Network Science Cheatsheet

UNIVERSITÉ DE LYON

Made by
Remy Cazabet

3 Nodes and Edges structural indices, neighborhoods similarity

Node Structural indices

Node structural indices, often called *Node centrality*, reflect how a node is characteristic of a given structural property. This is often summarized as *a measure of the node importance*, however *importance* and *centrality* are subjective/qualitative notions. Thus a centrality, despite its name, do not necessarily measure how *central* a node is, but rather how its position in the graph is typical of the property captured by this index.

Degree Centrality

Degree centrality is the most straighforward centrality. It can be interpreted as a measure of importance, of popularity, e.g., the more friends I have in a social network, the more *important* I am in this network.

Farness - Closeness - Harmonic centrality

The closeness of a node measures how close a node is from all other nodes, in term of shortest paths. To interpret it, we can make a parallel with a circle: the point which is the closest to all the other points of the circle is its center. The node of highest closeness is the equivalent of the center of the circle for this graph. Its formulation is easily understood as the inverse of the farness.

Farness: Average distance to all other nodes in the graph

$$\text{Farness}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \ell_{u,v}$$

Closeness: Inverse of the farness, i.e., how close the node is to all other nodes in term of shortest paths.

$$\text{Closeness}(u) = \frac{N-1}{\sum_{v \in V \setminus u} \ell_{u,v}}$$

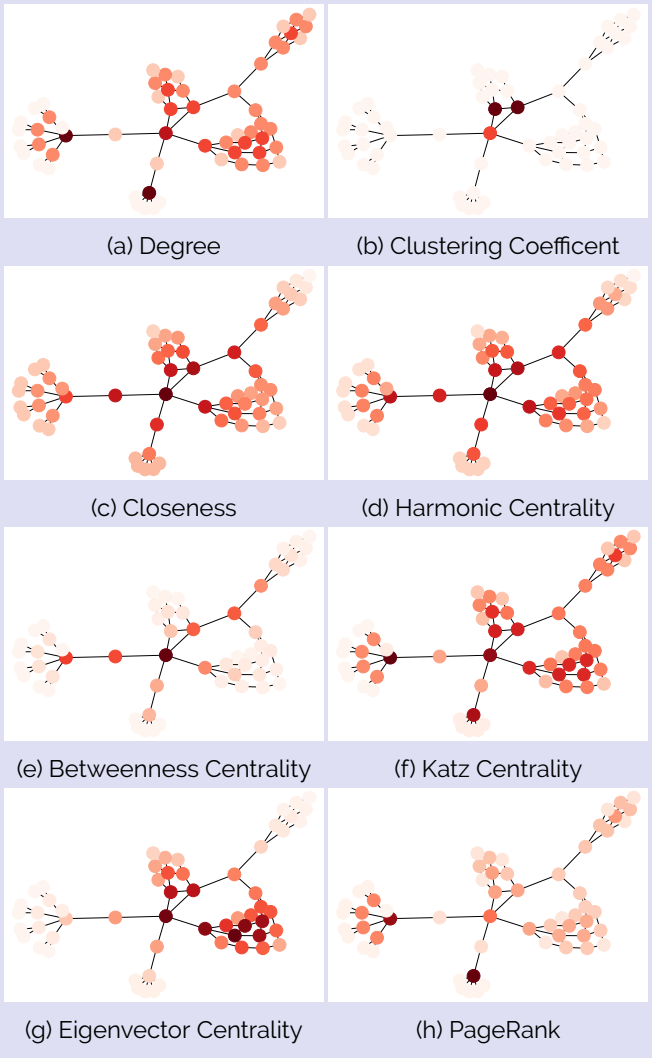
Harmonic centrality: A variant of the closeness defined as the average of the inverse of distance to all other nodes (Harmonic mean). Well defined on disconnected network with $\frac{1}{\infty} = 0$. Its interpretation is the same as the closeness.

$$\text{Harmonic}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \frac{1}{\ell_{u,v}}$$

Clustering Coefficient

This score, already defined, measure the *triadic closure* of a node. A high score is often interpreted as being well embedded in a particular community (friends of my friends are my friends because we all belong to the same group), a low score can be typical of a *bridge* node, e.g., few connections between my friends because they belong to different social circles.

