Algorithms for graph and network analysis: Graph indexes/descriptors

Paolo Lo Giudice[1] and Domenico Ursino[2*]

[1] DIIES, University "Mediterranea" of Reggio Calabria, Via Graziella, Località Feo di Vito, I-89122 Reggio Calabria, Italy
[2] DICEAM, University "Mediterranea" of Reggio Calabria, Via Graziella, Località Feo di Vito, I-89122 Reggio Calabria, Italy

Abstract

Graph and network analysis has been a multidisciplinary research field from its origin. As such, it has been widely employed also in bioinformatics and in biomedicine, where approaches based on graph indexes and descriptors have been proposed for addressing many challenges. In this chapter, first we show the main ways to represent a network. Then, we describe the most known network indexes, starting from the basic ones, continuing with the centrality based methods, and finishing with cohesion-based and other methods. Finally, we draw some closing remarks.

Keywords: Network representation, Degree, Density, Walk, Path, Cycle, Radius, Diameter, Centrality Indexes, Triads, Cliques, Components, Critical Mass

1 Introduction

Network Analysis (hereafter, NA) has been a multidisciplinary research field from its origin. Relationships represent the key concept in a network. Indeed, relationships, and not participants, play the biggest role in modeling a network [26, 47].

NA allows the identification of the relationship patterns that exist in a network. Moreover, it allows the detection, and the subsequent investigation, of the information (and/or other resource) flow among participants. Finally, it focuses on interactions among participants, which differentiates it from other kinds of analysis that mainly investigate the features of a single participant.

Network analysis-based approaches allow interactions in a group to be mapped, as well as the connectivity of a network to be visualized and investigated. Furthermore, they make it possible to quantify the processes taking place among network participants [40, 54, 61].

Three research fields where the use of NA is rapidly increasing are bioinformatics, biomedicine and QSAR/QSPR. Think, for instance, of Public Health [46, 7, 29]. According to [46], it is possible to find

*Corresponding Author – Phone Number: +39-347-7118918 – e-mail: ursino@unirc.it

three main types of network in this sector, namely: *(i)* transmission networks, *(ii)* social networks, and *(iii)* organizational networks. Transmission networks are particularly relevant and, therefore, widely investigated. They allow the analysis of the diffusion of both diseases [17, 18, 33, 2, 60] and medical information [60, 37, 58, 23, 59]. Two very relevant application contexts for transmission networks are social epidemiology [8, 27] and information diffusion on social networks [16, 16, 53, 28, 42, 65]. Social networks investigate how social structures and relationships influence both public health and human behavior [39, 6, 10, 35, 45]. Organizational networks represent the most recent research sector; in this case, researchers evaluate the impact of associations and/or agencies on public health [43, 9, 5, 48, 36].

In bioinformatics, an important investigation regards the usage of "information-based" tools to analyze medical problems. In this case, two very important research areas are molecular analysis [63, 12, 19, 25, 55] and brain analysis [52, 22, 1, 56, 67]. Another relevant topic concerns the definition of software packages and analytic tools allowing extensive studies on large datasets [31, 44, 68, 41, 38, 11].

In this analysis, two of the most used indexes are: *(i)* centrality indicators (adopted, for instance, in [66, 34]), and *(ii)* connection indicators (employed, for instance, in [21, 15, 64]). An overview on the usage of these indicators can be found in [20]. For instance, Closeness Centrality is used in [24] to study the evolution of protein-protein networks. In [49], the authors use eigenvector centrality to predict good candidate disease-related genes. In [14], the authors adopt 16 different centrality measures to analyze 18 metabolic networks. Finally, in [30], the authors employ both centrality and cohesion indexes for understanding how miRNAs influence the protein interaction network.

QSARs (Quantitative Structure Activity Relationships) and QSPRs (Quantitative Structure Property Relationship) represent mathematical correlations between a given biological activity or molecular property and one or more physicochemical and/or molecular structural properties. These are called descriptors since they "describe" the activity or the property under examination. Topological indices are very important descriptors. They provide a numerical representation of the topology of a molecule. One of the first topological indices was proposed by Wiener in 1947 [62]. This index correlates well with the boiling points of alkanes. Three important families of topological indices are molecular connectivities, electrotopological state (e-state) values, and information content indices [13, 3, 32, 4, 51, 50, 57].

This chapter is organized as follows. In Section 2, we describe how networks can be represented. In Section 3, we illustrate the main indexes employed in network analysis. Finally, in Section 4, we draw our conclusions.

2 Network Representation

A network $\mathcal{N} = \langle V, E \rangle$ consists of a set V of nodes and a set E of edges. Each edge $e_{ij} = (v_i, v_j)$ connects the nodes v_i and v_j. Edges can be either directed (when they can be traversed only in one direction) or undirected (when they can be traversed in both directions). Furthermore, networks can be weighted or unweighted. If a network is weighted, it can be represented as (v_i, v_j, w_{ij}), where w_{ij} denotes the weight of the corresponding edge. On the basis of the reference context, this weight could represent strength, distance, similarity, etc.

