Overview

An overview of social network analysis

Márcia Oliveira and João Gama*

Data mining is being increasingly applied to social networks. Two relevant reasons are the growing availability of large volumes of relational data, boosted by the proliferation of social media web sites, and the intuition that an individual's connections can yield richer information than his/her isolate attributes. This synergistic combination can show to be germane to a variety of applications such as churn prediction, fraud detection and marketing campaigns. This paper attempts to provide a general and succinct overview of the essentials of social network analysis for those interested in taking a first look at this area and oriented to use data mining in social networks. © 2012 Wiley Periodicals, Inc.

How to cite this article:
WIREs Data Mining Knowl Discov 2012, 2: 99–115 doi: 10.1002/widm.1048

INTRODUCTION

The world is a complex system of interconnected parts. Each part itself constitutes a smaller system whose networked structure can be, most of the times, analyzed through the lens of social network analysis (SNA).

SNA is an interdisciplinary methodology research area with contributions from Sociology, Social Psychology, Anthropology, Physics, Mathematics, Computer Science, among others, being a rich scientific field that has significantly benefited from the collaborative efforts of researchers from different scientific areas. Because networks were studied independently by distinct disciplines, for a considerable amount of time, each one developed its own jargon. To avoid ambiguity and clarify the adopted language, in Table 1 we present the network terminology used in different fields. Throughout this document, we will use these terms interchangeably.

The origins of SNA, as a basis for developing useful sociological concepts, can be traced back to the early 1930s, when Moreno[1] developed the sociometric approach as a way to conceptualize the structure of the social relations established among small groups of individuals. These interpersonal ties between members of a group were depicted using the so-called *sociograms*, which can be defined as charts where individuals are represented as nodes and the relations among them are represented by lines. Such diagrams revealed to be very useful in uncovering the hidden structures of groups, by means of the identification of, for instance, *stars*, alliances, and subgroups.

In a broader sense, a social network is constructed from relational data and can be defined as a set of social entities, such as people, groups, and organizations, with some pattern of relationships or interactions between them. These networks are usually modeled by graphs, where vertices represent the social entities and edges represent the ties established between them. The underlying structure of such networks is the object of study of SNA. SNA methods and techniques were thus designed to discover patterns of interaction between social actors in social networks.

Hence, the focus of SNA is on the relationships established between social entities rather in the social entities themselves. In fact, the main goal of this technique is to examine both the contents and patterns of relationships in social networks to understand the relations among actors and the implications of these relationships.

Common tasks of SNA involve the identification of the most influential, prestigious, or central actors, using statistical measures; the identification of hubs and authorities, using link analysis algorithms, and the discovery of communities, using community detection techniques. These tasks are extremely useful

*Correspondence to: jgama@fep.up.pt
Faculty of Economics, University of Porto, Porto, Portugal; The Laboratory of Artificial Intelligence and Decision Support, Institute for Systems and Computer Engineering of Porto, University of Porto, Porto, Portugal.

DOI: 10.1002/widm.1048

TABLE 1 | Network Terminology for Different Fields of Knowledge

Mathematics	Computer Science	Sociology	Physics	Others
Vertex/vertices	Node	Actor/agent	Site	Dot
Edge	Link/connection	Relational tie	Bond	Arc

in the process of extracting knowledge from networks and, consequently, in the process of problem solving. Because of the appealing nature of such tasks and to the high potential opened by this kind of analyses, SNA has become a popular approach in a myriad of fields, from Biology to Business. For instance, some companies use SNA to maximize positive word of mouth of their products by targeting the customers with higher network value (those with higher influence and support).[2–4] Other companies, such as the ones operating in the sector of mobile telecommunications, apply SNA techniques to the phone call networks and use them to identify customer's profiles and to recommend personalized mobile phone tariffs, according to these profiles. These companies also use SNA for churn prediction, i.e., to detect customers who may potentially switch to another mobile operator by detecting changes in the patterns of phone contacts.[5,6] Another interesting application emerges from the domain of fraud detection. For instance, SNA can be applied to networks of organizational communications (e.g., Enron company dataset) to perform an analysis of the frequency and direction of formal/informal email communication, which can reveal communication patterns among employees and managers. These patterns can help identify people engaged in fraudulent activities, thus promoting the adoption of more efficient forms of acting toward the eradication of crime.[7,8]

Besides social networks, there are other types of real-world structures that can be represented by networks. According to Newman,[9] real-world networks can be categorized into four main types: social networks, information networks (or knowledge networks), technological networks, and biological networks.

As previously mentioned, *social networks* are the ones that arise as a result of human and social interactions and encompass studies of friendship networks,[10] informal communication networks within companies,[11] collaboration networks[12] (e.g., networks of coappearance in movies by actors, in which two actors are connected if they appeared together in a movie, and networks of coauthorship among academics, in which individuals are linked if they coauthored one or more papers), among others.

In turn, *information networks* are based upon the exchange of information among entities usually aiming to enhance knowledge diffusion, business, or social aims. Examples include networks of citations between academic papers, commonly represented by an acyclic-directed graph where vertices represent papers and there is a direct edge if paper *A* cites paper *B*; and preference networks, which are generally modeled through bipartite graphs and represent individuals' consumption preferences for a given commercial product[13] (e.g., books). Another important example of an information network is the World Wide Web, which can be represented as a directed graph, in which vertices represent static Web pages and edges correspond to the hyperlinks between them.[14]

Technological networks are man-made networks designed for distribution of some commodity or resource (e.g., electricity, information). Some examples are networks of roads and railways, networks of airline routes, and networks of physical connections between computers (Internet).

The last type of networks are the so-called *biological networks*[15] and, as the name implies, are those that arise from biological processes, such as networks of chemical reactions among metabolites, protein interaction networks, genetic regulatory networks, real neural networks, and food webs or predator–prey networks.

Despite the fact the origins of network studies go back a few centuries ago, in recent years we witnessed an impressive advance in network-related fields, mainly because of the growing interest in social networks, which became a 'hot' topic and a focus of considerable attention. For this reason, a lot of students, practitioners and researchers are willing to enter the field and explore, even superficially, the potential of SNA techniques for the study of their problems. Bearing this in mind, in this paper our aim is to provide a general and succinct overview of the essentials of SNA for those interested in knowing more about the area and strongly oriented to use SNA in practical problems.

The remainder of this document is organized as follows. We begin by pointing out some types of representations that can be used to model social networks. Then, we introduce the best known statistical measures to analyze them, according to two levels of analysis: the actor level and the network level. Afterward, we talk a little about the link analysis task and explain how it can be used to identify influential and authoritative nodes. Then, we distinguish two important network models and introduce the main properties of real-world networks. Later, we devote a section to the problem of finding communities in

networks. After introducing the main concepts, we provide a list of the most popular SNA software and tools for those readers interested in applying network analysis in professional or academic problems. This overview ends with the identification of the current trends arising in the field of SNA.

REPRESENTATION OF SOCIAL NETWORKS

A social network consists of a finite set of actors and the relations, or ties, defined on them.[16] The established relationships can be of personal, or professional, nature and they can range from casual acquaintance to close familiar bonds. Besides social relations, links can also represent flow of information/goods/money, interactions, similarities, among others. The structure of such networks is usually represented by graphs. Therefore, networks are often regarded as equivalent to graphs.

A graph is composed of two fundamental components vertices and edges. Every edge is defined by a pair of vertices, also called its endpoints. Vertices represent a wide variety of individual entities (e.g., people, organizations, countries, papers, products, plants, and animals) according to the application field. In turn, an edge is the line that connects two vertices and, analogously, it can represent numerous kinds of relationships between individual entities (e.g., communication, cooperation, friendship, kinship, acquaintances, and trade). Edges may be directed or undirected, depending if the nature of the relation is asymmetric or symmetric.

