4 Networks and Connectivity: Metrics and Models

Luís Cavique

Abstract

This work explores network science to understand and visualize the intricate interconnectivity within organizations. The age of big data emphasizes the importance of deriving new insights by transforming data into networks to study their connections. The document introduces a three-step maturity framework for navigating network science, starting with the basics of network construction, moving on to standard metrics, and culminating in an examination of network topology and dynamics. The author aims to clarify the subject and encourage further exploration, suggesting that while network science may not have all the answers, it offers a critical analytical framework.

Keywords

Networks · Networks metrics · Networks models · Social networks analysis

4.1 Introduction

The skeleton of many structures can be represented by a network, which constitutes a set of objects connected in a specific way. Routers on the Internet, neurons in the brain, and aviation routes are examples of networks. Currently, the branch of mathematics known as Graph Theory, initially proposed by mathematician Leonhard Euler, constitutes the basis of mathematical modeling for the study of network science.

L. Cavique (✉)
Universidade Aberta, Lisbon, Portugal
e-mail: lcavique@uab.pt

O. Gomes et al. (eds.), *The Agent ID Model*, Integrated Science 32,
https://doi.org/10.1007/978-3-031-75996-3_4

Graph theory has been widely used in social network analysis due to its representational capacity and simplicity. A graph comprises vertices or nodes and links or edges that connect the nodes. In social networks, graph representation is also called sociogram, in which the nodes are the actors, and the connecting lines establish the set of relationships in a two-dimensional way.

A social network can be expressed mathematically by a graph comprising a set of nodes or vertices connected by edges that express a relationship between them. The social network concept can be described as a social structure composed of nodes related to other nodes using edges. The nodes represent people or organizations, and the edges represent their relationships.

A social network is a set of people or groups linked together by relationships that can be professional, family, or others. Studies in this field began in the thirties by Moreno and Jennings (1934) with the graphic representation of social networks, the sociograms. The introduction of mathematical models was initiated by Rapoport (1957) and continued by Erdös and Rényi (1959) with the creation of random graphs.

Another relevant study in sociology was developed by Milgram (1967), which led to the concept of a 'small world' and the notion of 'six degrees of separation'. The experience consisted of sending letters between two different points, from Nebraska, in the Midwest, to the east coast of the USA, in Massachusetts.

People were asked not to use the postal services but to use their contacts to send and resend the letters. The letters that reached their destination needed to pass, on average, through about six intermediary people, which led Milgram to conclude that, in this case, the Americans are approximately six steps away from each other.

The concept of social networks gained a new dynamic with the emergence of Web 2.0. In 2008, Facebook surpassed MySpace worldwide in total number of users, becoming the most extensive online social network.

Ellison et al. (2007) conducted a broad study indicating how Facebook allows actors to invest in new social ties. These characteristics show us that there is a change in the support of interaction, allowing conversations previously mapped only by observation.

Given the resources provided by Web 2.0, such as social networking sites, actors begin to record their conversations, their interactions, and their networks. With this, the mapping of these networks gains new potential in the sense that, for the first time, it is possible to map the tastes, acts, ideas, and connections of thousands of people, search for and establish patterns between these multiple networks, mainly through the interactions that are mediated by these tools.

A third school emerged after the mathematics and sociology approach to networks merged with new contributions from computer science, biology, physics, and economy: network science. The pioneering work by Watts and Strogatz (1998), published in the Nature journal, studies graphs with small-world properties and free-scale degree distribution. It follows the model of Barabási and Albert (1999) and the SIR epidemiological models that define new frontiers for networks in dynamic systems.

Problem The contributions in network science are numerous, diverse (articles, textbooks, and science popularization books), and generally with excellent quality. Therefore, tackling this subject presents a significant challenge.

A pioneering textbook (Barabási & Pósfai 2016) introduces network science to an interdisciplinary audience and includes real-world examples. Newman (2018) textbook integrates breakthroughs in network studies from multiple disciplines, covering network analysis, graph theory, mathematical models, and dynamical processes on networks. Scott (2000) provides an introduction and comprehensive guide to social network analysis, catering to beginners and more seasoned social network researchers in the social network. Easley and Kleinberg (2010) explore the new science of networks through an interdisciplinary lens, aiming to understand various modern societal phenomena. Additionally, the book examines how epidemics and financial crises can escalate quickly and with great intensity, reflecting the interconnected nature of contemporary society.

Regarding the science popularization books, some examples are highlighted. Barabási and Frangos (2014), the expert in the new science of networks, takes us on an intellectual adventure to prove that social networks, corporations, and living organisms are more similar than previously thought. Watts (2004) explores phenomena such as disease epidemics, financial market fluctuations, information search behaviors, organizational resilience, and the dynamics of personal relationships. Christakis and Fowler (2010) study the influence of third-degree connections within our social networks, revealing their remarkable ability to mold various aspects of life. Our actions can potentially alter the behaviors, thoughts, and even the health of individuals we have never encountered directly.

Objective This work aims to create an overview guided by a framework that allows insights into network science by mixing areas such as mathematics, sociology, computer science, biology, physics, and economics.

Contribution The contribution of this work is a guided tour of network science with the three-step maturity framework with increasing complexity. The pipeline of Fig. 4.1 first details what networks are and how to create them, then lists the most usual metric for networks, and finally introduces two models: the topology of networks and the network dynamics.

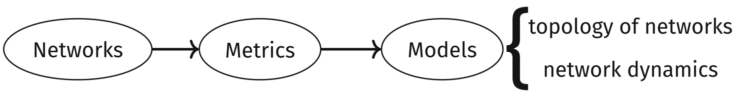

Fig. 4.1 Pipeline of this work. Source: adapted from https://gatton.uky.edu/about-us/stay-connected/news/2020/links-center-social-network-analysis-workshop-success

Organization The remainder of this chapter is organized as follows. Section 4.2 defines graphs and networks. Section 4.3 presents the most usual network metrics. Sections 4.4 and 4.4.2 detail the models: topologic models and dynamic models. Finally, in Sect. 4.5, we draw some conclusions.

4.2 Graphs and Networks

This section states some crucial definitions of networks and exemplifies the transformation of generic data into networks.

4.2.1 Definitions

A graph G is defined by the ordered pair (N, E), where N is a set of nodes / vertices and E is a set of edges. Each edge e belonging to the set E is denoted by $e = (u, v)$, a pair of nodes. The nodes u and v are the ends of the edge and are called adjacent nodes or neighbors. Berge (1970) is one of the foundational references in graph theory.

In this work, it is relevant to differentiate various dichotomies in graphs: (undirected and directed), (unweighted and weighted), and obviously (graphs and networks). An undirected graph features edges that connect vertices bidirectionally without orientation, allowing travel between any connected vertices in both directions. On the other hand, a directed graph, or digraph, consists of edges with a specific orientation, represented as ordered pairs (u, v), permitting travel only from u to v unless an opposite directed edge exists. Unweighted graphs have edges without associated values or weights, representing connections between vertices. Conversely, weighted graphs assign weights to edges, often denoting distances, costs, or other metrics, influencing traversal relations and analysis by considering the significance of connections between vertices.

A graph is a formal mathematical concept studied in graph theory. On the other hand, a network uses the structure of a graph associated with real-world systems with entities connected in some way, such as computer networks, social networks, transportation networks, or electrical networks.

4.2.2 From Data to Networks

Today, in the intricate big data ecosystem, we are inundated with overwhelming data from various sources, including barcode scanners, Web 2.0 platforms, and IoT devices. Now, the question is not how to collect data but how we can effectively utilize this vast data. This section shows a method for converting this data into structured networks (Scott 2000).