Example 2.1 Consider the networks in Figure 1. The one on the left is undirected and unweighted,

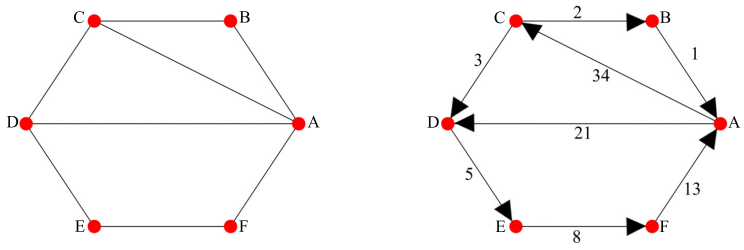

Figure 1: An example of undirected and unweighted network (on the left), and an example of directed and weighted network (on the right)

whereas the one on the right is directed and weighted. For instance, the edge $A - C$ in the network on the right indicates that there is a link from A to C and that the weight of this link is 34.

An important way to represent a network $\mathcal{N}$ employs the adjacency matrix $\mathcal{A}$. This is a $|V| \times |V|$ matrix. Each element corresponds to an edge between the nodes corresponding to the row and column. If $\mathcal{N}$ is unweighted, the generic element $\mathcal{A}[i, j]$ is set to 1 if there exists an edge from v_i to v_j; otherwise, it is set to 0. By contrast, if $\mathcal{N}$ is weighted, $\mathcal{A}[i, j]$ is set to the weight of the edge from v_i to v_j. Finally, if $\mathcal{N}$ is undirected, the corresponding adjacency matrix is conventionally shown as lower triangular matrix. The adjacency matrix is easily understood; however, in real cases, it is very sparse (i.e., most of its elements are set equal to 0) and, therefore, it wastes a lot of space.

To reduce the waste of space, $\mathcal{N}$ can be represented as an edge list $\mathcal{L}$. In this case, if $\mathcal{N}$ is unweighted, $\mathcal{L}$ consists of a list of pairs, each representing an edge with its starting and ending nodes. By contrast, if $\mathcal{N}$ is weighted, $\mathcal{L}$ consists of a list of triplets, each representing an edge with the corresponding starting node, ending node and weight. Clearly, an edge list is more compact, but less clear, than an adjacency matrix.

A further reduction of the space needed to represent $\mathcal{N}$ is obtained by adopting an adjacency list $\mathcal{L}^*$. If $\mathcal{N}$ is unweighted, $\mathcal{L}^*$ consists of a list of pairs $\langle v_i, \mathcal{L}'_i \rangle$, where v_i is a node of $\mathcal{N}$ and $\mathcal{L}'_i$ is the list of the nodes reachable from it. If $\mathcal{N}$ is weighted, $\mathcal{L}^*$ consists of a list of pairs $\langle v_i, \mathcal{L}''_i \rangle$, where v_i is a node of $\mathcal{N}$ and $\mathcal{L}''_i$ is, in turn, a list of pairs (v_j, w_{ij}), such that v_j is reachable from v_i and w_{ij} is the weight of the corresponding edge. Clearly, among the three structures presented above, the adjacency list is the most compact, but also the least clear.

Example 2.1 (...cnt'd)

Consider the networks shown in Figure 1. The corresponding adjacency matrixes are reported in Figure 2, and the associated edge lists are shown in Table 1. Finally, the corresponding adjacency lists are illustrated in Table 2.

Undirected and Unweighted Network	Directed and Weighted Network
(A,B), (A,C), (A,D), (A,E), (B,C), (C,D), (D,E), (E,F)	(A,C, 34), (A,D,21), (B,A,1), (C,B,2), (C,D,3), (D,E,5), (E,F,8), (F,A,13)

Table 1: The Edge Lists corresponding to the networks of Figure 1

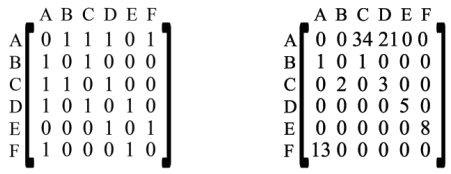

Figure 2: The Adjacency Matrixes corresponding to the networks of Figure 1

Undirected and Unweighted Network	Directed and Weighted Network
A - {B, C, D, E}	A - {(C,34), (D,21)}
B - {C}	B - {(A,1)}
C - {D}	C - {(B,2), (D,3)}
D - {E}	D - {(E,5)}
E - {F}	E - {(F,8)}
F - {}	F - {(A,13)}

Table 2: The Adjacency Lists corresponding to the networks of Figure 1

3 Index Description

3.1 Basic Indexes

The most basic, yet extremely important, indexes for a network $\mathcal{N}$ are its *order* and its *size*. The order of $\mathcal{N}$ indicates the number of its nodes, whereas the size of $\mathcal{N}$ denotes the number of its edges.

