Giovanni Scardoni,
giovanni.scardoni@gmail.com
Carlo Laudanna,
carlo.laudanna@univr.it
CBMC, University of Verona

Network centralities for Cytoscape

Definitions

Let $G = (V, E)$ a directed or undirected graph, with $n = |V|$ vertexes. $deg(v)$, $deg(v)$, $deg+(v)$ indicate, respectively, the degree, the in-degree and the out-degree of the vertex. $dist(v, w)$ is the shortest path between v and w. σ_{st} is the number of shortest paths between s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t passing through the vertex v. Notably:

- Vertex = nodes; edges = arches;
- The distance between two nodes, $dist(v, w)$, always indicates the shortest path between the two nodes;
- All calculated scores (but the Wiener index) are computed giving to higher values a positive meaning, where positive does refer to node proximity to other nodes. Thus, independently on the calculated node centrality, higher scores indicated proximity and lower scores indicate remoteness of a given node v from the other nodes in the graph G.

Centralities

Degree ($deg(k)$)

Is the simplest topological index, corresponding to the number of nodes adjacent to a given node v, where adjacent means directly connected. The nodes directly connected to a given node v are also called first neighbors of the given node. Thus, the degree also corresponds to the number of adjacent incident edges. In directed networks we distinguish in-degree, when the edges target the node v, and out-degree, when the edges target the adjacent neighbors of v. Calculation of the degree allows determining the degree distribution $P(k)$, which gives the probability that a selected node has exactly k links. $P(k)$ is obtained counting the number of nodes $N(k)$ with $k = 1, 2, 3 \ldots$ links and dividing by the total number of nodes N. Determining the degree distribution allows distinguishing different kind of graphs. For instance, a graph with a peaked degree distribution (Gaussian distribution) indicates that the system has a characteristic degree with no highly connected nodes. This is typical of random, non-natural, networks. By contrast, a power-law degree distribution indicates the presence of few nodes having a very high degree. Nodes with high degree (highly connected)

are called hubs and hold together several nodes with lower degree. Networks displaying a degree distribution approximating a power-law, $P(k)\ k^{\gamma}$, where γ is degree exponent and indicates proportional to, are called scale-free networks. Scale-free networks are mainly dominated by hubs (the smaller the value of γ the more important is the role of the hubs in the network) and are intrinsically robust to random attacks but vulnerable to selected alterations. Scale-free networks are typically natural networks.

In biological terms

The degree allows an immediate evaluation of the regulatory relevance of the node. For instance, in signaling networks, proteins with very high degree are interacting with several other signaling proteins, thus suggesting a central regulatory role, that is they are likely to be regulatory hubs. For instance, signaling proteins encoded by oncogenes, such as HRAS, SRC or TP53, are hubs. Depending on the nature of the protein, the degree could indicate a central role in amplification (kinases), diversification and turnover (small GTPases), signaling module assembly (docking proteins), gene expression (transcription factors), etc. Signaling networks have typically a scale-free architecture.

Diameter (Δ_G)

ΔG = the maximal distance (shortest path) amongst all the distances calculated between each couple of vertexes in graph G. The diameter indicates how much distant are the two most distant nodes. It can be a first and simple general parameter of graph compactness, meaning with that the overall proximity between nodes. A high graph diameter indicates that the two nodes determining that diameter are very distant, implying little graph compactness. However, it is possible that two nodes are very distant, thus giving a high graph diameter, but several other nodes are not. Therefore, a graph could have high diameter and still being rather compact or have very compact regions. Thus, a high graph diameter can be misleading in term of evaluation of graph compactness. In contrast a low graph diameter is much more informative and reliable. Indeed, a low diameter surely indicates that all the nodes are in proximity and the graph is compact. In quantitative terms, high and low are better defined when compared to the total number of nodes in the graph. Thus, a low diameter of a very big graph (with hundreds of nodes) is much more meaningful in term of compactness than a low diameter of a small graph (with few nodes). Notably, the diameter enables to measure the development of a network in time.

In biological terms

The diameter, and thus the compactness, of a biological network, for instance a protein-signaling network, can be interpreted as the overall easiness of the proteins to communicate and/or influence their reciprocal function. It could be also a sign of functional convergence. Indeed, a big protein network with

low diameter may suggest that the proteins within the network had a functional co-evolution. The diameter should be carefully weighted if the graph is not fully connected (that is, there are isolated nodes).