Table 4.1 Raw data and the incidence matrix product-customer

trans_id	Customer	Product
1	a	Bread
2	a	Butter
3	a	Milk
4	b	Oranges
5	b	Butter
6	b	Apples
7	c	Bread
8	c	Butter
9	c	Milk
10	c	Oranges

		Customer		
		a	b	c
Products	Apples		1	
	Bread	1		1
	Butter	1	1	1
	Milk	1		1
	Oranges		1	1

Table 4.2 Matrices of customer-customer and product-product

	a	b	c
a		1	3
b			2
c			

Matrix customer-customer

	Apples	Bread	Butter	Milk	Oranges
Apples			1		1
Bread			2	2	1
Butter				2	2
Milk					1
Oranges					

Matrix product-product

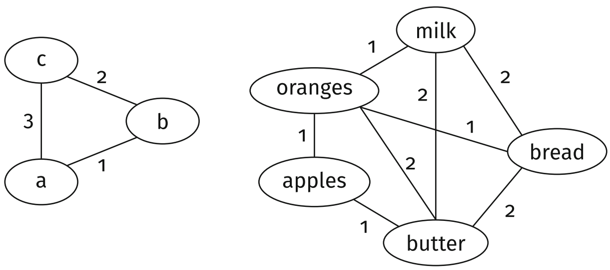

Fig. 4.2 Networks of customers and products. Source: prepared by the author

The following running example uses supermarket data. The method begins by selecting two specific columns: customer and product. Next, construct a product-customer incidence 5×3 matrix, as illustrated in Table 4.1.

With the information on the incidence matrix product-customer, two matrices can be generated: the customer-customer matrix and the product-product matrix, as shown in Table 4.2. In the matrix customer-customer, the pair (a, c) equals three, indicating that both customers purchased bread, butter, and milk. The same reasoning is applied to all elements of the matrices.]

Given the matrices of customers and products, the undirected networks can be quickly drawn, as shown in Fig. 4.2.

A set of observations can be drawn from the networks. We can detect and infer the following from a simple visual analysis of the networks. Analyzing the customers' network shows that customers a and c exhibit more significant similarities. On the contrary, a and b show fewer similarities.

In order to analyze the network of products, the clique concept must be defined. In graph theory, a clique is a subset of vertices within an undirected graph such that every two distinct vertices in the clique are adjacent; this means that every vertex in the clique is connected to every other vertex in the clique by an edge. The maximum clique in a graph refers to the largest clique within that graph. Two maximal cliques with three nodes with a weight equal to six are found, meaning that the most frequently bought together products are (milk, bread, butter) and (oranges, bread, and butter). The maximum weighted clique uses four nodes and weighs 10, meaning that the four most frequently bought together products are in the basket (milk, bread, butter, oranges).

4.3 Network Metrics

In this section, some of the most relevant network metrics are described. Centrality metrics are crucial for pinpointing influential nodes, while the degree of distribution provides insight into the network's overall connectivity framework. The clustering coefficients metric reveals local neighborhoods. Meanwhile, similarity, assortativity, and homophily explore the propensity for nodes to associate and form connections based on shared attributes.

4.3.1 Centrality Metrics

The measures of centrality arise in the context of network science. Intuitively, the most central vertices in a network are those from which we can reach any other more easily or quickly. The centrality measures identify an individual's position relative to others in their network. Thus, the more central the individual, the greater their influence and power in their network. This sub-section is based on the works of Landherr et al. (2010), Laranjeira and Cavique (2014), and Newman (2018).

4.3.1.1 Degree Centrality

The most straightforward and intuitive concept regarding the centrality of a vertex is the number of direct contacts it has. A person in a position that allows direct contact with many others is seen by others as a significant channel of information, which is why we say they are more central. Thus, degree centrality is the count of the number of adjacencies of a vertex.

Let G be any graph with N vertices, and let x be a vertex of G. The degree centrality of x, denoted by $\sigma_D(x)$, is the number of edges incident to x. Given the adjacency matrix of the graph $A(G)$ and their elements a_{ix}, the degree centrality is given by:

$$\sigma_D(x) = \sum_{i=1}^{N} a_{i,x} \tag{4.1}$$

4.3.1.2 Closeness Centrality

Another measure, known as closeness centrality, is based on the sum of the distances from a vertex to all other vertices in the graph (Sabidussi 1966). In various contexts, more important than having many connections is not being too far from the other elements.

Let G be a connected graph with N vertices, and let x be a vertex of G. The closeness centrality of x is given by the inverse of the sum of the distances from x to all other vertices in the graph, where $d_G(x, i)$ represents the distance from x to i:

$$\sigma_G(x) = \frac{1}{\sum_{i=1}^{N} d_G(x, i)} \tag{4.2}$$

4.3.1.3 Betweenness Centrality

Betweenness Centrality allows us to measure the capacity of an actor to influence their peers in a network. The idea behind this metric can be easily understood through the following analogy: a city integrated into various trade routes will undoubtedly have a strategic advantage, which motivates the idea that a critical node is part of many paths.

Let G be a graph (connected or not) with N vertices, and let x be a vertex of G. Given that g_{ij} represents the number of shortest paths from vertex i to vertex j, and $g_{ij}(x)$ indicates the number of those shortest paths that pass-through x, the betweenness centrality of x is given by:

$$\sigma_B(x) = \sum_{i=1}^{N} \sum_{i<j}^{N} \frac{g_{i,j}(x)}{g_{i,j}}, \quad i, j \neq x \tag{4.3}$$

4.3.1.4 Eigenvector Centrality

This measure is based on the concepts of eigenvalue and eigenvector of the adjacency matrix of the graph G. It suggests that another way to interpret centrality is to think that the centrality of a vertex is a function of its neighboring vertices, that is, an important node has important neighbors (Bonacich and Lloyd 2001).

Consider that the eigenvector centrality, $\sigma_E(x)$, is proportional (by a factor of λ^{-1}) to the sum of the centralities of its neighbors, $\lambda_{max}(A)$ is the largest eigenvalue (in magnitude) of the adjacency matrix A, and $v = (v_1, \ldots, v_n)^T$ is the eigenvector associated with it. The metric is given as follows:

$$\sigma_E(x) = \frac{1}{\lambda_{max}(A)} \sum_{j=1}^{N} a_{jx} \cdot v_j \tag{4.4}$$

4.3.1.5 Katz Centrality

The Katz centrality is a measure of centrality in a network that generalizes the concept of eigenvector centrality. Developed by Katz (1953), it is designed to measure the influence of a node within a network by considering not only the immediate neighbors of the node but also all other nodes in the network, connecting through various paths of different lengths.

Unlike measures that only consider direct connections (such as degree centrality) or the shortest paths (like betweenness centrality), Katz centrality considers the total number of walks between two nodes exponentially attenuated by the length of each walk. This way, shorter paths contribute more to the centrality score than longer ones, but longer paths are not ignored.

4.3.1.6 PageRank Centrality

PageRank centrality, often called PageRank, is a measure of the importance or centrality of a node within a network. It was initially developed by Page et al. (1998), the founders of Google, as part of the algorithm used by the Google search engine to rank web pages in their search results. In this centrality metric, a direct network is taken into account.

PageRank's underlying concept is that a page's importance is determined by the number and quality of links, assuming that more relevant pages will likely receive more links from other pages. The PageRank of a node (or web page) is defined recursively and depends on the number and PageRank scores of all other pages that link to it (inbound links). A page linked to many pages with high PageRank receives a high rank.