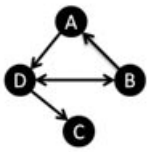
Formally, a graph G consists of a nonempty set $V(G)$ of vertices and a set $E(G)$ of edges, being defined as $G = (V(G), E(G))$. According to Diestel,[17] the *order* of a graph G is given by the total number of vertices n or, mathematically, $|V(G)| = n$. Analogously, the *size* of a graph G is the total number of edges $|E(G)| = m$. The maximum number of edges in a graph is $m_{\max} = \frac{n(n-1)}{2}$, for undirected graphs, and $m_{\max} = n(n-1)$, for directed ones.

In the literature, two main types of graph-theoretic data structures are referred to represent graphs: the first one are *list structures* and the second are *matrix structures*. These structures are appropriate to store graphs to further analyze them using automatic tools. List structures, such as incidence lists and adjacency lists, are suitable for storing sparse graphs because they reduce the required storage space. On the other hand, matrix structures such as incidence matrices ($A_{\downarrow}(n \times m)$), adjacency matrices or sociomatrices ($A_{\downarrow}(n \times n)$), Laplacian matrices (contains both adjacency and degree information), and distance matrices (identical to the adjacency matrices with the difference that the entries of the matrix are the lengths of the shortest paths between pairs of vertices) are appropriate to represent full matrices.

Several types of graphs can be used to model different kinds of social networks. For instance, graphs can be classified according to the direction of their links. This leads us to the differentiation between *undirected* and *directed* graphs. Undirected graphs (or undirected networks) are graphs whose edges connect unordered pairs of vertices or, in other words, each edge of the graph connects concomitantly two vertices. A more strict type of graph is the so-called directed graph (or directed network). Directed graphs, or in the abbreviation form *digraphs*, can be straightforward defined as graphs whose all edges have an orientation assigned (also called *arcs*), so the order of the vertices they link matters. Formally, a directed graph D is an ordered pair $(V(D), A(D))$ consisting of a nonempty set $V(D)$ of vertices and a set $A(D)$, disjoint from $V(D)$, of arcs. If e_{12} is an arc and v_1 and v_2 are vertices such that $e_{12} = (v_1, v_2)$, then e_{12} is said to join v_1 to v_2, being the first vertex v_1 called *initial vertex*, or *tail*, and the second vertex v_2 called the *terminal vertex*, or simply *head*. Graphically, directed edges are depicted by arrows, indicating the direction of the linkage. This type of graphs can be either *cyclic*, i.e., graphs containing closed loops of edges or 'ring' structures, or *acyclic* (e.g., trees). A typical example of an undirected graph is Facebook™ because, in this social network, the established friendship tie is mutual or reciprocal (e.g., if I accept a friend request from a given person then it is implicitly assumed that me and that person are friends of each other). Likewise, Twitter™ is an example of a directed graph because a person can be followed by others without necessarily following them. In this case, the tie between a pair of individuals is directed, with the tail being the follower and the head being the followed, meaning that a one-way relationship is established.

Regarding the values assigned to edges, we can make a distinction between *unweighted* and *weighted* graphs. Unless it is explicitly said, we always assume that graphs are unweighted. Unweighted graphs are binary since edges are either present or absent. On the other hand, weighted graphs are richer graphs because each edge has associated a weight $w \in \mathbf{R}_0^+$ providing the user with more information about, for instance, the strength of the connection of the pair of vertices it joins. According to Mark Granovetter,[18,19] in social networks the weight of a tie is generally a function of duration, emotional intensity, frequency of interaction, intimacy, and exchange of services.

Vertex	A	B	C	D
A	-	0	0	1
B	1	-	0	1
C	0	0	-	0
D	0	1	1	-



Vertex	Vertex
A	D
B	A
B	D
D	B
D	C

FIGURE 1 | A directed graph D represented by means of an adjacency matrix (left-hand side of the figure) and an adjacency list (right-hand side of the figure).

Therefore, strong ties usually represent close friends, and weak ties represent acquaintances. In other kinds of networks, the weight of a tie can represent a variety of things, depending on the context; for instance, a tie can represent the number of seats among airports, the number of exchanged products, and so on.

For undirected and unweighted graphs, adjacency matrices are binary (as a consequence of being unweighted) and symmetric (as a consequence of being undirected, meaning that $a_{ij} = a_{ji}$), with $a_{ij} = 1$ representing the presence of an edge between vertices i and j, and $a_{ij} = 0$ representing the absence of an edge between vertex pair (i,j). For directed and weighted graphs, the entries of such matrices take values from interval $[0, \max(w)]$ and are nonsymmetric. In both cases, we deal with nonnegative matrices.

In Figure 1, we provide an example of how a graph can be represented by an edge list and by an adjacency matrix.

ELEMENTARY STATISTICAL MEASURES

Mathematics is used to represent networks, while Statistics is mainly used to analyze them. In this section, we present some graph measures and popular metrics used in the analysis of social networks that arose from the field of Statistics. These measures are useful in the sense that they provide us insights about the structure of the network without the need to know its graphical representation. Studying the structure of these networks aims at understanding the behavior of the social systems that generated those networks, which is normally the final goal of such analysis.

The measures we will introduce in the following subsections can be divided according to the level of analysis one wants to perform: at the level of small units, such as actors, or at the level of the whole network. The former explores general measures of centrality as a way to understand how the position of a vertex is within the overall structure of the graph and, therefore, helps identify the key players in the network. The latter provides more compact information and allows the assessment of the overall structure of the network, giving insights about important properties of the underlying social phenomena.

Actor-Level Statistical Measures

Centrality, or *prestige*, is a general measure of how the position of an actor is within the overall structure of the social network and can be computed resorting to several metrics. The most widely used are *degree*, *betweenness*, *closeness*, and *eigenvector centrality*. The first three were proposed by Freeman[20] and were only designed for unweighted networks. Recently, Brin and Page[21] came up with extensions to weighted networks. The fourth metric—eigenvector centrality—was later proposed by Bonacich[22] and has its foundations on spectral graph theory. It became especially popular after being used as the basis of the well-known Google's Pagerank algorithm, which we will talk about in the next Section.

Although more actor-level statistical measures were proposed in the literature, in this subsection we will focus on explaining the mentioned measures of centrality. These measures determine the relative importance of an actor within the network, showing how the relationships are concentrated in a few individuals and, therefore, giving an idea about their social power. Higher centrality measures are associated to powerful actors in the network because their central position offers them several advantages, such as easier and quicker access to other actors in the network (useful for accessing resources such as information) and ability of exerting control over the flow between the other actors.[20] These central actors are also called 'focal points'. At the end of the section we will also introduce the concept of *transitivity* and explain how it can be computed using a *clustering coefficient*.

The reader must take into account that some of these actor-level metrics (e.g., degree, betweenness, and closeness) may need to be normalized to perform comparisons of networks with different orders and sizes.

Degree or Valency

The *degree*, or *valency*, of a node v, usually denoted as k_v, is a measure of the immediate adjacency and the involvement of the node in the network and is computed as the number of edges incident on a given node or, similarly, as the number of neighbors of node v. The neighborhood N_v is thus defined by the set of nodes that are directly connected to v. Degree can be computed in, at least, two different ways: based on

the adjacency matrix or based on the neighborhood of a node. In Eqs (1) and (2), we present each one of the alternatives, for undirected networks. Despite its simplicity, degree is an effective measure to assess the importance and influence of an actor in a social network. Yet, it has some limitations. The main one is that it does not take into consideration the global structure of the network:

$$k_i = \sum_{j=1}^{n} a_{ij}, \qquad 0 < k_i < n, \tag{1}$$

where a_{ij} is the entry of the ith row and jth column of the adjacency matrix **A**;

$$k_v = |N_v|, \qquad 0 < k_v < n, \tag{2}$$

where $|N_v|$ is the neighborhood of node v

For directed networks, there are two variants of degree centrality: *in degree*, denoted by k_v^+, and *out degree*, denoted by k_v^-. The former is given by the number of incoming nodes (i.e., number of edges beginning at vertex v) and the latter by the number of outgoing nodes (i.e., number of edges ending at vertex v), as defined in Eqs (3) and (4). The measure of degree in directed networks is also referred to as prestige. This expression is especially used in the literature of social networks because it was developed for measuring the prominence or importance of actors in the network. There are two types of prestige: *support* and *influence*. The first is related to the in-degree centrality, which is seen as a measure of support, and the second is related to the out-degree centrality, which is seen as a measure of influence:

$$k_i^+ = \sum_{j=1}^{n} a_{ji}, \tag{3}$$

$$k_i^- = \sum_{j=1}^{n} a_{ij}. \tag{4}$$

On weighted networks, *strength* is the equivalent of degree, being computed as the sum of the weights of the edges adjacent to a given node, as expressed by Eq. (5):

$$k_i^w = \sum_{j=1}^{n} a_{ij}^w. \tag{5}$$

A significant research effort was undertaken in studying the *degree distribution* of several types of networks, which turned it possible to classify a network based on this distribution. For instance, Barabási and coworkers[23,24] discovered that most real networks follow a *power-law* distribution, at least asymptotically. This means that, in these networks, the distribution of the vertex degree is very heterogeneous and highly right skewed, with a large majority of vertices having a low degree and a small number having a high degree. These networks are known as *scale free*, an expression coined by the same researchers. Other common functional forms are *exponential* (e.g., railways and power grids networks) and *power laws with exponential cutoffs* (e.g., networks of movie actors and some collaboration networks).