Given a node v_i of the undirected network, $\mathcal{N}$, the number of edges connecting v_i to the other nodes of $\mathcal{N}$ represents the *degree* of v_i. If $\mathcal{N}$ is *directed*, the number of edges incoming to and outgoing from v_i represents the *indegree* and the *outdegree* of v_i, respectively.

If $\mathcal{N}$ is *undirected*, it is possible to consider the *mean degree*, i.e., the ratio of the sum of the degrees of its nodes to the number of nodes. If $\mathcal{N}$ is *directed*, it is possible to define the *mean indegree* and the *mean outdegree* of $\mathcal{N}$ in an analogous fashion.

If $\mathcal{N}$ is unweighted, its *density* is simply the ratio of its edges to the number of all its possible edges. Recall that the number of all possible edges of $\mathcal{N}$ is $\frac{|V|\cdot(|V|-1)}{2}$, if $\mathcal{N}$ is undirected, whereas it is $|V| \cdot (|V| - 1)$, if $\mathcal{N}$ is directed. If $\mathcal{N}$ is weighted, its *density* can be defined as the ratio of the sum of the weights of the existing edges to the number of all its possible edges.

Example 2.1 (...cnt'd)

Consider the undirected network of Figure 1. Its order is 6 and its size is 8. The degree of node A of this network is 4. The mean degree of the network is 2.6, whereas its density is 0.53.

Consider, now, the directed network of the same figure. Its order is 6 and its size is 8. The indegree of node D is 2, whereas its outdegree is 1. The mean indegree and the mean outdegree of the network are 1.33 and 1.33, respectively. Finally, its density is 2.90.

Given a network, $\mathcal{N}$, a *walk* of $\mathcal{N}$ consists of an alternating sequence of nodes and edges that begins and ends with a node. If the starting and the ending nodes of a walk are different, it is said to be *open*; otherwise, it is said to be *closed*. If no node is visited twice, a walk is said to be *simple*.

Given a network, $\mathcal{N}$, a *path* is an open simple walk. A *cycle* is a closed simple walk. A *trail* is a walk that includes each edge no more than once. A *tour* is a closed trail, i.e. a walk that starts and

ends at the same node and has no repeated edges.

If $\mathcal{N}$ is unweighted, the *length of a walk* of $\mathcal{N}$ consists of the number of its edges. If $\mathcal{N}$ is weighted, the length of a walk of $\mathcal{N}$ is the sum of the weights of its edges. Given two nodes v_i and v_j of $\mathcal{N}$, their *geodesic distance* is the length of the shortest path from v_i to v_j. Given a node v_i, the *eccentricity* of v_i is the maximum geodesic distance between v_i and any other node of $\mathcal{N}$. Finally, the *radius* of $\mathcal{N}$ is the minimum eccentricity over all its nodes, whereas, if $\mathcal{N}$ is connected, the *diameter* of $\mathcal{N}$ is the maximum eccentricity over all its nodes.

Example 2.1 (...cnt'd)

Consider the directed weighted network of Figure 1. An example of a walk is the one linking nodes $A-D-E-F-A-C$. This is an open walk. Instead, an example of a closed walk is $B-A-C-B$. Since no node is crossed twice, this walk is simple.

The walk $A-C-D$ is an example of a path, whereas the walk $A-C-D-E-F-A$ is an example of a cycle. The walk $A-C-D-E$ is an example of a trail. Finally, the walk $B-A-D-E-F-A-C-B$ is an example of a tour. The length of the walk $A-D-E-F$ is 34. Consider, now, nodes C and A. Their shortest path is $C-B-A$. The distance of this shortest path is 3 and represents the geodesic distance from C to A. The eccentricity of the node C is 16. Finally, the diameter of this network is 62.

3.2 Centrality Indexes

Centrality indexes aim at measuring power, influence, or other similar features, for the nodes of a network. In real life, there is an agreement on the fact that power and influence are strictly related to relationships. By contrast, there is much less agreement about what power and influence mean. Therefore, several centrality indexes have been proposed to capture the different meanings associated with the term "power".

3.2.1 Degree Centrality

In a network, nodes having more edges to other nodes may have an advantage. In fact, they may have alternative ways to communicate, and hence are less dependent on other nodes. Furthermore, they may have access to more of the resources of the network as a whole. Finally, because they have many edges, they are often third-parties in exchanges among others, and can benefit from this brokerage. As a consequence of all these observations, a very simple, but often very effective, centrality index is node degree. The corresponding form of centrality is called *degree centrality*.

In an undirected network, the degree centrality of a node is exactly the number of its edges. In a directed network, instead, it is important to distinguish centrality based on indegree from centrality based on outdegree. If a node has a high indegree, then many nodes direct edges to it; as a consequence, a high indegree implies *prominence* or *prestige*. If a node has a high outdegree, then it is able to exchange with many other nodes; as a consequence, a high outdegree implies *influence*.

In many real-life networks (e.g., online social networks) degree centrality follows a power-law distribution. This implies that there are few nodes with a high degree centrality and many nodes with a low degree centrality.