Average Distance (AvD_G)

AvD_G = the average shortest path of a graph G, corresponding to the summa of all shortest paths between vertex couples divided for the total number of vertex couples. Often it is not an integer. As for the diameter, it can be a simple and general parameter of graph compactness, meaning with that the overall tendency of nodes to stay in proximity. Being an average, it can be somehow more informative than the diameter and can be also considered a general indicator of network navigability. A high average distance indicates that the nodes are distant (disperse), implying little graph compactness. In contrast a low average distance indicates that all the nodes are in proximity and the graph is compact. In quantitative terms, high and low are better defined when compared to the total number of nodes in the graph. Thus, a low average distance of a very big graph (with hundreds of nodes) is more meaningful in term of compactness than a low average distance of a small graph (with few nodes).

In biological terms

The average distance of a biological network, for instance a protein-signaling network, can be interpreted as the overall easiness of the proteins to communicate and/or influence their reciprocal function. It could be also a sign of functional convergence. Indeed, a big protein network with low average distance may suggest that the proteins within the network have the tendency to generate functional complexes and/or modules (although centrality indexes should be also calculated to support that indication).

Eccentricity ($C_{ecc}(v)$)

$$C_{ecc}(v) := \frac{1}{max\{dist(v,w) : w \in V\}}$$

The eccentricity is a node centrality index. The eccentricity of a node v is calculated by computing the shortest path between the node v and all other nodes in the graph, then the longest shortest path is chosen (let (v, K) where K is the most distance node from v). Once this path with length $dist(v, K)$ is identified, its reciprocal is calculated ($1/dist(v, K)$). By doing that, an eccentricity with higher value assumes a positive meaning in term of node proximity. Indeed, if the eccentricity of the node v is high, this means that all other nodes are in proximity. In contrast, if the eccentricity is low, this means that there is at least one node (and all its neighbors) that is far form node v. Of course, this does not exclude that several other nodes are much closer to node v. Thus, eccentricity is a more meaningful parameter if is high. Notably, high and low values are more

significant when compared to the average eccentricity of the graph G calculated by averaging the eccentricity values of all nodes in the graph.

In biological terms

The eccentricity of a node in a biological network, for instance a protein-signaling network, can be interpreted as the easiness of a protein to be functionally reached by all other proteins in the network. Thus, a protein with high eccentricity, compared to the average eccentricity of the network, will be more easily influenced by the activity of other proteins (the protein is subject to a more stringent or complex regulation) or, conversely could easily influence several other proteins. In contrast, a low eccentricity, compared to the average eccentricity of the network, could indicate a marginal functional role (although this should be also evaluated with other parameters and contextualized to the network annotations).

Closeness ($C_{clo}(v)$)

$$C_{clo}(v) := \frac{1}{\sum_{w \in V} dist(v, w)}$$

The closeness is a node centrality index. The closeness of a node v is calculated by computing the shortest path between the node v and all other nodes in the graph, and then calculating the summa. Once this value is obtained, its reciprocal is calculated, so higher values assume a positive meaning in term of node proximity. Also here, high and low values are more meaningful when compared to the average closeness of the graph G calculated by averaging the closeness values of all nodes in the graph. Notably, high values of closeness should indicate that all other nodes are in proximity to node v. In contrast, low values of closeness should indicate that all other nodes are distant from node v. However, a high closeness value can be determined by the presence of few nodes very close to node v, with other much more distant, or by the fact that all nodes are generally very close to v. Likewise, a low closeness value can be determined by the presence of few nodes very distant from node v, with other much closer, or by the fact that all nodes are generally distant from v. Thus, the closeness value should be considered as an average tendency to node proximity or isolation, not really informative on the specific nature of the individual node couples. The closeness should be always compared to the eccentricity: a node with high eccentricity + high closeness is very likely to be central in the graph.

In biological terms

The closeness of a node in a biological network, for instance a protein-signaling network, can be interpreted as the probability of a protein to be functionally relevant for several other proteins, but with the possibility to be irrelevant for few other proteins. Thus, a protein with high closeness, compared to the average closeness of the network, will be easily central to the regulation of other proteins

but with some proteins not influenced by its activity. Notably, in biological networks could be also of interest to analyze proteins with low closeness, compared to the average closeness of the network, as these proteins, although less relevant for that specific network, are possibly behaving as intersecting boundaries with other networks. Accordingly, a signaling network with a very high average closeness is more likely organizing functional units or modules, whereas a signaling network with very low average closeness will behave more likely as an open cluster of proteins connecting different regulatory modules.