Betweenness

Node betweenness b_v measures the extent to which a node lies between other nodes in the network and can be computed using the formula presented in Eq. (6). Nodes with high betweenness occupy critical roles in the network structure, since they usually have a network position that allow them to work as an interface between tightly knit groups, being 'vital' elements in the connection between different regions of the network. In the social networks perspective 'interactions between two nonadjacent actors might depend on other actors in the set of actors, especially the actors who lies on the paths between the two',[16] which stresses out the importance of a good value of betweenness. These actors are also called *gatekeepers* because they tend to control the flow of information between communities.

$$b_v = \sum_{s,t \in V(G)\setminus v} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{6}$$

where σ_{st} denotes the number of shortest paths between vertices s and t (usually $\sigma_{st} = 1$) and $\sigma_{st}(v)$ expresses the number of shortest paths passing through node v.

This quantity can also be computed for edges. The *betweenness of an edge* b_e is commonly defined as the number of shortest paths between nodes that run along a given edge of the network. It is quite useful in SNA since it allows discovering *bridges* and *local bridges* which are, by definition, edges with high betweenness. In the context of SNA, bridges are connections outside an individual's circle of acquaintances. These connections are of great interest for individuals seeking to access new information and resources, since they ease the diffusion of information across entire communities.[25] However, situations like these are quite rare in real-world scenarios and, even if they happen, the advantages they confer are usually temporary, due to the temporal instability of such edges. A more common and realistic situation is local bridges. Equation (7) indicates how this measure can

be computed:

$$b_e = \sum_{u,v \in V(G)} \frac{\sigma_{uv}(e)}{\sigma_{uv}}, \quad (7)$$

where $\sigma_{uv}(e)$ expresses the number of shortest paths passing through edge e. The sum indicates that this fraction needs to be computed for every pair of nodes u and v in the network.

Closeness

Closeness is a rough measure of the overall position of an actor in the network, giving an idea about how long it will take to reach other nodes from a given starting node. Formally, it is the mean length of all shortest paths from one node to all other nodes in the network. Because of its definition, usually this measure is only computed for nodes within the largest component of the network, using the formula presented in Eq. (8). In the social networks context, closeness is a measure of reachability that measures how fast a given actor can reach everyone in the network:

$$Cl_v = \frac{n-1}{\Sigma_{u \in V(G)\setminus v} d(u,v)}. \quad (8)$$

Eigenvector Centrality

This metric is based on the assignment of a relative score to each node and measures how well a given actor is connected to other well-connected actors. This score is given by the first eigenvector of the adjacency matrix. The basic idea behind eigenvector centrality is that the power and status of an actor is recursively defined by the power and status of his/her *alters*. *Alters* is a term frequently used in the egocentric approach of social networks analysis, and it refers to the actors that are directly connected to a specific actor, called *ego*. In other words, we can say that the centrality of a given node i is proportional to the sum of the centralities of i's neighbors. This is the assumption behind the eigenvector centrality formula, which is as follows:

$$x_i \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} x_j, \quad (9)$$

where x_i/x_j denotes the centrality of node i/j, a_{ij} represents an entry of the adjacency matrix $\mathbf{A}$ ($a_{ij} = 1$ if nodes i and j are connected by an edge and $a_{ij} = 0$ otherwise) and λ denotes the largest eigenvalue of $\mathbf{A}$.

Eigenvector centrality is a more elaborated version of the degree, once it assumes that not all connections have the same importance by taking into account not only the quantity, but especially the quality of these connections.

Local Clustering Coefficient

Social networks are naturally transitive, which means that a given actor's friends are also likely to be friends. This property of transitivity is quantified by a *clustering coefficient* that can be global, i.e., computed for the whole network, or local, i.e., computed for each node. Watts and Strogatz[26] proposed a local version of the clustering coefficient, denoted c_i ($i = 1,\ldots, n$). In this context, transitivity is a local property of a node's neighborhood that indicates the level of cohesion between the neighbors of a node. This coefficient is, therefore, given by the fraction of pairs of nodes, which are neighbors of a given node v that are connected to each other by edges [see Eq. (10)]:

$$c_i = \frac{2|e_{jk}|}{k_i(k_i-1)} : v_j, v_k \in N_i,\ e_{jk} \in E, \quad (10)$$

where N_i is the neighborhood of node v_i, e_{jk} represents the edge that connects node v_j to node v_k, k_i is the degree of node v_i, and $|e_{jk}|$ indicates the proportion of links between the nodes within the neighborhood of node v_i.

Network-Level Statistical Measures

Before explaining each one of the network-level statistical measures, there are three fundamental concepts that should be first introduced: *path*, *geodesic distance* between two nodes, and *eccentricity* of a vertex.

A *path* is a sequence of nodes in which consecutive pairs of nonrepeating nodes are linked by an edge; the first vertex of a path is called the *start vertex* and the last vertex of the path is called the *end vertex*. Of particular interest is the concept of *geodesic distance*, or *shortest path*, between nodes i and j, denoted as $d(i,j)$. The *geodesic distance* can be defined as the length of the shortest path, or the minimal path, between nodes i and j.

In turn, the *eccentricity* is the greatest geodesic distance between a given vertex v and any other in the graph, as defined in Eq. (11). These three concepts are formed on the basis of most of the network-level metrics we are going to introduce, namely, the *diameter/radius*, the *average geodesic distance*, the *average degree*, the *reciprocity*, the *density*, and the *global clustering coefficient*.

$$\epsilon_v = \max_{i \in V(G)\setminus v} d(v,i). \quad (11)$$

Diameter and Radius

The *diameter D* is given by the maximum eccentricity of the set of vertices in the network and, analogously,

the *radius* R can be defined as the minimum eccentricity of the set of vertices, as defined in Eqs (12) and (13). Sparser networks have generally greater diameter than full matrices, due to the existence of fewer paths between pairs of nodes. Leskovec et al.[27] discovered that, for certain types of real-world networks, the effective diameter shrinks over time, contradicting the conventional wisdom of increasing diameters. In the context of SNA, this metric gives an idea about the proximity of pairs of actors in the network, indicating how far two nodes are, in the worst of cases:

$$D = \max\{\in_v : v \in V\}, \tag{12}$$

$$R = \min\{\in_v : v \in V\}. \tag{13}$$

Average Geodesic Distance

The *average geodesic distance* for all combinations of vertex pairs in a network is usually denoted by l and is given by Eq. (14). This metric gives an idea of how far apart nodes will be, on average. For instance, in the SNA context the average geodesic distance can be used to measure the efficiency of the information flow within the network:

$$l = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} d(i, j), \tag{14}$$

where $d(i, j)$ is the *geodesic distance* between nodes i and j, and $1/2n(n–1)$ is the number of possible edges in a network comprising n nodes.