4.3.1.7 Limitations of Centrality Metrics

Centrality measures in networks each capture different aspects of a node's importance, leading to varied results depending on the specific measure applied, exemplified in Fig. 4.3. For instance, degree centrality focuses on immediate connections. In contrast, betweenness centrality emphasizes a node's role as a bridge in the shortest paths between others.

Therefore, no single centrality measure provides a comprehensive solution across all network types and contexts, making it crucial to select the most appropriate measure based on the specific characteristics and requirements of the analyzed network. In the following sub-sections, new metrics are explored, and in the following sections, concepts like the topology and dynamics of the network are developed.

4.3.2 Degree Distribution

In network science, the degree of a node is a fundamental metric that quantifies the number of connections or edges a node has with other nodes in the network. It measures how connected a node is within the network's structure (Newman 2018).

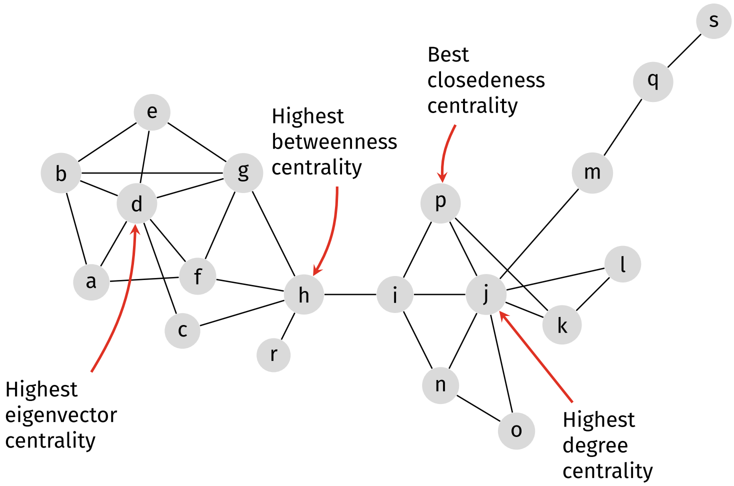

Fig. 4.3 Results of different centrality metrics. Source: https://ultrabpm.wordpress.com/2013/03/25/social-network-analysis-part-two/

In undirected networks, the degree of a node is simply the count of direct connections it has. In this case, the edges do not have a direction (or can be considered bi-directional), and the degree indicates the number of edges connected to the node. In directed networks, the degree concept is split into two distinct types. In-degree is the number of incoming edges to a node. It represents how many other nodes directly connect to the given node. On the other hand, the out-degree is the number of outgoing edges from a node. It shows how many nodes the given node is directing connections.

The degree distribution of a network is a fundamental concept in network science that characterizes the connectivity of nodes within a network. It provides insight into the structure and dynamics of complex networks, which are systems of interconnected elements or nodes.

Figure 4.4 shows an undirected network and its degree distribution. Some nodes have a degree equal to zero, most have a degree equal to two, and a few have a degree equal to seven.

Degree distribution shapes network resilience, with scale-free networks vulnerable to targeted attacks but resistant to random failures. It also affects the speed of information or disease spread, with network hubs accelerating this process. Additionally, the degree to which distribution influences network dynamics and community structure impacts network growth, flow efficiency, and community formation and communication. This understanding is vital for analyzing and optimizing networks across diverse fields.

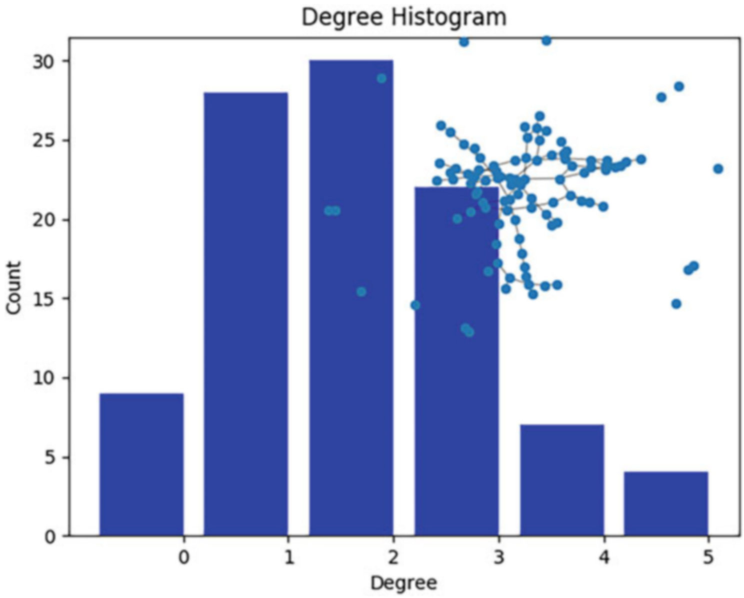

Fig. 4.4 Degree distribution of a network (Source: https://networkx.org/documentation/networkx-2.3/auto_examples/drawing/plot_degree_histogram.html)

4.3.3 Clustering Coefficient

The clustering coefficient is a measure that quantifies the degree to which nodes in a network tend to cluster together. It provides an insight into the extent to which nodes in a network tend to form tightly knit groups characterized by a relatively high density of ties (Newman 2018). The clustering coefficient is based on triplets of nodes, where a triplet is three nodes connected by three edges. The clustering coefficient can be defined as the ratio between the number of closed triplets and the number of all possible triplets (open and closed). In other words, for an individual node, the local clustering coefficient C_i is calculated as follows:

$$C_i = \frac{(\text{\# pairs of neighbors of } i \text{ that are connected})}{(\text{\#pairs of neighbors of } i)} \tag{4.5}$$

The clustering coefficient can be defined for individual nodes and the entire network. For an individual node, the local clustering coefficient is calculated as follows:

$$\overline{C} = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{4.6}$$

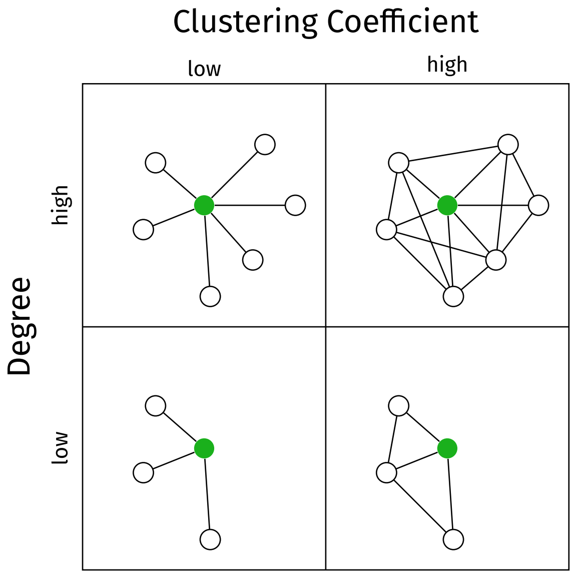

Fig. 4.5 Clustering coefficient versus degree (Adapted from https://www.researchgate.net/figure/Representation-of-the-clustering-coefficient-and-degree-Networks-have-nodes-with_fig5_261995698)

Figure 4.5 shows the comparison of two metrics for the same node. A low-degree (equal to 3) and high-degree (equal to 6) are shown regarding the degree metric. Concerning the clustering coefficient, the low values are equal to zero, and the high coefficient shows an example with a value equal to one.