3.2.2 Closeness Centrality

A weak point of degree centrality is that it considers only the immediate edges of a node or the edges of the neighbors of a node, rather than indirect edges to all the other nodes. Actually, a node might be linked to a high number of other nodes, but these other nodes might be rather disconnected from the network as a whole. In this case, the original node could be quite central, but only in a local neighborhood.

Closeness centrality emphasizes the distance (or, better, the closeness) of a node to all the other nodes in the network. Depending on the meaning we want to assign to the term "close", a number of slightly different closeness measures can be defined.

In order to compute the closeness centrality of the nodes of a network, first the length of the shortest path between every pair of nodes must be computed. Then, for each node: *(i)* the average shortest distance to all the other nodes is computed; *(ii)* this distance is divided by the maximum distance; *(iii)* the obtained value is subtracted from 1. The result is a number between 0 and 1; the higher this number, the higher the closeness and, consequently, the lower the distance.

As for the distribution of the values of closeness centrality in a network, in many real-life network (e.g., online social networks) few nodes form a long tail on the right but all the other nodes form a bell curve residing at the low end of the spectrum.

3.2.3 Betweenness Centrality

Betweenness centrality starts from the assumption that a node v_i of a network $\mathcal{N}$ can gain power if it presides over a communication bottleneck. The more nodes of $\mathcal{N}$ depend on v_i to make connections with other nodes, the more power v_i has. On the other side, if two nodes are connected by more than one shortest path and v_i is not on all of them, it loses some power. The betweenness centrality of v_i considers exactly the proportion of times v_i is on the shortest path between other nodes; the higher this proportion the higher betweenness centrality.

Betweenness centrality also allows the identification of boundary spanners, i.e., nodes acting as bridges between two or more subnetworks that, otherwise, would not be able to communicate to each other. Finally, betweenness centrality also measures the "stress" (in the sense of a higher usage) which v_i must undergo during the activities of $\mathcal{N}$.

Betweenness centrality can be measured as follows: first, for each pair of nodes of $\mathcal{N}$, the shortest path is computed. Then, for each node v_i of $\mathcal{N}$, the number of the shortest paths, which v_i is involved on, is computed. Finally, if necessary, the obtained results can be normalized to the range $[0, 1]$.

3.2.4 Eigenvector Centrality

Eigenvector centrality starts with the assumption that, in order to evaluate the centrality of a node, v_i, in a network, $\mathcal{N}$, instead of simply adding the number of edges to compute degree, one should weight each of the edges by the degree of the node at the other end of the link (i.e., well connected nodes are worth more than badly connected ones). In this case, v_i is central if it is connected to other nodes that, in turn, are central. A node with a high eigenvector centrality is connected to many nodes that are themselves connected to many nodes.

Eigenvector centrality allows the identification of the so called "gray cardinals", i.e., nodes representing, for instance, advisors or decision makers operating secretely and unofficially. For instance, Don Corleone was a "gray cardinal" because he had an immense power, since he surrounded himself with sons and his trusted "capos", who handled his affairs. By knowing well connected people, "gray cardinals" can use these relationships to reach their objectives while staying largely in the shadow.

The eigenvector centrality of v_i can be computed as follows: *(i)* a centrality score of 1 is assigned to all nodes; *(ii)* the score of each node is computed as a weighted sum of the centralities of all the nodes of its neighborhood; *(iii)* the obtained scores are normalized by dividing them by the largest score; *(iv)* steps *(ii)* and *(iii)* are repeated until the node scores stop changing.

3.2.5 PageRank

PageRank extends the idea of centrality. In fact, instead of outgoing edges, PageRank centrality is determined by incoming edges. PageRank was originally developed for indexing web pages. In fact, it was extensively used by Google to rank web pages. However, it can be applied to all directed networks.

PageRank follows the same ideas as eigenvector centrality, i.e., the PageRank of v_i depends on the number of edges incoming to it, weighted by the PageRank of the nodes at the other end of the edge.

Analogously to what happens for the computation of the eigenvector centrality, the computation of PageRank is iterative. However, differently from Eigenvector Centrality, PageRank computation is local in nature, because only immediate neighbors are taken into consideration; however, its iterative nature allows global influence to propagate through the network, although much more attenuated than in the case of eigenvector centrality. As a consequence of its local nature, the computation of PageRank scales much better to very large networks. Furthermore, at any time, it returns a result, but if more iterations are performed, the quality of results greatly improves.

3.3 Cohesion Indexes

One of the main issues in NA is the identification of cohesive subgroups of actors within a network. Cohesive subgroups are subsets of actors linked by strong, direct, intense, frequent and/or positive relationships. Cohesion indexes aim at supporting the identification of cohesive subnetworks in a network.

To introduce cohesion indexes, we must start with the concept of a subnetwork. A *subnetwork* consists of a subset of nodes of a network and of all the edges linking them. An *ego-network* is a subnetwork consisting of a set of nodes, called "alters", connected to a focal node, called "ego", along with the relationships between the ego and the alters and any relationships among the alters. These networks are important because the analysis of their structure provides information useful to understand and predict the behavior of ego.