When there is the case of a network having more than one connected component, the previous formula does not hold, because the geodesic distance is conventionally defined as infinite when there is no path connecting two vertices. In such situations, it is more appropriate to use the *harmonic average geodesic distance*, defined in Eq. (15), once it turns infinite distances into zero nullifying their effect on the sum:

$$l^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} \frac{1}{d(i, j)}. \tag{15}$$

Average Degree

The *average degree* is simply the mean of the degrees of all vertices in a network, as represented in Eq. (16). According to Costa et al.[28] the average degree can be used to measure the global connectivity of a network:

$$\bar{k} = \frac{1}{n} \sum_{i=1}^{n} k_i. \tag{16}$$

Reciprocity

Reciprocity r is a specific quantity for directed networks that measures the tendency of pairs of nodes to form mutual connections between each other. There are several ways to compute this metric. The most popular and intuitive way is to compute the ratio of the number of mutual connections in the network to the number of all connections, as shown in Eq. (17). Adopting this definition, the value of reciprocity represents the probability that two nodes in a directed network point to each other. By definition, in an undirected network, reciprocity is always maximum $r = 1$ because all pairs of nodes are symmetric:

$$r = \frac{\#\text{mut}}{\#\text{mut} + \#\text{asym}}, \qquad 0 < r < 1, \tag{17}$$

where #mut denotes the number of mutual dyads and #asym the number of asymmetric dyads.

Taking the definitions of Wasserman and Faust,[16] we say that an *asymmetric* dyad is a pair of nodes that has an arc going in the direction of one node or the other, but not both directions. In turn, a mutual dyad is defined by a pair of nodes connected by two arcs, each one going in a different direction (e.g., $a \rightarrow b$ and $b \rightarrow a$, being a and b two nodes in a network).

Density

Density ρ is an important network-level measure, which is able to explain the general level of connectedness in a network. It is given by the proportion of edges in the network relative to the maximum possible number of edges, as defined in Eq. (18). Density is a quantity that goes from a minimum of 0, when a network has no edges at all, to a maximum of 1, when the network is perfectly connected (also called *complete graph* or *clique*). Therefore, high values of ρ are associated to dense networks, and low values of density are associated to sparse networks:

$$\rho(G) = \frac{m(G)}{m_{\max(G)}}, \qquad 0 < \rho < 1, \tag{18}$$

where m is the number of edges in the network and $m_{\max(G)}$ denotes the number of possible edges, which is $\frac{n(n-1)}{2}$ for undirected networks and $n(n–1)$ for directed ones.

Global Clustering Coefficient

There are several ways to compute the global version of the *clustering coefficient*. We adopt the one proposed by Watts and Strogatz[26] that obtains the global clustering coefficient c, for the whole network, through the computation of the average of all local

values c_i ($i = 1,\ldots, n$), as shown in Eq. (19). Small-world networks,[26] such as the ones we find in real social contexts, are characterized by high global clustering coefficients, meaning that the property of transitivity among nodes emerges more often and in a stronger way, increasing the probability of clique formation:

$$c = \frac{1}{n}\sum_i c_i. \tag{19}$$

LINK ANALYSIS

In certain network settings, such as the Web, one may be interested in finding the most valuable, authoritative or influential node (e.g., web page), or a list of them. To perform this task, several link analysis algorithms were devised, being the HITS[29] and the Google Pagerank[30] algorithms the most popular ones. These algorithms explore the relationship between links and the content of web pages, to improve the task of information retrieval in the Web, being of extreme importance for the design of efficient search engines. As the development of these methods was motivated by the problem of web queries, for the sake of simplicity, we will explain them in this context.

Before introducing any of these algorithms, it is first necessary to define some elementary concepts, namely, the concepts of *hubs* and *authorities*. In the context of Web, a hub can be understood as a web page that points to many other web pages or, in other words, as a compilation of web pages that address a specific topic. The quality of a hub is usually determined by the quality of the authorities it points to. On the other hand, authorities are web pages cited by many different hubs, which means that their relevance is measured by the number of inward links they receive. Typically, good authoritative pages are reliable sources of information about a given topic.

In the following subsection, we explain the foundations of Pagerank algorithm.

Pagerank Algorithm

Pagerank is a link analysis algorithm based on the concept of eigenvector centrality. This algorithm is used by Google™ Internet search engine to rank web pages according to the value of the information they carry, so the most valuable ones appear at the top of the search results.

The idea of the algorithm is that information on Web can be ranked according to link popularity (the more web pages are linked to a given web page the more popular that web page is). Nevertheless, in this process of weighting web pages, not only the number of links or, equivalently, the degree of a node is relevant, but also the importance of the web pages linking to them. Therefore, Pagerank measures the relative importance of a set of web pages based not only on the quantity but especially the quality of their links.

The basic Pagerank is computed as follows (according to the definition provided by Easley and Kleinberg[31]):

Initialization: In a network of n nodes (or web pages), assign a Pagerank value of $1/n$ to each node, and choose the number of iterations k of the algorithm.

1. Update the Pagerank values of each node by sequentially applying the following rule: *Basic Pagerank update rule*: divide the actual Pagerank value of node p by the number of its outgoing links and pass these equal *shares* to the nodes it points to. Note that if a node p has no outgoing links, the Pagerank *share* is passed to itself. The update of a node's Pagerank value is performed by summing the shares it receives in each iteration.
2. Apply this rule until the kth iteration, or until convergence has occurred.

To illustrate, consider the following example: in a network comprised six nodes termed A, B, C, D, E, and F. How can we find the most influential node, using the Pagerank algorithm? First, according to the initialization step of the algorithm, each node is assigned an equal Pagerank of $PR = \frac{1}{n} = \frac{1}{6}$, as represented in Figure 2(a). Then, these values are updated k times (for the sake of simplicity, we consider only two iterations) by applying the *basic Pagerank update rule*.

To apply the rule, first is necessary to compute the shares of all nodes. Then, for each node we sum all shares the node receives. The result of this sum will be its new Pagerank value, as shown in Table 2 and Figure 2(b). For instance, the share of node D, which has only one outgoing link, is computed as $share(D) = \frac{1/6}{1} = \frac{1}{6}$. Its new Pagerank value is given by the sum of the shares of its ingoing links, namely, those coming from nodes A, C, E, and F:

$$PR(D) = \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{6} = +\frac{5}{12}. \tag{20}$$

After computing these values for all nodes in the network, we repeat the process for the second

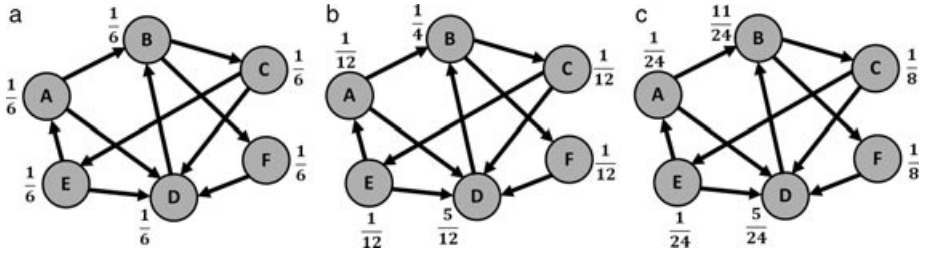


FIGURE 2 | Illustration of the process behind Pagerank algorithm in a network comprised six nodes. The first network (a) corresponds to the initialization step. In network (b) are shown the updated Pagerank values at the end of the first iteration of the algorithm. Note that node D is so far the most authoritative node, with a Pagerank value of 5/12.The rightmost network (c) corresponds to the second (and last) iteration of Pagerank. Here, we notice that node B overtakes the position of node D in terms of Pagerank values.