Radiality ($C_{rad}(v)$)

$$C_{rad}(v) := \frac{\sum_{w \in V}(\Delta_G + 1 - dist(v,w))}{n-1}$$

The radiality is a node centrality index. The radiality of a node v is calculated by computing the shortest path between the node v and all other nodes in the graph. The value of each path is then subtracted by the value of the diameter $+1$ ($\Delta G + 1$) and the resulting values are summated. Finally, the obtained value is divided for the number of nodes -1 ($n-1$). Basically, as the diameter is the maximal possible distance between nodes, subtracting systematically from the diameter the shortest paths between the node v and its neighbors will give high values if the paths are short and low values if the paths are long. Overall, if the radiality is high this means that, with respect to the diameter, the node is generally closer to the other nodes, whereas, if the radiality is low, this means that the node is peripheral. Also here, high and low values are more meaningful when compared to the average radiality of the graph G calculated by averaging the radiality values of all nodes in the graph. As for the closeness, the radiality value should be considered as an average tendency to node proximity or isolation, not definitively informative on the centrality of the individual node. The radiality should be always compared to the closeness and to the eccentricity: a node with high eccentricity + high closeness+ high radiality is a consistent indication of a high central position in the graph.

In biological terms

The radiality of a node in a biological network, for instance a protein-signaling network, can be interpreted as the probability of a protein to be functionally relevant for several other proteins, but with the possibility to be irrelevant for few other proteins. Thus, a protein with high radiality, compared to the average radiality of the network, will be easily central to the regulation of other proteins but with some proteins not influenced by its activity. Notably, in biological networks could be also of interest to analyze proteins with low radiality, compared to the average radiality of the network, as these proteins, although less relevant for that specific network, are possibly behaving as intersecting boundaries with other networks. Accordingly, a signaling network with a very high average radiality is more likely organizing functional units or modules, whereas a signaling network with very low average radiality will behave more likely as an open

cluster of proteins connecting different regulatory modules. All these interpretations should be accompanied to the contemporary evaluation of eccentricity and closeness.

Centroid value ($C_{cen}(v)$)

$$C_{cen}(v) := min\{f(v,w) : w \in V\{v\}\}$$

Where $f(v,w) := \gamma_v(w) - \gamma_w(v)$, and $\gamma_v(w)$ is the number of vertex closer to v than to w. The centroid value is the most complex node centrality index. It is computed by focusing the calculus on couples of nodes (v,w) and systematically counting the nodes that are closer (in term of shortest path) to v or to w. The calculus proceeds by comparing the node distance from other nodes with the distance of all other nodes from the others, such that a high centroid value indicates that a node v is much closer to other nodes. Thus, the centroid value provides a centrality index always weighted with the values of all other nodes in the graph. Indeed, the node with the highest centroid value is also the node with the highest number of neighbors (not only first) if compared with all other nodes. In other terms, a node v with the highest centroid value is the node with the highest number of neighbors separated by the shortest path to v. The centroid value suggests that a specific node has a central position within a graph region characterized by a high density of interacting nodes. Also here, high and low values are more meaningful when compared to the average centrality value of the graph G calculated by averaging the centrality values of all nodes in the graph.

In biological terms

The centrality value of a node in a biological network, for instance a protein-signaling network, can be interpreted as the probability of a protein to be functionally capable of organizing discrete protein clusters or modules. Thus, a protein with high centroid value, compared to the average centroid value of the network, will be possibly involved in coordinating the activity of other highly connected proteins, altogether devoted to the regulation of a specific cell activity (for instance, cell adhesion, gene expression, proliferation etc.). Accordingly, a signaling network with a very high average centroid value is more likely organizing functional units or modules, whereas a signaling network with very low average centroid value will behave more likely as an open cluster of proteins connecting different regulatory modules. It can be useful to compare the centroid value to other algorithms detecting dense regions in a graph, indicating protein clusters, such as, for instance, MCODE.