Centrality - Examples



(a) Degree (b) Clustering Coefficent

(c) Closeness (d) Harmonic Centrality

(e) Betweenness Centrality (f) Katz Centrality

(g) Eigenvector Centrality (h) PageRank

Katz centrality

Katz centrality is said to be a measure of the influence potential of a node. For a node u, it is defined as the sum, for all path length distance ℓ, of the number of nodes located at distance exactly ℓ of u, discounted of a factor decreasing as ℓ increases. The intuition is that, the more nodes can be accessed in few steps, the higher the value. More formally, it is expressed as

$$C_{\text{Katz}}(u) = \sum_{\ell=1}^{\infty} \sum_{v=1}^{N} \alpha^{\ell} (A^{\ell})_{vu}$$

in which A^{ℓ}_{vu} means the number of paths of length ℓ from v to u, and $\alpha < \frac{1}{\lambda_1}$ a parameter smaller than the reciprocal of the largest eigenvalue of A, allowing to compute with matrix form:

$$C_{\text{Katz}}(u) = ((I - \alpha A^T)^{-1} - I)\vec{1}$$

Note that in a directed network, Katz centrality must be interpreted as a *vote* mechanism: a highest centrality of u means that more nodes can reach u quickly, and not that u can reach many nodes quickly.

Betweenness centrality

The betweenness centrality measures how much the node plays the role of a bridge. The highest the betweenness, the more the node is essential to move quickly in the graph. More formally, the betweenness of u is defined as the fraction of the shortest paths between all pairs of nodes in the graph (but u) that go through u. As a consequence, if we remove a node of high betweenness, many shortest paths will become longer, and the graph harder to navigate. The extreme situation is a node on the only path between otherwise disconnected components: if we remove this node, some nodes becomes unreachable from others. Those nodes thus tend to have high betweenness. It is defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

with σ_{st} the number of shortest paths between nodes s and t and $\sigma_{st}(v)$ the number of those paths passing through v.
The betweenness tends to grow with the network size. A normalized version can be obtained by dividing by the number of pairs of nodes, i.e., for a directed graph: $C_B^{\text{norm}}(v) = \frac{C_B(v)}{(N-1)(N-2)}$.

Eigenvector centrality

Eigenvector centrality is a recursive definition of importance: a node is important if it is connected to other important nodes. In practice, it is defined in the following way: the eigenvector centrality C_u for every node u of the graph is such that if each node *sends* its centrality score to its neighbors, then the sum of scores received by each node will be equal to λC_u (with λ a constant). More formally,

$$C_u^{t+1} = \frac{1}{\lambda} \sum_{v \in N_u^{in}} C_v^t \quad (1)$$

with λ a normalisation constant. This recursive definition can be interpreted in term of eigenvectors and eigenvalues, which is defined as $Ax = \lambda x$, with x an eigenvector, λ the corresponding eigeinvalue. The eigenvector centrality is defined as the leading invector, i.e., the eigenvector associated with the highest eigenvalue, the only solution for which all centrality values are positive.
A simple way to compute this eigenvalue is called the power method: one start with random values on nodes, and iterate equation 1. After some time, it can be proven that the values converge to the eigenvector centrality.

Eigenvector centrality cannot in general be computed on directed networks, because of source nodes, i.e., $k^{in} = 0$. Those nodes have by definition a, eigenvector centrality of 0 at t+1, and thus *send* a value of 0 at t+2, which might in turn result in a score of 0 for its successors, and so on and so forth.