TABLE 2 | Updated Pagerank Values after the First Iteration $k = 1$

Node	A	B	C	D	E	F
Shares	1/12	1/12	1/12	1/6	1/12	1/6
Updated Pagerank	1/12	1/4	1/12	5/12	1/12	1/12

TABLE 3 | Updated Pagerank Values at the End of the Second (and Last) Iteration $k = 2$

Node	A	B	C	D	E	F
Shares	1/24	1/8	1/24	5/12	1/24	1/12
Updated Pagerank	1/24	11/24	1/8	5/24	1/24	1/8

iteration $k = 2$, obtaining the results shown in Table 3 and Figure 2(c). This rule is applied iteratively until the convergence of the Pagerank values, or until the kth iteration. As we consider only two iterations, we can try to draw some conclusions and interpret the results based only on the information available in Tables 2 and 3. Therefore, at the end of the first iteration, node D seemed to be the most promising one, with a Pagerank of 5/12; nevertheless, at the second iteration node B overtakes the position of D, being now assigned to the first place of the ranking of nodes. This sudden change befits the idea behind Pagerank algorithm that measures the quality, instead of the quantity, of a node's connections. Therefore, and besides node D is the one receiving more incoming links, the importance of the nodes linking to them is not that significant. On the other hand, node B has only two incoming links, but one of them is of great importance, namely, node D. This is the main reason why node B receives the larger Pagerank value at the end of the second iteration, turning into the most influential, or authoritative, node in the network. If B was an actor, he/she would be considered the most important one, once this Pagerank value means that a great part of the information that flows through the network passes through it.

Although Pagerank and other link analysis algorithms were originally motivated by the necessity of extracting and understanding the information yielded by Web, they are also used in other domains, such as social sciences. In the social sciences field, links can be analyzed from two distinct, but somehow interrelated, perspectives: the information centered and the actor centered. These perspectives are typically used to help understand the underlying social phenomena, by means of the identification of the most valuable sources of information or, alternatively, the most important actors. Nevertheless, there is still some lack of consensual guiding principles about how to interpret link analysis results in a social science context. Thelwall[32] stresses out the importance of developing guidelines for improving the process of interpreting these results and proposes a theoretical framework for link analysis interpretation.

PROPERTIES OF REAL-WORLD NETWORKS

Real-world problems are an inexhaustible source of inspiration for network theories. The great majority of real-world events and activities we play, observe, and study can be easily modeled using graphs and can be analyzed through the lens of network analysis.

In this overview, we focus on social networks, which are those arising as a result of human and social interactions, though there are other types of real-world networks. Given the diversity of such networks, researchers classified them into main types, although this classification is not fully consensual. One possibility is the one provided by Newman,[9] which was presented in the Introduction section. Although they stem from distinct real-world problems and knowledge

fields, they all share a set of common properties which make them peculiar, thus opposing to the two well-known network models: random networks and regular networks.

The simplest and best known model of network is the *random graph*.[33,34] This type of graph is characterized by the random placement of edges between a fixed number n of vertices, to create a network in which each of the $1/2n(n–1)$ possible edges is independently present with some probability p. When $p = 0$, we obtain a graph of perfect order—*regular graph*; and when $p = 1$, we obtain a random graph, which embodies the total chaos. Because both regular graphs and random graphs represent extremes, they are not realistic. Additional properties[9] are required to model complex and atypical networks such as the ones we find in real world. In short, we can say that real-world networks are nonrandom and nonregular graphs with unique features, where 'order coexists with disorder'.[35] In this section, we introduce and explain some of these properties, which are as follows:

1. Small-world effect;
2. Transitivity or clustering;
3. Power-law degree distributions;
4. Network resilience;
5. Mixing patterns;
6. Community structure.

Property 1: the Small-World Effect

Stanley Milgram,[36] an American social psychologist, was the first to point out the existence of small-world effects in real social networks, through a series of famous experiments which are today known as the *Milgram experiment*. This experiment was done to test the speculative idea of the small-world effect and is one of its first direct demonstrations. The main hypothesis of the study was that pairs of apparently distant individuals are connected by a short path, i.e., by a few number of acquaintances, through the network. To probe the distribution of the path lengths, Milgram asked some random participants (about 300) to pass a letter to someone they knew in a first-name basis in an attempt to get it to an assigned target person. With this experiment, it was shown that the median length of the paths that succeeded in reaching the target was six, which clearly explains the origins of the concept *six degrees of separation*.

The overall conclusion of Milgram and its colleagues has been accepted in a broad sense. In fact, the small-world effect has been widely observed in real-world networks and it is manifested by the existence of shortcuts between most of vertex pairs in a network. In social settings, this means that two apparently disconnected people can quickly get in touch with each other through an incredible low number of acquaintances or friends. This finding has several implications in dynamic processes because it implies, for instance, that the spread of a contagious disease throughout the population will be faster than one would expect.

In mathematical terms, the small-world effect means that the average geodesic distance (i.e., the shortest path) between pairs of vertices scales logarithmically, or slower, with the network size with a fixed mean degree.[9] This property is also observed in random graphs, where the diameter is very small, only growing logarithmically with n, and the vertices have all about the same degree.

Property 2: Transitivity or Clustering

According to Wasserman and Faust,[16] *transitivity* is a property that considers *triples of nodes* (i.e., sets of three vertices, in which at least one is connected to both others) in a graph or, in other words, measures the density of *triangles* (i.e., three nodes connected to each other by three edges in the network, which means that every node is fully connected to the remaining two nodes) in a network. In the social network parlance, it means that a friend of your friend is also likely to be your friend.

This property is quantified using a *clustering coefficient* that can be global or local, as mentioned in the section *Elementary Statistical Measures*.

Property 3: Power-Law Degree Distributions

The *degree distribution* $P(k)$ is the probability distribution of the degrees of nodes over the whole network. Therefore, $P(k)$ represents the probability that a vertex chosen uniformly at random has degree k, and is defined by the fraction of nodes in the network that have degree k. This means that, if the total number of nodes in the network is n, and n_k of these nodes have degree k then, for this value of the degree, we have a probability of $P(k) = \frac{n_k}{n}$. Computing this probability for each degree value k, of the set of degree values appearing in a given network, we obtain the probability distribution of the degree in this network.

Random graphs, such as the ones studied by Erdös and Rényi,[34] show a binomial degree distribution because the presence, or absence, of an edge is equiprobable (i.e., equal for all possible vertex pairs). In the limit of large graph size, this degree

distribution goes from binomial to Poisson distribution. Therefore, in this class of graphs, the degree distribution is highly homogeneous as most vertices have similar, or equal, degree.

Real-world networks are, in turn, quite different from random graphs with respect to degree distribution. Barabási and Albert[24] discovered that, in real graphs, the distribution of the vertex degree is very heterogeneous and highly right-skewed, with a large majority of vertices having a low degree and a small number having a high degree. This finding comes to reinforce the previous work of Price[37] on networks of citations between scientific papers. In both cases, they state that the degree distribution of real networks, such as citation networks, follows a power-law (at least asymptotically) and, therefore, these networks are sometimes referred to as *scale-free networks*.[23] Power-law distributions usually arise when the amount you get of something depends on the amount you already have. A common analogy is 'the rich get richer'. Price[38] used the term *cumulative advantage* to refer to this mechanism, which is believed to be the most probable explanation for the power-law degree distributions in several real-world networks, which include, but are not restricted to, collaboration networks and the World Wide Web. Today, this process is best known as *preferential attachment*, a name coined by Barabási and Albert.[24] In their seminal paper,[24] the authors describe a network growth model which became known as the *Barabási–Albert model*. This work shows that network growing with preferential attachment will indeed become scale free, due to the 'the-rich-gets-richer' strategy employed in the model.

Property 4: Network Resilience

Network resilience measures the impact on the connectivity of the network when one or more vertices are removed and it is an indicator of the cohesion of the network. Different kinds of networks exhibit different levels of resilience. Most networks are robust against random vertex removal but considerably less robust to targeted removal of the highest degree vertices. Also, when the endpoints of a bridge are deleted there are strong changes in the network with respect to the ability of communication between pairs of vertices, as some of them become disconnected. Betweenness centrality can also be seen as a measure of resilience as it tells us how many geodesic paths will get longer when a vertex is removed from the network. However, in real settings the removal of a single node is not usually cause for alarm, since the networks comprise millions or even billions of nodes. In such cases, it is more appropriate to test the resilience of the network the removal of a given percentage of nodes.