3.3.1 Triads

A *triad* is a triple of nodes and of the possible edges existing among them. With *undirected networks*, there are four possible kinds of relationship among three nodes (see Figure 3), i.e., no edges, one edge, two edges or three edges. A triad census aims at determining the distribution of these four

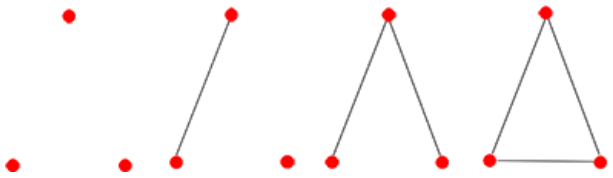

Figure 3: The possible kinds of relationship involving a triad in an undirected network

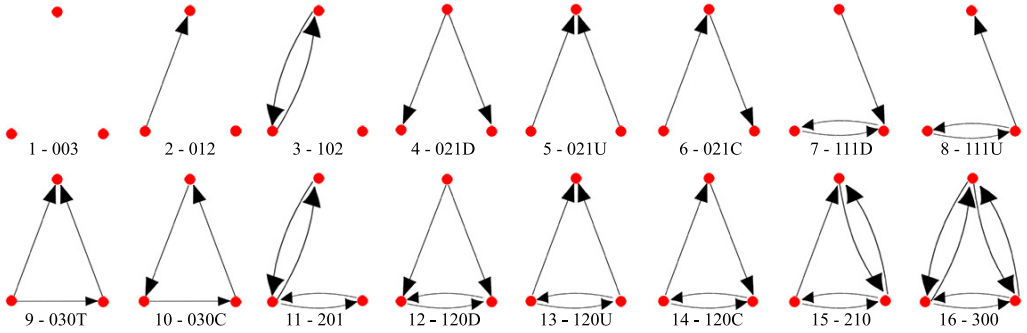

Figure 4: The possible kinds of relationship involving a triad in a directed network

kinds of relationship across all the possible triads. It can give a good approximation of how much a population is characterized by "isolation", "couples only", "structural holes" or "closed triads". In this context, "structural holes" represent a very interesting concept. Given three nodes v_i, v_j and v_k, there exists a structural hole if v_i is connected to v_j, v_j is connected to v_k but v_i is not connected to v_k. Structural holes may have important implications in a network; in fact, they are nodes capable of using and handling asymmetric information; furthermore, they can also bridge two communities. The *ratio between structural holes and closed triads* is a very important index, because, if it is high, then the corresponding network tends to be hierarchic, whereas, if it is low, then the corresponding network tends to be egalitarian.

Starting from triads, it is possible to define clustering coefficient. It is an important indicator of the density of a network. Given a node v, the clustering coefficient of v is defined as the ratio of the number of the existing closed triads involving v to the maximum possible number of triads involving v.

With a *directed network*, there are 16 possible kinds of relationship among three nodes (see Figure 4), including those exhibiting hierarchy, equality, the formation of exclusive groups or clusters. A very important scenario is that of *transitive triads* (i.e., relationships where, if there are edges from v_i to v_j and from v_j to v_k, then there is also an edge from v_i to v_k). Such triads represent the "equilibrium" toward which triadic relationships tend.

3.3.2 Cliques

In its most general definition, a clique is a subnetwork of a network, $\mathcal{N}$, in which its nodes are more closely and intensely linked to each other than they are the other nodes of $\mathcal{N}$. In its most formal and rigorous form, a *clique* is defined as a maximal complete subnetwork of a given network. The smallest

clique is the dyad, consisting of two nodes linked by an edge. Dyads can be extended to become more and more inclusive in such a way as to form strong or closely connected regions in graphs. A clique with four or more nodes consists of several overlapping closed triads and, as such, it inherits many of the properties of closed triads.

Taking into account that the concept of clique is an extension of the concept of a closed triad when the number of involved nodes is higher than 3, it is possible to introduce the concept of a k-clique coefficient as an extension of the concept of clustering coefficient. Specifically, given a node v, the k-clique coefficient of v is defined as the number of existing cliques of k nodes involving the maximum possible number of cliques of k nodes involving v.

This rigorous definition of a clique may be too strong in several situations. Indeed, in some cases, at least some members are not so strongly connected. To capture this cases, the definition of clique can be relaxed.

One way to do so is to define a node as a member of a clique if it is connected to each node of the clique at a distance greater than 1. In this case, the path distance 2 is used. This definition of clique is called *N-clique*[1], where N stands for the maximum length of the allowed path. The definition of N-clique presents some weaknesses. For instance, it tends to return long and stringy N-cliques. Furthermore, N-cliques have properties undesirable for many purposes. For instance, some nodes of N-cliques could be connected by nodes that are not, themselves, members of the N-clique. To overcome this last problem, it is possible to require that the path distance between any two nodes of an N-clique satisfies a further condition, which forces all links among members of an N-clique to occur by way of other members of the N-clique. The structure thus obtained is called *N-clan*.