Pagerank centrality

Pagerank centrality is famous for being the method originally used by google to rank web-pages: all pages containing the researched words are ordered according to their Pagerank score in the graph of the WWW, in which nodes are webpages and edges are hyperlinks.
It is a variant of the Eigenvector centrality, solving the problem of source nodes.
Pagerank introduces two improvements: 1) at each step t, each node gain a small constant value. 2) The values *sent* are divided equally among successors (normalization by degree). Equation 1 thus becomes:

$$C_u^{t+1} = \alpha \sum_{v \in N_u^{in}} \frac{C_v^t}{k_v^{out}} + \beta \quad (2)$$

with, by convention, $\beta = 1$,$\alpha \in [0, 1]$ a parameter.
Pagerank centrality can also be expressed as the leading eigenvector of the so-called *Google matrix* G, defined as $G_{ij} = \alpha S_{ij} + (1 - \alpha)/n$, with S_{ij} the adjacency matrix normalized by column.

Pagerank & Random Walk

Pagerank can be interpreted in term of **random walks**. If you consider a random walker moving from nodes to nodes following randomly chosen out-going links, which starts on a random node and moves an infinite number of times. Consider that at each step, this random walker can *teleport* to any other node with a probability $/alpha$ instead of following an outgoing edge. Then, the probability for this random walker to be on each particular node corresponds to its Pagerank score.
We can note that the average length of a walk before restart is $\frac{\alpha}{1-\alpha}$. The *typical* value $\alpha = 0.85$ thus means that random walkers move in average 5.7 times before restart, a typical value of average distance in real graphs.

Edge Structural indices

Edges situation in the network can also be described using structural properties, most of them being similar to node centralities.
Edge Clustering C^e of an edge (u, v) is the fraction of the neighbors of at least one of the two nodes which are neighbors of both of them, i.e.,

$$C^e(u, v) = \frac{|N_u \cap N_v|}{|N_u \cup N_v| - 2}$$

High clustering edges are said *Integrative*, low values nodes are said *Dispersive*.
Edge betweenness Is defined exactly as node betweenness, but counting shortest paths going through each edge instead of each node, i.e.,

$$C_B(u, v) = \sum_{s \neq t \in V} \frac{\sigma_{st}(u, v)}{\sigma_{st}}$$

with σ_{st} the number of shortest paths between nodes s and t and $\sigma_{st}(u, v)$ the number of those paths passing through edge (u, v).

Node Similarity

When studying a network, one might be interested in comparing nodes between themselves, for instance to discover the most similar nodes in the network, or to assess if two nodes they are interested in share a similar network location.

A first approach is to define the similarity between nodes u and v, $\sigma_{u,v}$ as: $\sigma_{u,v} = |N_u \cap N_v|$.

A weakness of this approach is that high degree nodes tends to be considered similar to low degree nodes. A variant consists in normalizing by nodes degrees, thus computing the Jaccard Coefficient of neighborhoods:

$$\sigma_{u,v} = \frac{|N_u \cap N_v|}{|N_u \cup N_v| - 2}$$

Cosine Similarity

Cosine similarity $\sigma^{\cos}$ is a standard method to compare *vectors*. It is defined for two vectors x, y as :

$$\sigma_{xy}^{\cos} = \frac{x.y}{|x||y|}$$

This score can be used to measure the similarity between nodes neighborhoods by using as vector x_u of node u the row of the adjacency matrix corresponding to this node, i.e., $x_u = A_u$.
Cosine similarity of nodes then simplifies to:

$$\sigma_{uv}^{\cos} = \frac{|N_u \cap N_v|}{\sqrt{k_u k_v}}$$

Pearson coefficient

Pearson coefficient is a standard measure of correlation between variables X and Y, which is defined as :

$$r_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

with cov the covariance and σ the standard deviation.
Much as for Cosine Similarity, we can adapt this measure to nodes similarities by considering A's rows as discrete variables. The result can be understood intuitively by observing that the numerator becomes:

$$cov(u, v) = |N_u \cap N_v| - \frac{k_u k_v}{N}$$

which can be interpreted as the **number of common neighbors minus the expected number of common neighbors** in a randomized network, given nodes degrees.

$cov(u, v) = 0$ means that the number of common neighbors is exactly what we would expect by chance given their degrees, while positive values means that they have more than expected (resp. for negative values).