Property 5: Mixing Patterns

Some networks are made up of different types of vertices. In these kinds of networks, the linking between vertices or, in other words, the probability of connection between vertex pairs, tends to be selective and highly dependent on vertex types (e.g., food web, whose vertices may be herbivores, carnivores, and plants). In social networks, this is also evident because individuals tend to interact with other similar to them. This selective linking is usually called *assortative mixing*, or *homophily*, and a classic example is mixing by race. Real networks tend to show higher tendencies for assortative mixing.

Newman[39] proposed an *assortative coefficient* to quantify the assortative mixing of a network. This coefficient replaces the previously proposed by Gupta et al.[40] and allows us to distinguish a randomly mixed network from a perfectly assortative one.

Property 6: Community Structure

The great majority of real social networks show community structure, which means that we can find groups of densely connected vertices that are low connected to other groups of vertices in the network. This topic will be deepened in the following section.

COMMUNITY DETECTION

One of the unique features of social networks is that they tend to show *community structure*. This property usually arises as a consequence of both global and local heterogeneity of edges distribution in a graph. Thus, we often find high concentrations of edges within certain regions of the graph, called *communities*, and low concentration of edges between those regions.

Communities, also known as *modules* or *clusters*, can be straightforward defined as similar groups of nodes. A more complete definition is built upon the concept of density: communities can be understood as densely connected groups of vertices in the network, with sparser connections between them.

According to Newman and Girvan,[41] there are two main lines of research in discovering communities in network data. The first has its origins in Computer Science and is known as *graph partitioning*, whereas the second has been mainly pursued by sociologists and is usually referred as *blockmodeling*, *hierarchical clustering*, or *community structure detection*. The

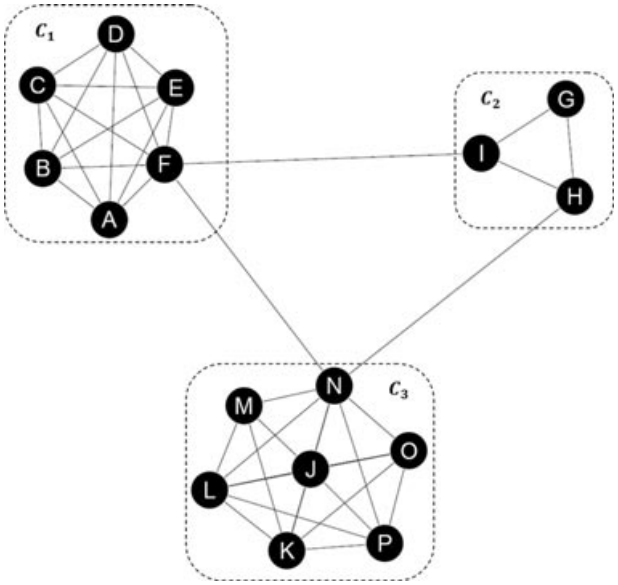


FIGURE 3 | Illustration of a network with three distinct communities: $C_1 = \{A, B, C, D, E, F\}$, $C_2 = \{G, H, I\}$, and $C_3 = \{J, K, L, M, N, O, P\}$.

former originally arose in the Computer Sciences field because of the necessity of finding the best way to allocate tasks to processors so as to minimize the communications between them. This network optimization task aimed at enhancing the computation, in a parallel computing environment. The latter was motivated by the discovery of community groups within society, to simplify the analysis of social phenomena through the arrangement of people according to their similarities. The main process behind community detection algorithms is based on dividing the original graph, into a set of disjoint subgraphs, through the optimization of a given objective function (e.g., modularity). The aim of both approaches is to discover groups of related vertices in the network and, if possible, the corresponding hierarchical organization, based only on the information provided by network topology. This is usually done by iteratively removing the *bridges* between groups of vertices, as suggested by Girvan and Newman.[42]

To better understand the introduced concepts, in Figure 3 is depicted a simple network comprising three communities, named C_1, C_2, and C_3. In this picture, we represent an ideal situation since each community is itself a complete graph, or a clique, of varying size ($C_1 = K_6$, $C_2 = K_3$, and $C_3 = K_7$). Also, the density of ties between communities is very low. The few ties that exist are bridges, since they are the only available connections between different parts of the network.

In real life, we can find several examples of such tight groups. There is a long list of examples, so we will only name a few. Society is a rich environment for finding communities, once people have the natural tendency to form groups. These groups can be families, circles of friends, working and/or religious groups, towns, nations, and so on. If we also consider groups formed by companies, or by customers of a given product, we can identify communities with relevance to Economics and Business fields. Biology is another activity where methods for finding communities are useful, especially within the scope of metabolic networks. For instance, in protein–protein interaction networks we can find groups of proteins with similar functions within the cell. We can also find virtual online communities in the network of Internet, or groups of topic-related web pages, which may be useful for the development of automatic and efficient recommendation systems.

The importance of studying these communities is intuitive in domains such as SNA. To highlight this importance, Fortunato[35] has stated that the analysis of the structural position of nodes, in each module, can help identify *central actors* (those within central positions), often associated to group control and stability functions, as well as *intermediate actors*, who are those who lie at the boundaries of communities and play a key role in the spread and exchange of new ideas and information, creating bridges between communities. Other interesting possibility opened by the task of discovering communities is the one that focus on the analysis of coarse-grained descriptions of the original graph. An example is the study of graphs obtained by considering vertices as communities and edges between them as an indicator of overlap between communities. This strategy is used by Oliveira and Gama[43] for the detection of transitions in clusters.

The following subsections are devoted to the introduction of the most popular (not necessarily the best) methods to solve the problem of finding communities. The great majority of these traditional algorithms assume partitions of vertices, instead of *covers*,[1] i.e., they do not allow overlap of communities, so each vertex is assigned to a single community. However, if one suspects that the nature of his/her network implies the existence of overlapping communities, a possible choice is the *clique percolation method*, proposed by the physicists Palla et al.[44] The main feature of this prominent approach is its ability to find overlapping communities in a network, by allowing vertices to belong to more than one group. This characteristic is especially appealing in social sciences, as people tend to belong to more than one

community (e.g., family, work, friends, etc.) at the same time.

For those interested in using clique percolation method to detect overlapping communities, Palla et al.[44] developed the CFinder software package, which is freely available at www.cfinder.org.

Hierarchical Clustering

Hierarchical clustering is a popular class of methods for finding clusters, since it does not require any assumptions regarding their number, membership, and size. Hierarchical clustering algorithms produce a flexible nested structure (smaller clusters within larger clusters which, in turn, are embedded in even larger clusters), typically represented by means of a dendrogram, that uncovers the multilevel structure of the network. Such features are highly desired in domains where little information is available concerning the community structure of a network. In addition, these methods proved to be quite effective in solving cluster analysis problems, thus becoming attractive for graph partitioning and community detection purposes.

The procedure of traditional hierarchical clustering is quite intuitive, being strongly based on the definition of similarity. Usually, the first step is the selection of the similarity measure that will be used to assess how alike two nodes are according to a given global, or local, property. Examples of such measures are the *cosine similarity*, the *Jaccard index*, the *Euclidean*, or *Manhattan* distances, the *Hamming* distance between pairs of rows in an adjacency matrix, among others. The next step is to compute the similarity matrix between all pairs of nodes, regardless of the fact that those nodes are, or not, connected to each other. Then, one chooses the approach to group them—the agglomerative or the divisive—and, depending on the choice, selects a given distance measure to compute the similarity between clusters (e.g., single linkage, complete linkage, Ward's method, etc.). The result is a dendrogram illustrating the arrangement of clusters returned by the hierarchical algorithm. To select the best partition, i.e., the best number of communities k, a typical strategy is to compute the value of *modularity*[41] for every possible number of clusters and select the number that maximizes this function.

As mentioned before, there are two general approaches for hierarchical clustering, which are as follows:

1. *Divisive methods*: this class of methods focuses on identifying and removing the spanning links between densely connected regions,[31] namely, bridges and local bridges. A well-known algorithm exploring this method is the one proposed by Girvan and Newman.[42]
2. *Agglomerative methods*: this class of methods focuses on the tightly knit parts of the network, rather on the connections at their boundaries. Walktrap[45] is an example of an algorithm based on this method.