An alternative way of relaxing the rigorous definition of clique consists of allowing nodes to be members of a clique even if they have edges to all but k other members. In other words, a node is a member of a clique of size N if it has edges to at least $N-k$ nodes of that clique. This relaxation leads to the definition of *k-plex*. While the N-clique approach often leads to large and stringy groupings, the k-plex approach often returns large numbers of smaller groupings. Depending on the goals of the analysis, both N-clique and k-plex could provide useful information about the sub-structure of groups.

A *k-core* is a maximal group of nodes all of whom are connected to at least k other nodes of the group. The k-core approach is more relaxed than the k-plex one; indeed, it allows a node to join the group if it is connected to k other nodes, regardless of how many other nodes it may not be connected to. By varying the value of k different group structures can emerge. K-cores usually are more inclusive than k-plexes.

If a network is weighted, it is possible to introduce a last concept, i.e., the concept of *F-groups*. F-groups return the maximal groups formed by "strongly transitive" and "weakly transitive" triads. A strongly transitive triad exists if there is an edge (v_i, v_j, w_{ij}), an edge (v_j, v_k, w_{jk}) and an edge (v_i, v_k, w_{ik}) and $w_{ij} = w_{jk} = w_{ik}$. A weakly transitive triad exists if w_{ij} and w_{jk} are both higher than w_{ik}, but this last is greater than some cut-off value.

[1]Observe that, in this definition, an N-clique is different from a clique composed by N nodes. In fact, on this definition, an N-clique is a subgraph where all nodes are connected to each other either directly or by means of paths comprising at most N edges.

3.3.3 Components

Components of a network are subnetworks that are internally connected, but disconnected from the other subnetworks. For directed networks, it is possible to define two different kinds of component. A weak component is a set of nodes that are connected, regardless of the direction of edges. A strong component also considers the direction of edges.

Rather as the strict definition of clique may be too strong to capture the concept of a maximal group, the notion of component may be too strong to capture all the meaningful weak points, holes and locally low density sub-parts of a larger network. Therefore, also for components, some more flexible definitions have been proposed. Due to space limitations, we do not illustrate these definitions in detail. The interested reader can find them in [26].

3.4 Other Indexes

In this section, we present some other concepts about NA that can be very useful in Bioinformatics and Computational Biology. The first concept regards the diffusion in some kinds of network, like online social networks. Several past studies have shown that the diffusion rate in a network is initially linear. However, if a *critical mass* is reached, this rate becomes exponential until the network is saturated. The same investigations have shown that the critical mass is reached when about 7% of nodes are reached by the diffusion process. From an economic point of view, in a diffusion process, critical mass is reached when benefits start outweighing costs. If benefits do not balance costs, the critical mass is not obtained, and the diffusion process itself will eventually fail. This concept is also valid in contexts closer to biomedicine. For instance, it is valid to model the spread of an epidemic in a population.

If diffusion regards information, there are several indexes that can help to foresee if a node v_i will contribute to the diffusion process. These indexes are: *(i) relevance* (does v_i care at all?); *(ii) saliency* (does v_i care right now?); *(iii) resonance* (does the information content agrees with what the actor associated with v_i believes in?); *(iv) severity* (how good or bad is the information content?); *(v) immediacy* (does information require an immediate action?); *(vi) certainty* (does information cause pain or pleasure?); *(vii) source* (which node did the information come from and does v_i trust this source?); *(viii) entertainment value* (is the information funny?).

To understand the behavior of actors in a network, a key concept is *homophily*. It states that two actors, who share some properties, will more likely form links than two actors, who do not. Other ways to express the same concept state that: *(i)* two actors that are very close to a third one in a network often tend to link to each other; *(ii)* two actors sharing attributes are likely to be nearer to one another in networks.

In some kinds of networks, such as online social networks or epidemiological networks, homophily is a major force, which, if left alone, would lead communities to become excessively uniform and, at the same time, extremely segregated. To avoid this risk, two important elements act in real life, i.e., curiosity and weak ties. In particular, it has been shown that weak ties are much more powerful than strong ties in stimulating innovation in the behavior of an actor or of a whole network.

4 Closing Remarks

In this chapter, we have provided a presentation of several graph indexes and descriptors. We have seen that network analysis is largely employed in bioinformatics and biomedicine. Then, we have illustrated the most common network representations proposed in the past. Finally, we have presented a large variety of both basic and advanced indexes and descriptors. We think that the usage of graph-based indexes and descriptors in bioinformatics did not come to an end. On the contrary, in the future, the availability of large amounts of data in these contexts, along with the development of more and more powerful hardware, will lead to more and more complex and effective approaches for facing the new challenges that will appear in these sectors.

Acknowledgement

This work was partially supported by Aubay Italia S.p.A.

References

[1] S. Achard, R. Salvador, B. Whitcher, J. Suckling, and E.D. Bullmore. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of Neuroscience*, 26(1):63–72, 2006. Soc Neuroscience.

[2] S.O. Aral. Sexual network patterns as determinants of std rates: paradigm shift in the behavioral epidemiology of stds made visible, 1999.