4.3.4 Similarity, Assortativity, and Homophily

Similarity refers to a measure of how alike two nodes are within a network based on specific attributes or patterns of connections. Similarity can be measured in different ways, taking into account various attributes like demographic characteristics in social networks or functional properties in biological networks, as well as the overall structure of the network (Newman 2018). For example, in a social network, two individuals might be considered similar if they share common interests, work in the same field, or have a similar social circle.

The notion of similarity between nodes encompasses a variety of metrics, each tailored to capture distinct aspects of node characteristics or their interconnectedness. Cosine similarity evaluates the alignment of vectors representing node attributes or connections, offering insight into their mutual orientation in a multidimensional space. Jaccard similarity, on the other hand, quantifies the overlap between the neighbor sets of two nodes, emphasizing their shared connections relative to their combined network footprints.

Pearson correlation delves into the linear relationship between node attributes, revealing the degree to which they co-vary. While Euclidean distance typically conveys dissimilarity through spatial separation, in some contexts, it inversely relates to similarity, with closer nodes being more alike. The structural similarity extends beyond individual attributes or pairs of nodes, incorporating broader network patterns to assess how nodes fit within the overall network architecture. Each metric offers a distinct way to understand a network's complex web of connections, helping to focus the analysis on particular qualities or patterns that matter.

Note that the metrics of similarity are related to distance metrics. However, similarity and distance are inversely related: as the distance between two points decreases, their similarity increases, and vice versa.

Assortativity is a property of networks that measures the tendency of nodes to connect with other nodes that are similar to them. Assortativity can be in terms of node degree (number of connections), called degree assortativity, or in terms of some nodal attribute, known as attribute assortativity. In assortative networks, high-degree nodes are more likely to be connected to other high-degree nodes.

Similarly, nodes with similar attributes tend to be connected more often than by chance. This concept is fundamental in understanding how networks form and evolve and how information or diseases might spread.

On the other hand, negative assortativity, or disassortative mixing, characterizes a network pattern where nodes preferentially connect to others that are unlike themselves, either in node degree or other attributes. In such networks, nodes with many connections often link to nodes with few connections and vice versa. This pattern contrasts with positive assortativity, where similar nodes tend to connect.

Homophily is closely related to assortativity and refers to the principle that contact between similar people occurs more commonly than among dissimilar people. Homophily can be based on various attributes, including but not limited to age, gender, ethnicity, education, and social status. It is a driving force behind the formation of social networks. It influences how networks are structured, affecting the flow of information, behaviors, and even the spread of diseases within the network.

4.4 Topology of Networks

This section discusses the Erdös-Rényi model of random graphs, where graphs are generated by connecting nodes randomly. Then, it introduces the Barabási-Albert model, a method for generating scale-free networks through preferential attachment. It highlights its applicability in representing real-world social, biological, and technological networks. After that, describe the Watts-Strogatz model, which bridges the gap between highly ordered and completely random networks, explaining the small-world phenomenon observed in real-world networks. It explores the concepts of cliques and their relaxations in network topology, focusing on identifying cohesive sub-groups within more extensive networks.

Additionally, it addresses the role of brokers in bridging structural holes between communities facilitating the exchange of information and innovation. Finally, it highlights the importance of visualization in network topology, presenting it as a crucial tool for understanding the complex relationships within networks.

4.4.1 Erdös-Rényi Model

Random graphs are types of graphs that are generated using a random process. Given N nodes and E edges, the algorithm to generate a random graph $G(N, E)$ uses a random seed number and links a random pair of nodes.

The link density p can quantify the overall connectivity, the number of links divided by the maximum number of edges $E_{max} = N(N-1)/2$. The density $p = E/E_{max}$, in other words, $p = 2E/[N(N-1)]$. The degree k of a node is the number of its edges to neighbors. The mean degree is related to the density by $\overline{k} = p(N-1) = 2E/N$.

The Erd"os-Rényi model is a specific type of random graph model. The Erdös-Rényi model, named after Paul Erdös and Alfréd Rényi, is one of the earliest and simplest models for generating random graphs. It comes in two versions, denoted $G(N, E)$ and $G(N, p)$. In the $G(N, E)$ model, a graph is constructed by choosing exactly E unique edges from the possible combinations $(N, 2)$. The resulting graph has N vertices and E edges. The $G(N, p)$ model is the more commonly used version of the Erdös-Rényi model. It starts with N isolated vertices, and then each possible pair of vertices is connected with an edge with probability (or density) p, independently from every other pair. The parameter p controls the density of the graph, where a higher p results in a graph with more edges.

The properties of graphs generated by the Erdös-Rényi model have been extensively studied. For example, one of the well-known results is about the phase transition related to the emergence of a giant connected component:

- When the average degree $\overline{k} < 1$, the fraction of nodes in its largest connected component is close to 0;
- When the average degree $\overline{k} > 1$, the fraction of nodes in its largest connected component grows as the average degree $\overline{k}$ grows; that is, the probability of a 'giant component' goes to 1.

Erdös, one of the most prolific mathematicians, wrote over 1500 papers with more than 500 co-authors. Erdos represents the number zero in his social network, and the researchers who worked with him are called Erdos's number 1. Erdos' number of 1 co-authors is called Erdos's number of 2, and so on, building one of the oldest small world networks known. The work of Erdös and Rényi (1959) presents interesting random graph properties. A relevant example of the "Erdos Number" can be found in Grossman et al. (2024).

4.4.2 Barabási-Albert Model

The Barabási and Albert (1999) model generates random scale-free networks using a preferential attachment mechanism. Albert-László Barabási and Réka Albert introduced it in their seminal work on scale-free networks. The model marked a significant shift in how scientists understand the structure and evolution of complex networks, such as the Internet, social networks, and biological networks.

The Barabási-Albert model starts with a small number of nodes (initial network) and grows the network by adding one node at a time. Each new node creates links to existing nodes with a probability proportional to the number of links the existing nodes already have. This 'rich-get-richer' mechanism ensures that already well-connected nodes are more likely to receive new links. The critical steps in the model are:

- Growth: The network starts with an initial small number of interconnected nodes. At every step, a new node is added to the network.
- Preferential Attachment: Each new node forms a fixed number of links to existing nodes. The probability that a new node will connect to an existing node is proportional to the degree (number of connections) of that existing node.

The term 'scale-free' comes from the characteristic degree distribution of these networks, which follows a power law, at least asymptotically. In a scale-free network, the probability $P(k)$ that a randomly selected node has k connections (or degree k) falls off as k raises to a negative power, typically $P(k) \sim k^{-\gamma}$ where γ is a parameter usually in the range $2 < \gamma < 4$. A scale-free network has a few highly connected nodes (hubs) and many nodes with few connections.

Unlike in Erdös-Rényi random graphs, where the degree distribution is binomial or Poisson in the limit of large network sizes, leading to a characteristic average degree, scale-free networks do not have a characteristic scale for the degree. This property has profound implications for the network's robustness, its vulnerability to targeted attacks, and the dynamics of processes, such as spreading phenomena on the network. Figure 4.6 shows the degree distributions and the topology of both models.

Regarding real-world applicability in social, biological, or technological networks, the Barabási-Albert model is more realistic than the previous one. Newman et al. (2006) exemplify the exponents of real-world networks of cinema actors, the World Wide Web, and electric power, $\gamma_{actor} = 2.3$, $\gamma_{www} = 2.1$, and $\gamma_{power} = 4$.