Stress ($C_{str}(v)$)

$$C_{str}(v) := \sum_{s \neq v \in V} \sum_{t \neq v \in V} \sigma_{st}(v)$$

The stress is a node centrality index. Stress is calculated by measuring the number of shortest paths passing through a node. To calculate the stress of a node v, all shortest paths in a graph G are calculated and then the number of shortest paths passing through v is counted. A stressed node is a node traversed by a high number of shortest paths. Notably, and importantly, a high stress values does not automatically implies that the node v is critical to maintain the connection between nodes whose paths are passing through it. Indeed, it is possible that two nodes are connected by means of other shortest paths not passing through the node v. Also here, high and low values are more meaningful when compared to the average stress value of the graph G calculated by averaging the stress values of all nodes in the graph.

In biological terms

The stress of a node in a biological network, for instance a protein-signaling network, can indicate the relevance of a protein as functionally capable of holding together communicating nodes. The higher the value the higher the relevance of the protein in connecting regulatory molecules. Due to the nature of this centrality, it is possible that the stress simply indicates a molecule heavily involved in cellular processes but not relevant to maintain the communication between other proteins.

S.-P. Betweenness ($C_{spb}(v)$)

$$C_{spb}(v) := \sum_{s \neq v \in V} \sum_{t \neq v \in V} \delta_{st}(v)$$

where

$$\delta_{st}(v) := \frac{\sigma_{st}(v)}{\sigma_{st}}$$

The S.-P. Betweenness is a node centrality index. It is similar to the stress but provides a more elaborated and informative centrality index. It is calculated considering couples of nodes $(v1, v2)$ and counting the number of shortest paths linking $v1$ and $v2$ and passing through a node n. Then, the value is related to the total number of shortest paths linking $v1$ and $v2$. Thus, a node can be traversed by only one path linking $v1$ and $v2$, but if this path is the only connecting $v1$ and $v2$ the node n will score a higher betweenness value (in the stress computation would have had a low score). Thus, a high S.-P. Betweenness score means that the node, for certain paths, is crucial to maintain node connections. Notably, to know the number of paths for which the node is critical it is necessary to look at the stress. Thus, stress and S.-P. Betweenness can be used to gain complementary information. Further information could be gained by referring the S.-P. Betweenness to node couples, thus quantifying the importance of a node for two connected nodes. Also here, high and low values are more meaningful when compared to the average S.-P. Betweenness value of the graph G calculated by averaging the S.-P. Betweenness values of all nodes in the graph.

In biological terms

The S.-P. Betweenness of a node in a biological network, for instance a protein-signaling network, can indicate the relevance of a protein as functionally capable of holding together communicating proteins. The higher the value the higher the relevance of the protein as organizing regulatory molecule. The S.-P. Betweenness of a protein effectively indicates the capability of a protein to bring in communication distant proteins. In signaling modules, proteins with high S.-P. Betweenness are likely crucial to maintain functionally and coherence of signaling mechanisms.

Wiener index ($W_{index}(v)$)

$$W_{index}(v) := \sum_{w \in V} dist(v, w)$$

[B The Wiener index is a node centrality index. The Wiener index is the summa of the all distances (shortest paths) between a node v and all other nodes in the graph. Basically, it is similar to the closeness but here the score has the opposite meaning, as the reciprocal is not calculated. Thus a low Wiener index means that a node is closer to all other nodes. High and low values are more meaningful when compared to the average Wiener index of the graph G calculated by averaging the Wiener index values of all nodes in the graph. Notably, the graph Wiener index is the summa of the Wiener indexes of all nodes in the graph. Also here, a low graph Wiener index indicates a graph whose nodes are likely to be highly connected. As for the closeness, the Wiener index provides a general evaluation of the average tendency of the node to stay in proximity or isolation, not really informative on the specific nature of the individual node couples. The Wiener index should be always compared to the eccentricity, closeness and radiality.

In biological terms

The Wiener index of a node in a biological network, for instance a protein-signaling network, can be interpreted as the probability of a protein to be functionally relevant for several other proteins, but with the possibility to be irrelevant for few other proteins. Thus, a protein with low Wiener index, compared to the average Wiener index of the network, will be easily central to the regulation of other proteins but with some proteins not influenced by its activity. Notably, in biological networks could be also of interest to analyze nodes with high Wiener index, compared to the average Wiener index of the network, as these nodes, although less relevant for that specific network, are possibly behaving as intersecting boundaries with other networks. Accordingly, a signaling network with a very low average Wiener index or a low graph Wiener index is more likely organizing functional units or modules, whereas a signaling network with very high average Wiener index or graph Wiener index will behave more likely as an open cluster of proteins connecting different regulatory modules.