[3] A.T. Balaban. Applications of graph theory in chemistry. *Journal of chemical information and computer sciences*, 25(3):334–343, 1985. ACS Publications.

[4] S.C. Basak. Philosophy of mathematical chemistry: A personal perspective. *HYLE–International Journal for Philosophy of Chemistry*, 19(1):3–17, 2013.

[5] T. Becker, M. Leese, P. McCrone, P. Clarkson, G. Szmukler, and G. Thornicroft. Impact of community mental health services on users' social networks. PRiSM Psychosis Study. 7. *The British Journal of Psychiatry*, 173(5):404–408, 1998. RCP.

[6] L. Berkman. Assessing the physical health effects of social networks and social support. *Annual Review of Public Health*, 5(1):413–432, 1984. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

[7] L. Berkman and T. Glass. Social integration, social networks, social support, and health. *Social Epidemiology*, 1:137–173, 2000.

[8] L.F. Berkman, I. Kawachi, and M.M. Glymour. *Social Epidemiology*. Oxford University Press, 2014.

[9] S. Borgatti and P. Foster. The network paradigm in organizational research: A review and typology. *Journal of Management*, 29(6):991–1013, 2003. Sage Publications Sage CA: Thousand Oaks, CA.

[10] J. Cassel. The contribution of the social environment to host resistance. *American Journal of Epidemiology*, 104(2):107–123, 1976. Oxford University Press.

[11] H. Chen, L. Ding, Z. Wu, T. Yu, L. Dhanapalan, and J. Chen. Semantic web for integrated network analysis in biomedicine. *Briefings in Bioinformatics*, 10(2):177–192, 2009. Oxford Univ Press.

[12] M. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A. Carvunis, N. Simonis, and J. Rual. Literature-curated protein interaction datasets. *Nature Methods*, 6(1):39–46, 2009. Nature Publishing Group.

[13] J.C. Dearden. The Use of Topological Indices in QSAR and QSPR Modeling. In *Advances in QSAR Modeling*, pages 57–88. 2017. Springer.

[14] G. del Rio, D. Koschutzki, and G. Coello. How to identify essential genes from molecular networks? *BMC Systems Biology*, 3(1):102, 2009. BioMed Central.

[15] E. Estrada. Generalized walks-based centrality measures for complex biological networks. *Journal of Theoretical Biology*, 263(4):556–565, 2010. Elsevier.

[16] G. Eysenbach. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *Journal of Medical Internet Research*, 10(3):e22, 2008. JMIR Publications Inc.

[17] S. Friedman and S. Aral. Social networks, risk-potential networks, health, and disease. *Journal of Urban Health*, 78(3):411–418, 2001. Springer.

[18] S. Friedman, A. Neaigus, B. Jose, R. Curtis, M. Goldstein, G. Ildefonso, R. Rothenberg, and D. Des Jarlais. Sociometric risk networks and risk for HIV infection. *American Journal of Public Health*, 87(8):1289–1296, 1997. American Public Health Association.

[19] T. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. Chandrika, S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, and B. Periaswamy. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293, 2006. Nature Publishing Group.

[20] M. Ghasemi, H. Seidkhani, F. Tamimi, M. Rahgozar, and A. Masoudi-Nejad. Centrality measures in biological networks. *Current Bioinformatics*, 9(4):426–441, 2014. Bentham Science Publishers.

[21] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Science of the United States of America*, 99(12):7821–7826, 2002.

[22] M. Greicius, B. Krasnow, A. Reiss, and V. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, 2003. National Acad Sciences.

[23] X. Guardiola, A. Diaz-Guilera, C. Perez, A. Arenas, and M. Llas. Modeling diffusion of innovations in a social network. *Physical Review E*, 66(2):026121, 2002. APS.

[24] M. Hahn and A. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005. SMBE.

[25] J. Han. Understanding biological functions through molecular networks. *Cell research*, 18(2):224–237, 2008. Nature Publishing Group.

[26] R. Hanneman and M. Riddle. *Introduction to social network methods.* `http://faculty.ucr.edu/~hanneman/nettext/` , 2005. University of California, Riverside.

[27] S. Haustein, I. Peters, C. Sugimoto, M. Thelwall, and V. Lariviere. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, 65(4):656–669, 2014. Wiley Online Library.

[28] C. Hawn. Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health Affairs*, 28(2):361–368, 2009. Health Affairs.

[29] J. House, K. Landis, and D. Umberson. Social relationships and health. *Science*, 241(4865):540, 1988. The American Association for the Advancement of Science.

[30] C. Hsu, H. Juan, and H. Huang. Characterization of microRNA-regulated protein-protein interaction network. *Proteomics*, 8(10):1975–1979, 2008. Wiley Online Library.

[31] D. Huang, B. Sherman, and R. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009. Nature Publishing Group.

[32] O. Ivanciuc and A.T. Balaban. The graph description of chemical structures. *Topological indices and related descriptors in QSAR and QSPR*, pages 59–167, 1999.