4.4.3 Watts-Strogatz Model

The Watts and Strogatz (1998) model, proposed by Duncan J. Watts and Steven Strogatz in their paper, is a mathematical model designed to explain the small-world phenomenon in networks. This model represents a bridge between the highly

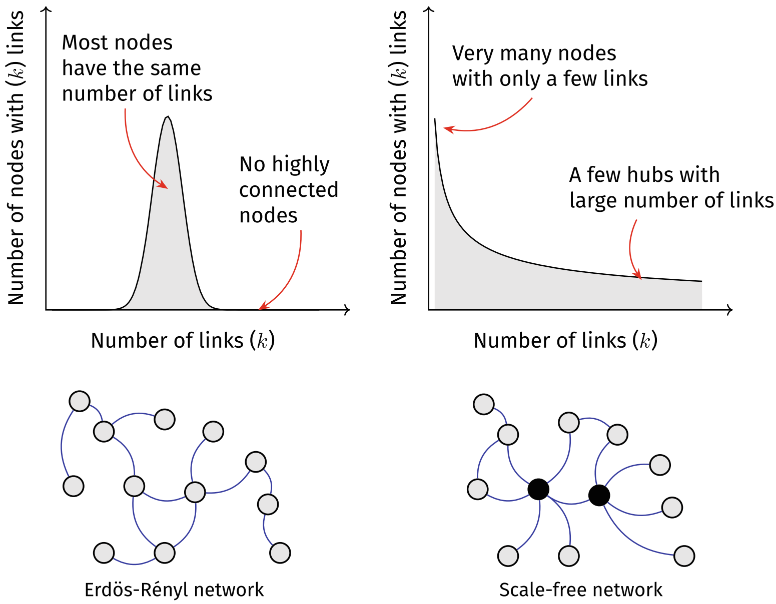

Fig. 4.6 Degree distribution and topology of Erdös-Rényi and scale-free networks (Adapted from https://edgeperspectives.typepad.com/edge_perspectives/2007/05/the_power_of_po.Html and https://pt.wikipedia.org/wiki/Ficheiro:Scale-free_network_sample.png)

ordered lattice networks and the completely random networks described by the Erdös-Rényi model, capturing the small-world properties observed in many real-world networks.

The Watts-Strogatz model starts with a regular lattice network, where each node is connected to its k nearest neighbors in a ring topology. The model then introduces a small amount of randomness by rewiring each edge with probability p, where p ranges from 0 (no rewiring, and thus a regular lattice) to 1 (complete rewiring, resulting in a random graph).

Figure 4.7 shows the critical steps in the process are as follows:

- Start with a regular lattice with a ring of N nodes, each connected to its k nearest neighbors.
- Small-world is found by rewiring it with probability p to a node chosen uniformly at random over the entire ring, avoiding self-loops and duplicate edges. The average shortest path length between nodes in a small-world network is shorter than expected for a regular lattice.
- A random network is found for probability $p = 1$. Given that an edge between two nodes is independent of any other edge in the network, the clustering coefficient is lower than in the small-world network.

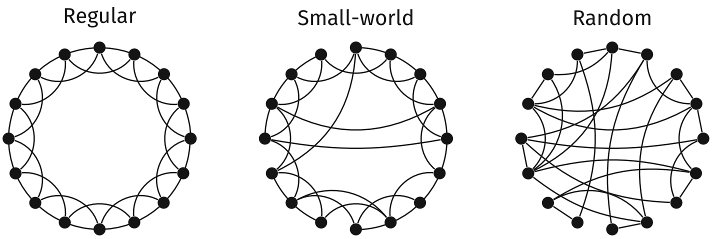

Fig. 4.7 The three types of networks and probability of rewiring (Adapted from https://www.researchgate.net/figure/Algorithm-of-Watts-Strogatz-model-which-can-be-tuned-by-parameter-p-0-1_fig1_258120851)

By bridging the gap between regular and random networks, the Watts-Strogatz model highlights the importance of network topology in determining functional properties such as information transmission and robustness. The model's ability to capture the essence of small-world networks makes it fundamental to studying complex systems and network science.

4.4.4 Cliques, Relaxations and Structural Holes

Cavique et al. (2018) study the themes described in this sub-section, focusing on data reduction to visualize communities and brokers in social networks.

A clique in a graph is defined as a complete sub-graph where each member has direct connections with every other member or node. It becomes a maximal clique if it cannot be included within any other clique. The maximum clique, denoted as the clique with the largest number of nodes, poses an NP-hard problem. To establish a lower bound for this maximization problem, researchers have proposed a heuristic method by Johnson (1974) and a meta-heuristic approach employing Tabu Search, which was developed by Soriano and Gendreau (1996). Further insights into these methods and related works can be found in the studies by Cavique et al. (2002) and Cavique and Luz (2009).

The requirement for every pair of vertices to have an edge in a clique structure imposes significant limitations, which may not be realistic in social networks where such fully connected structures are rare. Alternative approaches for more flexible and cohesive groups have been proposed, including concepts such as k-clique, k-clan/k-club, and k-plex, as suggested by Scott (2000). This study introduces a combination of k-cliques to provide a more adaptable and realistic representation of network structures.

Relaxations: k-clique, k-clan/k-club and k-plex Distance and diameter serve as valuable metrics for analyzing social networks. The distance between vertices u and v in the graph $G(N, E)$ is symbolized as $d(u, v)$, representing the shortest

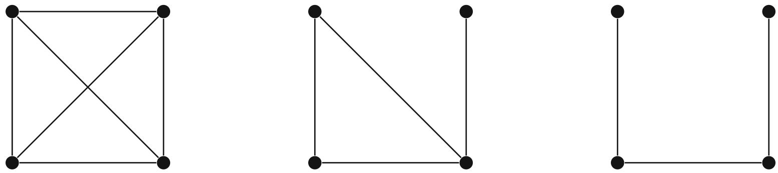

Fig. 4.8 Examples of a 4-node graph of 1-clique, 2-clique and 3-clique

path length between them. The diameter of G, denoted as $diam(G)$, is defined as $diam(G) = \max d(u, v)\ \forall\, u, v \in \mathbb{N}$.

Luce (1950) introduced the concept of a dense group based on distance, known as a k-clique, where k represents the maximum path length between any pair of vertices. A k-clique is a subset of vertices denoted as C, for which the distance $d(i, j) \leq k$. Specifically, a 1-clique is essentially equivalent to a regular clique, as the distance between its vertices is limited to just one edge. On the other hand, a 2-clique corresponds to a maximal complete sub-graph with a path length of either one or two edges. To illustrate a path distance of two, consider it the "friend of a friend" connection in social relationships. In social networking platforms like LinkedIn, each member can access their immediate connections and those two or three degrees away. As the value of k increases, it leads to a gradual relaxation of the criteria for clique membership. Refer to Fig. 4.8 for visual representation.

In graph theory, the concept of the k-th power of a graph, denoted as G_k, creates a new graph where two vertices are connected if their distance in the original graph G is at most k. Utilizing the k-th power of graph G in conjunction with a maximum clique algorithm allows us to identify all the maximal k-cliques within the graph. During this transformation process, additional edges are introduced to ensure that the distance between every pair of nodes is at most k.