In the next subsection, we present one of the best known and widely used divisive hierarchical algorithm for finding communities, especially in social networks: the algorithm of Girvan and Newman.

Girvan–Newman Algorithm

Among the most popular algorithms, or even the most popular one, for solving community detection problems is the one devised by Girvan and Newman[42] and known as the *Girvan–Newman algorithm*.

The Girvan–Newman algorithm is a divisive hierarchical technique that deconstructs the initial full network into progressively smaller connected pieces, until the point where there are no edges to remove and each node represents itself a community. Bearing in mind that communities are cohesive groups of nodes, with sparser connections between them, the criterion to remove the edges, proposed by Girvan and Newman,[42] is the graph-theoretic centrality measure *edge betweenness*. The reason behind this choice is related to the fact that this centrality measure is able to identify edges that lie on a large number of shortest paths between nodes and, therefore, are believed to connect different nonoverlapping communities. Thus, the main idea of this algorithm is that if we identify and remove bridges, we isolate the existing communities in a network.

Because it is based on the concept of betweenness, it is only suitable for networks of moderate order (up to a few thousand nodes), due to the high cost of computing it. The input of the algorithm is a full graph and the output is a hierarchical structure, such as a dendrogram, where communities at any level correspond to a horizontal cut through this hierarchical tree. The steps of the algorithm can be summarized as follows:

- Compute the betweenness of all edges in the network;
- Remove the edge with highest betweenness. This step may cause the network to split into separate disconnected parts, which constitute

the first level of regions in the partitioning of the graph.

- Repeat the previous steps until there are no edges to remove in the graph. Note that the obtained smaller components, within larger components, are the regions nested within the larger regions found in the first steps.

Because of its popularity, almost all standard software libraries have this algorithm implemented. For instance, in R[46] we can use the edge.betweenness.community function, provided by library igraph, to apply the *Girvan–Newman algorithm*.

Modularity Optimization

A widely used and very popular class of methods to detect communities in networks is *modularity maximization*. Modularity Q is a quality function that attempts to measure the merit of a given partition of the network into communities. It has been used not only to compare the quality of the partitions obtained by different community detection methods, but also as an objective function to optimize. According to Newman,[47]

> Modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random.

Based on this definition, we can deduce that modularity is a measure that explicitly takes into account the heterogeneity of the edges. The basic idea is that a network shows meaningful community structure if the number of edges between communities is fewer than expected on the basis of random choice. By assumption, the higher its value the better the partition, meaning that the found communities are internally densely connected and externally sparsely connected, because there are more edges falling within groups than what would be expected by chance. Modularity is computed as

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \qquad (21)$$

where m is the number of edges, k_i and k_j represent, respectively, the degree of nodes i and j, A_{ij} is the entry of the adjacency matrix that gives the number of edges between nodes i and j, $\frac{k_i k_j}{2m}$ represents the expected number of edges falling between those nodes, c_i and c_j denote the groups to which nodes i and j belong, and $\delta(c_i, c_j)$ represents the Kronecker delta.

BOX 1 SOFTWARE AND TOOLS FOR SOCIAL NETWORK ANALYSIS

To fulfill the growing need for social network mining and visualization, a considerable collection of software and tools were developed for SNA. Some of these tools are more technical and targeted to users with strong programming background and skills (e.g., UCInet[51]; `igraph`, `sna`, and `NetworkX`[52] libraries for `R`[46] environment), and others are more intuitive and, thus, more suitable for users from the social sciences (e.g., Gephi[53] and NodeXL[54]). Though these tools have different characteristics, most of them allow: the computation of metrics that provide a local (actor level) and global (network level) description of the network; the graphical visualization of the network; and the detection of communities. On the basis of the study of Combe et al.,[55] these constitute the expected functionalities of an SNA tool. In general, the main functionalities that can be embedded in SNA tools are the following:

- Creation of networks;
- Visualization and manipulation of networks;
- Qualitative and quantitative/statistical analysis of networks;
- Community detection;
- Predictive analysis (peer influence/contagion modeling, homophily models, and link prediction).

Despite the quantity and diversity of available tools in the Web, some of the most popular are

- Pajek[56]: freely available software for the analysis and visualization of large-scale networks.
- Gephi[53]: open source software for network manipulation and exploration, endowed with a three-dimensional render engine to display real-time evolving networks.
- UCInet[51]: commercial social network analysis which makes use of Pajek and NETDRAW for visualization and it especially suitable for statistical and matricial analyses.
- NodeXL[54]: freely available add-in to Microsoft Excel 2007 for the overview, discovery and exploration of networks, which does not require any programming skills because it is very user friendly. Not suitable for the analysis of large networks.
- `R`[46] libraries (`igraph`, `sna`, `tnet`, `statnet`, and `NetworkX`[52]): freely available packages for R environment which are very comprehensive (e.g., significant number of implemented algorithms for community detection; analysis of longitudinal networks, as well as two-mode networks) and with good two- and three-dimensional visualization capabilities.

From the formula, we can deduce that $Q \in [-1, 1]$, being either negative or positive. If positive, then there is possibility of finding community structure on the network. If Q is not only positive, but also large, then the corresponding partition may reflect the real community structure. According to Clauset et al.,[48] in practice, it was found that a modularity of about 0.3 is a good indicator of the existence of meaningful communities.

Following this reasoning, and knowing that the higher the modularity, the best the obtained network division is, a natural approach would be maximizing this measure, by computing it for every possible partition of the network and selecting the partition returning the higher value. This simple idea gave rise to a new class of methods whose foundations are set on the maximization of modularity. Albeit this approach is quite attractive, the exhaustive search over all possible divisions is usually intractable. This undesired effect of computational inefficiency has been circumvented by adapting a number of heuristic methods to this specific optimization problem. Following this strategy, one can obtain a fairly good approximation of the global optimum (in this case, the maximum value of modularity) in an acceptable time. Algorithms that employ this strategy are, for instance, the one proposed by Blondel et al.,[49] which performs a hierarchical optimization of modularity by exploring greedy techniques, and the one proposed by Guimerá and Amaral,[50] that applies the *simulated annealing* procedure to the modularity optimization problem.

Those interested in knowing more about the problem of finding communities in networks, can refer to the recently released survey by Fortunato.[35]

CONCLUSIONS AND CURRENT TRENDS

At the beginning, analysis of social networks was based in single small graphs. The first studies used data collected using direct questionnaires asking respondents to detail their interactions with others. The gathered data was then represented using the mathematical graph model, where vertices correspond to individuals (the respondents) and edges to interactions between them. These traditional studies usually entailed some problems such as inaccuracy, subjectivity, and lack of generality due to small sample size.

The substantial advance of technology, the increasing availability of computers and the arising of communication networks contributed to the emergence of new movements in network research. These new approaches started to focus in the analysis of the statistical properties of large-scale complex networks, which were easily gathered using computers and other electronic devices. This change of scales forced upon a corresponding change in the traditional analytical approach and data mining statistical methods became of great importance, due to their ability to extract patterns and knowledge from massive quantities of data.[9] Nowadays, the wide availability of software and SNA tools and libraries (more than 50) is a reflection of this evolution in SNA (Box 1).

Because of these technological advances and consequent impact in the availability of networked data, new challenges are being posed to the research field of SNA, and a new paradigm is emerging. This paradigm takes into consideration new factors in the analysis of social networks, such as the size of data, which is incredibly getting large, and changes in space and time.

The first issue has implications on existing methods for SNA, which now need to be improved to scale well to large-scale social networks. For instance, finding communities in large-scale social networks will continue to be a dynamic research challenge. Also in the area of community detection, the urge to develop new methods which are not only scalable and efficient but also fully automatic, in the sense that they do not need any user-specified parameter (e.g., number and size of communities), will continue to be an important research problem.

The second issue is of extreme relevance because the speed at which data is collected is turning obsolete static analyses. Therefore, the study of the dynamics and evolution of social networks, which include but are not restricted to both the discovery of the general properties that govern the temporal evolution of social networks and the detection and understanding of temporal and spatial changes in these networks, will continue to be a significant strand of research in SNA.