[33] A. Jolly, S. Muth, J. Wylie, and J. Potterat. Sexual networks and sexually transmitted infections: a tale of two cities. *Journal of Urban Health*, 78(3):433–445, 2001. Springer.

[34] B. Junker. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, 7(1):219, 2006. BioMed Central.

[35] B. Kaplan, J. Cassel, and S. Gore. Social support and health. *Medical Care*, 15(5):47–58, 1977. JSTOR.

[36] N. Kapucu. Interorganizational coordination in dynamic context: Networks in emergency response management. *Connections*, 26(2):33–48, 2005.

[37] E. Katz and P. Lazarsfeld. *Personal influence*. 1955. New York: Free Press.

[38] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, and C. Duran. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012. Oxford Univ Press.

[39] R. Kessler, R. Price, and C. Wortman. Social factors in psychopathology: Stress, social support, and coping processes. *Annual Review of Psychology*, 36(1):531–572, 1985. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

[40] D. Knoke and S. Yang. *Social Network Analysis*, volume 154. Sage, 2008.

[41] P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008. BioMed Central.

[42] L. Laranjo and A. Arguel. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. *Journal of the American Medical Informatics Association*, pages amiajnl–2014, 2014. The Oxford University Press.

[43] S.J. Leischow and B. Milstein. Systems thinking and modeling for public health practice, 2006.

[44] P. Librado and J. Rozas. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11):1451–1452, 2009. Oxford Univ Press.

[45] N. Lin, X. Ye, and W. Ensel. Social support and depressed mood: a structural analysis. *Journal of Health and Social Behavior*, pages 344–359, 1999. JSTOR.

[46] D. Luke and J. Harris. Network analysis in public health: history, methods, and applications. *Annual Review of Public Health*, 28:69–93, 2007. Annual Reviews.

[47] M.Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. 2011. O'Reilly Media, Inc.

[48] N. Mueller, M. Krauss, and D. Luke. Interorganizational Relationships Within State Tobacco Control Networks: A Social Network Analysis. *Preventing Chronic Disease*, 1(4), 2004. Centers for Disease Control and Prevention.

[49] A. Ozgur, T. Vu, G. Erkan, and D. Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. 24(13):i277–i285, 2008. Oxford Univ Press.

[50] D.H. Rouvray and R Bruce R.B. King. *Topology in chemistry: Discrete mathematics of molecules*. 2002. Elsevier.

[51] K. Roy, S. Kar, and R.N. Das. *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. 2015. Academic Press.

[52] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010. Elsevier.

[53] D. Scanfeld, V. Scanfeld, and E. Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, 38(3):182–188, 2010. Elsevier.

[54] J. Scott. *Social Network Analysis*. Sage, 2012.

[55] T. Sevimoglu and K. Arga. The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, 11(18):22–27, 2014. Elsevier.

[56] K. Supekar, V. Menon, D. Rubin, M. Musen, and M.D. Greicius. Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. *PLoS Computational Biology*, 4(6):e1000100, 2008. Public Library of Science.

[57] R. Todeschini and V. Consonni. *Molecular descriptors for chemoinformatics*, volume 41. John Wiley & Sons, 2009.

[58] T. Valente. *Network models of the diffusion of innovations*. 1995. Hampton Press.

[59] T. Valente and R. Davis. Accelerating the diffusion of innovations using opinion leaders. *The Annals of the American Academy of Political and Social Science*, 566(1):55–67, 1999. Sage Publications Sage CA: Thousand Oaks, CA.

[60] T. Valente and R. Fosados. Diffusion of innovations and network segmentation: the part played by people in promoting health. *Sexually Transmitted Diseases*, 33(7):S23–S31, 2006. LWW.

[61] S. Wasserman and J. Galaskiewicz. *Advances in social network analysis: Research in the social and behavioral sciences*, volume 171. Sage Publications, 1994.

[62] H. Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20, 1947. ACS Publications.

[63] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T. Makela, and S. Hautaniemi. Integrated network analysis platform for protein-protein interactions. *Nature Methods*, 6(1):75–77, 2009. Nature Publishing Group.

[64] K. Wu, Y. Taki, K. Sato, Y. Sassa, K. Inoue, R. Goto, K. Okada, R. Kawashima, Y. He, and A. Evans. The overlapping community structure of structural brain network in young healthy individuals. *PLoS One*, 6(5):e19608, 2011. Public Library of Science.

[65] W. Xu, I. Chiu, Y. Chen, and T. Mukherjee. Twitter hashtags for health: applying network and content analyses to understand the health knowledge sharing in a Twitter-based community of practice. *Quality & Quantity*, 49(4):1361–1380, 2015. Springer.

[66] J. Yoon, A. Blumer, and K. Lee. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics*, 22(24):3106–3108, 2006. Oxford Univ Press.

[67] A. Zalesky, A. Fornito, and E. Bullmore. Network-based statistic: identifying differences in brain networks. *Neuroimage*, 53(4):1197–1207, 2010. Elsevier.

[68] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1128, 2005.