A limitation of the k-clique concept is that some vertices may be distant from the group. The distance between two nodes may correspond to a path involving nodes not belonging to the k-clique. The cohesion groups based on diameter, called k-club and k-clan, were introduced to overcome this handicap. To find all k-clans, all the k-cliques S_i must be found first, and then the restriction $diam(G[S]) \leq k$ must be applied to remove the undesired k-cliques. In Fig. 4.9, the left graph shows the 2-clique $\{1, 2, 3, 4, 5\}$ is not a 2-clan because $d(4, 5) = 3$. Path 4-6-5 is not possible as node 6 does not belong to the sub-graph with the 2-cliques. Another approach to these diameter models is the k-club, which is defined as a subset of vertices S such that $diam(G[S]) \leq k$. In the same graph two 2-cliques can be found: $\{1, 2, 3, 4, 5\}$ and $\{2, 3, 4, 5, 6\}$, one 2-clan: $\{2, 3, 4, 5, 6\}$ and three 2-clubs: $\{1, 2, 3, 4\}$, $\{1, 2, 3, 5\}$ and $\{2, 3, 4, 5, 6\}$.

An alternative way of relaxing a clique is the k-plex concept, which considers the degree of the vertices. The degree of a graph's vertex is the number of edges incident to the vertex, denoted by $deg(v)$. The maximum degree of graph G is the maximum degree of its vertices and is denoted by $\Delta(G)$. On the other hand, the minimum degree is the minimum degree of its vertices and is denoted by $\delta(G)$. A

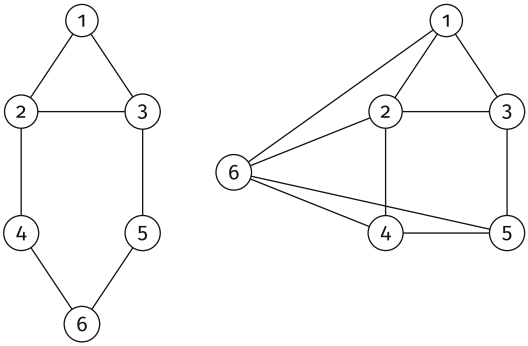

Fig. 4.9 2-clans, 2-clubs (left), and 3-plex (right)

subset of vertices S is a k-plex if the minimum degree in the induced sub-graph $\delta(G[S]) \geq |S| - k$. In Fig. 4.9, the right graph shows $|S| = 6$, and the degree of vertices 1, 2, 3, 4, and 5 does not exceed the value 3. Thus, the minimum degree in the induced sub-graph $\delta(G[S])$ is 3. For $|S| = 6$, $k = 3$ is obtained.

Connectivity of Graphs and Brokerage A critical concept in community structure metrics is the notion of graph connectivity. This sub-section discusses node-connectivity and line-connectivity in graphs (Harary 1969). In a connected graph, all nodes are reachable. On the contrary, a graph is disconnected if no path exists between any pair of nodes. A graph with only one node is also considered connected. Graph G can lose its connectivity when specific lines or nodes are deleted.

A node n is a cut-point (or cut-vertex) if the number of components in G_1 is fewer than the number of components in G_2 that results from deleting node n. Analogously, a bridge (or line-cut) is an edge critical to the graph's connectedness. Line l is a line-cut if the number of components in G_1 is fewer than the number of components in G_2 that results from deleting line l. Generalizing, a k-line-cut is a set of lines that, if deleted, disconnects the graph. A bridge is a 1-line-cut. Similarly, the k-cut-point is a set of nodes that, if removed, makes the graph disconnected.

Following the Triadic Closure property, Burt (1992) developed a complementary approach called 'Structural Holes', which refers to the absence of links in a connected organization. He also introduced the term brokerage, meaning nodes that connect two dense groups.

Figure 4.10 shows two ways of spanning structural holes using a bridge or a broker. Structural hole, bridge, and broker can be defined as follows:

- Structural hole refers to the lack of edges between components or communities;
- Bridge is an edge whose removal increases the number of components in the network;
- Broker or cut-vertex is a vertex whose deletion increases the number of components in the network.

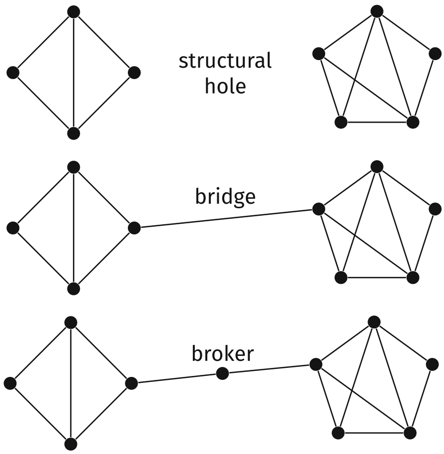

Fig. 4.10 Structural hole, bridge, and broker

Brokerage across structural gaps opens up fresh opportunities and fosters the exchange of innovative ideas. It is pretty standard to find structural holes between business unit organizations, which often result in rewards for the intermediary. Two distinct broker strategies come into play: the altruistic broker, referred to as the “Tertius iungens”, embodies the idea of joining and connecting different communities, while its counterpart, the egoistic broker, known as the “Tertius gaudens,” focuses on personal gain within these connections.

For many years, the centrality of actors within networks has been a significant concern in social network analysis. Central nodes, also known as hubs, can be identified using various metrics such as degree centrality, betweenness centrality, closeness centrality, or eigenvector centrality. Brokers share similarities with hubs, as both exhibit high centrality scores. However, the key distinction lies in the role of brokers, which involves bridging between different communities, whereas hubs are typically integral parts of a specific community.

4.4.5 Network Visualization

Visualization is crucial in the topology of networks, highlighting the intricate relationships within various types of networks. Mapping connections visually provides immediate insight into the network’s structure and key nodes, which might be otherwise hidden in complex data. They also bridge complex network theory and practical application, allowing for more informed decision-making. We propose five tools commonly used for network visualization and analysis.

- **Gephi**: An open-source network analysis and visualization software package written in Java, primarily for exploring and understanding graphs.
- **Cytoscape**: An open-source software platform for visualizing complex networks and integrating these with any attribute data.

Table 4.3 Tools performance concerning the number of nodes

	100 nodes	10,000 nodes	1,000,000 nodes
Gephi	Excellent for visualization	Performance issues	Not practical
Cytoscape	Excellent for visualization	Performance issues	Not practical
NetworkX	Excellent for visualization	Performance issues	Not practical
igraph	Excellent for Visualization	Suitable for this scale	Can handle some analyses
Graph-tool		Very efficient	Strong for analysis

- **NetworkX**: A Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.
- **igraph**: The igraph software is available in C/C++ and for the Python and R programming languages. It is a versatile library used across different platforms and in various programming environments.
- **Graph**-tool: An efficient Python module for manipulation and statistical analysis of graphs. It is built on top of the C++ Boost Graph Library, which makes it highly efficient and suitable for large networks.

Table 4.3 shows the tools' performance concerning the number of nodes. The tools Gephi, Cytoscape, and NetworkX perform better until 10,000 nodes are reached. On the other hand, the tolls igraph and Graph-tool are most appropriate for large networks.

4.5 Network Dynamics

This section covers percolation and diffusion models, relating percolation in physics to the spread of ideas, behaviors, or products in networks. Then, it elaborates on the SIR (Susceptible-Infected-Recovered) model, a cornerstone epidemiological model used to analyze how infectious diseases spread through populations. Finally, discusses the role of intelligent agents in the evolution of complex networks, highlighting their significance in modeling, simulating, and understanding network dynamics and emergent behaviors.

4.5.1 Percolation and Diffusion Models

In physics, percolation refers to the movement and filtering of fluids through porous materials. It is a critical concept in studying the behavior of complex systems and phase transitions, such as the transition from a non-conductive to a conductive state in materials as connections between conducting components reach a critical threshold. This concept is used in various fields, including materials science, geology, environmental engineering, and the propagation of forest fires. Figure 4.11 shows an example of percolation in its final phase.

