortativity coefficient

$$r = \frac{\sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) k_i k_j}{\sum_{ij} (k_i \delta_{ij} - \frac{k_i k_j}{2m}) k_i k_j}$$  \hspace{1cm} (67)
As an example, we show in Fig. 8 the giant components of two graphs of this type generated using the Monte Carlo method. One of them, graph (a), is assortatively mixed by degree, while the other, graph (b), is disassortatively mixed. The difference between the two is clear to the eye. In the first case, because the high degree vertices prefer to attach to one another, there is a central "core" to the network, composed of these high-degree vertices, and a straggling periphery of low-degree vertices around it. In epidemiology a dense central portion of this type is called a "core group" and is thought to be capable of acting as a reservoir for disease, keeping diseases circulating even when the density of the network as a whole is too low to maintain endemic infection. In social network analysis one also talks of "core/periphery" distinctions in networks, another concept that mirrors what we see here. In the second graph, which is disassortative, a contrasting picture is evident: the high-degree vertices prefer not to associate with one another, and are as a result scattered widely over the network, producing a more uniform appearance.

To shed more light on the effects of assortativity, we show in Fig. 9 the size of the largest component in networks of this type as the degree distribution parameter \( \kappa \) is varied, for various values of \( p \). For low values of \( \kappa \) the mean degree of the network is small, and the resulting density of edges is too low to produce percolation in the network, so there is no giant component. As \( \kappa \) increases, however, there comes a point, clearly visible on the plot, at which the edge density is great enough to form a giant component. Figure 9 reveals two interesting features of this transition. First, the position of the transition, the value of the parameter \( \kappa \) at which it takes place, is smaller in assortatively mixed networks than in disassortative ones. In other words, it appears that the presence of assortativity in the degree correlation pattern allows the network to percolate more easily. This result is intuitively reasonable: the core group of the assortative network seen in Fig. 8a has a higher density of edges than the...
Homophily

Measures Notebook

- Jupyter Notebook example
- http://kti.tugraz.at/staff/denis/courses/netsci/measures.zip