It is also expected that the popularity of SNA will continue to increase, attracting more and more researchers to the field and impelling an increasing number of companies to incorporate SNA methods into their business processes and generalize their use as strategic tools.

NOTES

[a] *Cover* is a synonym of *partition* for soft assignment of vertices to communities.

ACKNOWLEDGMENTS

We are grateful to the referees for their helpful comments and suggestions on an earlier draft of this paper. We are also grateful for the financial support of the project *Knowledge Discovery from Ubiquitous Data Streams* (PTDC /EIA-EIA/098355/2008). The work of Márcia Oliveira was also supported by the Portuguese Foundation for Science and Technology (FCT), under the PhD grant SFRH/BD/81339/2011.

REFERENCES

1. Moreno JL. *Who Shall Survive?* New York: Beacon House; 1953.
2. Domingos P, Richardson M. Mining the network value of customers. In: *Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM 2001, 57–66.
3. Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM 2002, 61–70.
4. Leskovec J, Adamic LA, Huberman BA. The dynamics of viral marketing. *ACM Trans Web* 2007, 1:228–237.
5. Dasgupta K, Singh R, Viswanathan B, Chakraborty D, Mukherjea S, Nanavati AA, Joshi A. Social ties and their relevance to churn in mobile telecom networks. In: *Eleventh International Conference on Extending Database Technology: Advances in Database Technology*. New York, NY: ACM 2008, 668–677.
6. Wei C-P, Chiu I-T. Turning telecommunications call details to churn prediction: a data mining approach. *Expert Syst Appl* 2002, 23:103–112.
7. Xu J, Che H. Criminal network analysis and visualization. *Commun ACM* 2005, 48:101–107.
8. Shetty J, Adibi J. The Enron Email Dataset Database Schema and Brief Statistical Report. Technical Report, University of Southern California, 2004.
9. Newman MEJ. The structure and function of complex networks. *SIAM Rev* 2003, 45:167–228.
10. Van De Bunt GG, Van Duijn MAJ, Snijders TAB. Friendship Networks Through Time: An Actor-Oriented Dynamic Statistical Network Model. *Comput Math Org Theory* 1999, 5:167–192.
11. Ritter T. The networking company: antecedents for coping with relationships and networks effectively. *Ind Mark Manage* 1999, 28:467–479.
12. Newman MEJ. The structure of scientific collaboration networks. *Proc Natl Acad Sci USA* 2001, 98:404–409.
13. Truyen TT, Phung DQ, Venkatesh S. Preference networks: probabilistic models for recommendation systems. In: *Sixth Australasian Conference on Data Mining and Analytics*. Darlinghurst, Australia: Australian Computer Society, Inc. 2007. 70:195–202.
14. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J. Graph structure in the Web. *Comput Netw* 2000, 33:309–320.
15. Alon U. Biological networks: the tinkerer as an engineer. *Science* 2003, 301:1866–1867
16. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press; 1994.
17. Diestel R. *Graph Theory*. 3rd ed. Heidelberg: Spring-Verlag; 2005.
18. Granovetter M. The strength of weak ties. *Am J Sociol* 1973, 78:1360–1380.
19. Granovetter M. *Getting a Job: A Study of Contacts and Careers*. Cambridge, MA: Harvard University Press; 1974.
20. Freeman LC. Centrality in social networks: conceptual clarification. *Soc Netw* 1979, 1:215–239.
21. Brin S, Page L. Node centrality in weighted networks: generalizing degree and shortest paths. *Soc Netw* 2010, 32:245–251.
22. Bonacich P. Power and centrality: a family of measures. *Am J Sociol* 1987, 92:1170–1182.
23. Barabási AL, Bonabeau E. Scale-Free Networks. *Sci Am* 2003, 288:60–69.
24. Barabási A-L, Albert R. Emergence of scaling in random networks. *Science* 1999, 286:509–512.
25. Kossinets G, Watts DJ. Empirical analysis of an evolving social network. *Science* 2006, 311:88–90.
26. Watts DJ, Strogatz SH. Collective dynamics of small-world networks. *Nature* 1998, 393:440–442.
27. Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: densification laws, shrinking diameters and possible explanations. In: *Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. New York, NY: ACM 2005, 177–187.
28. Costa L, Oliveira O, Travieso G, Rodrigues F, Villas Boas P, Antiqueira L, Viana M, Rocha L. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Adv Phys* 2011, 60:329–412.
29. Kleinberg J. Authorative sources in a hyperlinked environment. *J ACM* 1999, 46:604–632.

30. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. *Comput Netw ISDN Syst* 1998, 30:107–117.
31. Easley D, Kleinberg J. *Networks, Crowds and Markets: Reasoning about a Highly Connected World*. Cambridge, UK: Cambridge University Press; 2010.
32. Thelwall M. Interpreting social science link analysis research: a theoretical framework. *J Am Soc Inf Sci Technol* 2006, 57:60–68.
33. Rapoport A. Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *Bull Math Biophys* 1953, 15:523–533.
34. Erdos P, Renyi A. On the evolution of random graphs. *Publ Math Inst Hungarian Acad Sci* 1960, 5:17–61.
35. Fortunato S. Community detection in graphs. *Phys Rep* 2010, 486:75–174.
36. Milgram S. The small world problem. *Psychol Today* 1967, 1:61–67.
37. Price DDS. Networks of scientific papers. *Science* 1965, 149:510–515.
38. Price DDS. A general theory of bibliometric and other cumulative advantage processes. *J Am Soc Inf Sci* 1976, 27:292–306.
39. Newman MEJ. Mixing patterns in networks. *Phys Rev E* 2003, 67:026126.
40. Gupta S, Anderson RM, May RM. Networks of sexual contacts: implications for the pattern of spread of HIV. *AIDS* 1989, 3:807–817.
41. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E* 2004, 69:026113.
42. Girvan M, Newman MEJ. Community structure in social and biological networks. *Proc Natl Acad Sci USA* 2002, 99:7821–7826.
43. Oliveira M, Gama J. MEC—monitoring clusters' transitions. In: *Proceedings of the 5th Starting AI Researchers' Symposium*. Lisbon, Portugal: IOS Press; 2010.
44. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 2005, 435:814–818.
45. Pons P, Latapy M. Computing communities in large networks using random walks. *J Graph Algorithms Appl* 2006, 10:191–218.
46. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2011. Available at: http://www.R-project.org. (Accessed January 14, 2012)
47. Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci USA* 2006, 103:8577–8582.
48. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. *Phys Rev E* 2004, 70:066111.
49. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp* 2008, 2008:P10008.
50. Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. *Nature* 2005, 433:895–900.
51. Borgatti SP, Everett MG, Freeman LC. Ucinet for Windows: software for social network analysis. *Harv Anal Technol* 2002, 2006.
52. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: *Seventh Python in Science Conference*; 2008, 11–15.
53. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: *Third International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA: Association for the Advancement of Artiïcial intelligence. 2009, 361–362.
54. Smith M, Shneiderman B, Milic-Frayling N, Rodrigues EM, Barash V, Dunne C, Capone T, Perer A, Gleave E. Analyzing (social media) networks with NodeXL. In: *Fourth International Conference on Communities and Technologies*. New York, NY: ACM 2009, 255–264.
55. Combe D, Largeron C, Egyed-Zsigmond E, Géry M. A comparative study of social network analysis tools. *Soc Netw* 2010, 2:1–12.
56. Batagelj V, Mrvar A. Pajek—program for large network analysis. *Connections* 1998, 21:47–57.

FURTHER READING/RESOURCES

Doreian P, Stockman FN, eds. *Evolution of Social Networks*. London: Routledge; 1997.

Degenne A, Forsé M. *Introducing Social Networks*. London/Thousand Oaks, CA/New Delhi: Sage Publications; 1999.

Freeman LC. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver, Canada: Empirical Press; 2004.

Carrington PJ, Scott J, Wasserman S, eds. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press; 2005.

Knoke D, Yang S. *Social Network Analysis*. 2nd ed. London/Thousand Oaks, CA/New Delhi: Sage Publications; 2008.