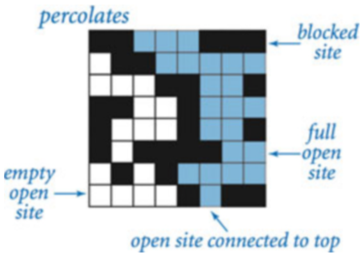

Fig. 4.11 Percolation in the final phase (Source: https://medium.com/@mind_sFlight/union-find-data-type-disjoint-set-percolation-1028b31e036d)

Percolation is related to the concept of model diffusion. Model diffusion in networks refers to the process by which ideas, behaviors, or products spread through a population via individual connections (Easley and Kleinberg 2010). This process is influenced by the structure of the network and the rules governing how an entity decides to adopt something new based on its neighbors' choices. Diffusion models explore how these dynamics lead to the widespread adoption or rejection of innovations within a network.

An example of model diffusion is the spread of a new technology like smartphones. Initially, a few individuals adopt smartphones because they see a unique value in them (innovators). As more people observe these innovators and see the benefits, they, too, start using smartphones (early adopters). Gradually, the trend picks up, and a more significant segment of the population, influenced by the adoption of their network, begins using smartphones (early majority and late majority). Eventually, even the most resistant individuals (laggards) may adopt due to network pressures or seeing widespread utility.

In the diffusion of ideas, it is relevant to differentiate between cascades and clustering. The cascades refer to the rapid spread of ideas or behaviors across a network facilitated by a set of initial adopters. Clustering, on the other hand, represents tightly-knit groups within the network where members are densely connected. A complete cascade, where an idea spreads throughout the entire network, can occur if no clusters exist. The clusters act as barriers to the spread of cascades; their dense interconnections make it difficult for new ideas to penetrate these groups. Therefore, clusters in a network can halt or significantly slow down the spread of a cascade, making them a crucial factor in understanding the dynamics of idea diffusion.

4.5.2 SIR Contagion Model

The SIR model is a fundamental compartmental model used in epidemiology to understand how infectious diseases spread through populations (Newman 2018). The acronym SIR stands for Susceptible, Infected, and Recovered, representing the three compartments into which the population is divided based on their disease status:

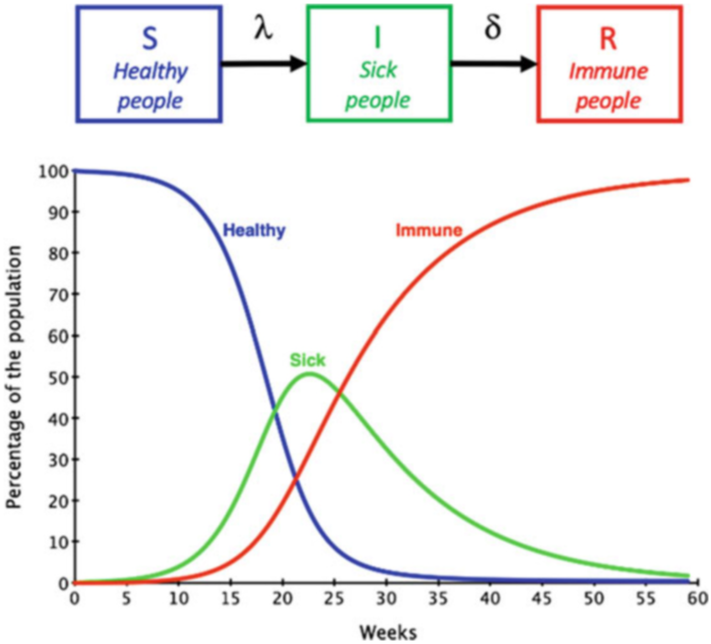

Fig. 4.12 SIR model (Source: https://www.researchgate.net/figure/SIR-Model-with-infection-density-Here-b-represents-the-infection-density-y0-represents_fig1_359517106)

- Susceptible (S): Individuals who have not yet been infected with the disease are at risk of infection if they contact an infected person.
- Infected (I): Individuals currently infected with the disease can transmit it to susceptible individuals. These individuals are in the stage where they can spread the disease to others.
- Recovered (R): Individuals who have been infected and have either recovered from the disease, gained immunity to it, or have died. In the context of the SIR model, recovered individuals are typically assumed to be immune and no longer at risk of being reinfected or spreading the infection.

Figure 4.12 shows the percentage of the population in each status. The pipeline with the rates of infection and recovery are also exhibited.

The model describes the flow of individuals between these compartments over time, driven by the rates of infection and recovery:

- Transmission rate (λ): The rate at which susceptible individuals become infected after contacting an infected individual.
- Recovery rate (δ): The rate at which infected individuals recover (and gain immunity) or die, thereby moving from the Infected compartment to the Recovered compartment.

The dynamics of the disease are typically described using a set of differential equations that represent the rate of change of the number of individuals in each compartment. These equations consider the transmission and recovery rates and the interactions between susceptible and infected individuals.

Several epidemiological models are related to the SIR model, each extending or modifying the basic framework to capture the nuances of different diseases better or incorporate additional factors affecting disease spread. Some of these related models include:

- SIS Model: In the Susceptible-Infected-Susceptible (SIS) model, individuals move from being susceptible to infected and back to susceptible again without entering a recovered state with immunity. This model applies to diseases where infection does not confer lasting immunity, allowing individuals to be reinfected repeatedly.
- SIRS Model: In the Susceptible-Infected-Recovered-Susceptible (SIRS) model, individuals can lose their immunity over time and become susceptible again. This model is relevant for diseases where immunity wanes after recovery, allowing for the possibility of reinfection.
- SVIR Model: The Susceptible-Vaccinated-Infected-Recovered (SVIR) model includes a vaccinated class to account for individuals who have been immunized against the disease and are, therefore, less likely to become infected.
- SIRD Model: The Susceptible-Infected-Recovered-Deceased (SIRD) model explicitly accounts for deaths due to the disease, separating those who die from those who recover with immunity.

Each model tailors the basic SIR framework to reflect specific diseases' characteristics more accurately or include additional epidemiological or demographic factors. The choice of model depends on the disease being studied, the available data and the specific aspects of the disease spread that researchers or public health officials are most interested in understanding.

4.5.3 Intelligent Agents

In the evolution of complex networks, intelligent agents play a significant role in modeling, simulating, and understanding these networks' dynamics and emergent behaviors. Intelligent agents autonomously make decisions and act within their environments to achieve specific goals, characterized by their ability to perceive, react, and adapt. They interact socially with other agents or humans and can learn from their experiences to improve over time. These agents are crucial in robotics, virtual assistants, and simulations, driving advancements and efficiencies across diverse applications.

To implement intelligent agents, NetLogo, as a software platform, is a powerful tool for implementing and visualizing models of complex networks. Figure 4.13 shows the classic survival struggle of the grass, sheep, and wolves.

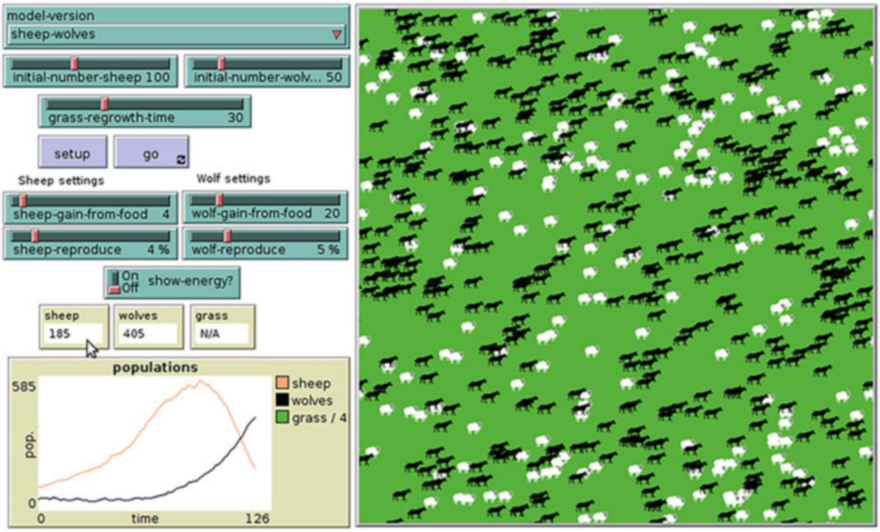

Fig. 4.13 Grass, sheep, and wolves' model in NetLogo (Source: https://edutechwiki.unige.ch/en/NetLogo_Wolf_Sheep_Predation_model)

NetLogo allows the simulation of environments and the visualization of the results. NetLogo provides an accessible environment for creating models of complex networks populated by intelligent agents. It allows users to define agents' behavior, interaction rules, and the network's structure, making it an ideal platform for simulating complex systems. Moreover, NetLogo offers powerful visualization tools that help understand complex network dynamics. Users can visually track the evolution of the network and the agents' behaviors over time, facilitating the analysis of emergent patterns and network properties.

The LabMAg (agent modeling laboratory) in FCUL (faculty of sciences at the University of Lisbon) has worked in this area. An example of their work can be found in Lemos et al. (2017) in ProtestLab. ProtestLab is an agent-based model for simulating street protests, featuring diverse agents such as protesters, police, and media. It includes scenario elements like obstacles and attraction points, and agents can exhibit various behaviors and states, including violence. The model offers quantitative metrics to analyze crowd dynamics, protest intensity, and police effectiveness. It was tested on a scenario resembling anti-austerity protests in Lisbon, accurately replicating observed real-life protest behaviors and interactions.

4.6 Conclusions

Network science captures the essence of the interconnectivity in an organization, allowing us to visualize it as a skeleton of the system. Nowadays, in the big data era, extracting new knowledge is more and more relevant. One way to do it is to map that data into networks and explore their connectivities. Network science is characterized by a vast and varied array of high-quality contributions, making any new approach a challenge.

This work offers a structured exploration of network science through a three-step maturity framework that progresses in complexity. Initially, it outlines the fundamentals of networks and their construction, followed by an overview of standard network metrics. The final stage explores two key concepts: network topology and dynamics. Software packages for topological visualization and dynamic understanding are provided.

We believe this work clarifies the field and inspires readers to explore it. While network science may not solve every problem, it serves as a valuable framework.

References

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Barabási, A. L., & Frangos, J. (2014). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. Basic Books. ISBN: 9780465038619.

Barabási, A. L., Pósfai, M. (2016). *Network science*. Cambridge University Press. ISBN: 9781107076266.

Berge, C. (1970). *Graphes et hypergraphes*. D'dunod université: Mathématiques. Dunod.

Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks, 23*(3), 191–201.

Burt, R. S. (1992). *Structural holes: The social structure of competition*. Harvard University Press. ISBN: 9780674843714.

Cavique, L., & Luz, C. J. (2009). A heuristic for the stability number of a graph based on convex quadratic programming and tabu search. *Journal of Mathematical Sciences, 161*, 944–955.

Cavique, L., Marques, N. C., & Gonçalves, A. (2018). A data reduction approach using hypergraphs to visualize communities and brokers in social networks. In *Social network analysis and mining* (vol. 8, issue 1, p. 60). Springer.

Cavique, L., Rego, C., & Themido, I. (2002). A scatter search algorithm for the maximum clique problem. In *Essays and surveys in metaheuristics* (pp. 227–244). Springer. ISBN: 978-1-4615-1507-4. https://doi.org/10.1007/978-1-4615-1507-4_11.

Christakis, N., & Fowler, J. (2010). *Connected: The amazing power of social networks and how they shape our lives*. Harper Collins Publishers. ISBN: 9780007356423.

Easley, D., & Kleinberg. J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press. ISBN: 9781139490306.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication, 12*(4), 1143–1168.

Erdös, P., & Rényi, A. (1959). On random graphs I. *Publ. Math. Debrecen, 6*, 290–297.

Grossman J., Ion, P., & Castro, R. D. (2024). The Erdös number project. Retrieved Jan 2024 from http://www.oakland.edu/enp/.

Harary, F. (1969). *Graph theory. Addison-Wesley series in mathematics*. Addison-Wesley. ISBN: 9788185015552.

Johnson, D. S. (1974). Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences, 9*(3), 256–278.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39–43.

Landherr, A., Friedl, B., & Heidemann, J. (2010). A critical review of centrality measures in social networks. *Business & Information Systems Engineering, 2*, 371–385.

Laranjeira, P. A., & Cavique, L. (2014). Métricas de centralidade em redes sociais (in Portuguese). *Revista de Ciências da Computação, 9*, 1–20.

Lemos, C. M., Coelho, H., & Lopes, R. J. (2017). Protestlab: A computational laboratory for studying street protests (pp. 3–29). Springer.

Luce, R. D. (1950). Connectivity and generalized cliques in sociometric group structure. *Psychometrika, 15*(15), 159–190.

Milgram, S. (1967). The small world problem. *Psychology Today, 2*(1), 60–67.

Moreno, J. L., & Jennings, H. H. (1934). *Who shall survive? A new approach to the problem of human interrelations*. Nervous and Mental Disease Monograph Series. Beacon House.

Newman, M. (2018). *Networks*. OUP Oxford. ISBN: 9780192527493.

Newman, M., Barabási, A. L., & Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton Studies in Complexity. Princeton University Press. ISBN: 9780691113579.

Page, L., et al. (1998). The pagerank citation ranking: Bringing order to the web. Tech. Rep. Stanford University. https://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/pagerank.pdf.

Rapoport, A. (1957). Contribution to the theory of random and biased nets. *The Bulletin of Mathematical Biophysics, 19*, 257–277.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika, 31*(4), 581–603.

Scott, J. (2000). *Social network analysis: A handbook*. SAGE. ISBN: 9780761963394.

Soriano, P., & Gendreau, M. (1996). Tabu search algorithms for the maximum clique. *Clique, Coloring and Satisfiability, Second Implementation Challenge DIMACS* (pp. 221—242). American Mathematical Society.

Watts, D. J. (2004). *Six degrees: The science of a connected age*. W. W. Norton. ISBN: 9780393076127.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440–442.

Luís Cavique is a tenured Assistant Professor at the Computer Science Section in the Department of Sciences and Technology at Universidade Aberta. He worked in the Polytechnic Education System from 1991 to 2008, namely as Adjunct Professor in the Setubal and in the Lisbon Polytechnic Institute. He received the degree in Computer Science Engineering from the New University of Lisbon (FCT-UNL), the MSc degree in Operational Research and Systems Engineering from the Technical Lisbon University and the PhD degree in Systems Engineering from the Technical Lisbon University (IST-UTL) in 2002. His research areas are in the intersection of Computer Science and Systems Engineering, particularly in Data Science. He is author of over 200 scientific works in peer reviewed journals and conferences. He is an integrated member of LASIGE in Lisbon University.