Network Science–A Brief Overview

Natalie Nagata
Dr. William DeMeo
University of Hawai'i at Mānoa

May 16, 2017

1 Introduction

Science and technology has advanced at ever increasing rates over the past few decades. As a result, the study of complex networks has become a field of huge interest for having wide-ranging potential and applicability. By using modern graph theory, we can analyze the structure of networks to understand characteristics that weren't previously well understood. We shall provide a brief discussion of some different models ised in network analysis and their applications. This paper is not meant to be comprehensive but aims to provide a short overview for readers who are new to network science. But first, before discussing some basic models in network science, it is helpful to go over some basic concepts in graph theory.

2 Basics

The fundamentals of network science are based in graph theory. A graph $G = (V, E)$ is defined as an ordered pair which consists of a nonempty set of vertices V and a set of edges E.

For the remainder of this paper, we shall denote the number of vertices as $N = |V|$ and the number of edges as $L = |E|$. Each vertex can be labeled such that $V = \{v_1, v_2, \ldots, v_N\}$ and, similarly, each edge can be labeled such that $E = \{e_1, e_2, \ldots, e_L\}$. Each edge must be connected to one or two vertices (sometimes called *endpoints* [1]). Thus each edge may be represented as Visually, you may think of graphs/networks as a diagram that consists of points (vertices) and there may be lines (edges) connected between two points which represent relationships between two points. (See Fig 1) Note that graph and network theorists essentially study the same systems but describe them within their respective fields using different terminology. Graph theorists tend to consider "vertices" and "edges" whereas network theorists study "nodes" and "links'.' All of these terms shall be used interchangeably from here on (although the symbolic representation shall remain as mentioned previously).

2.1 Types of Graphs

Graphs and networks are categorized in a variety of ways. A *simple graph* is a graph such that there exists at most only one edge between any pair of vertices and does not contain any looped edges (edges that begin and end at the same vertex). Note that a graph does not necessarily contain edges. Nodes that have no adjacent edges are referred to as *isolates*. Graphs that are not simple are known as *multigraphs* since they may contain multiple links between the same pair of vertices. Multigraphs may also contain loops. (See Fig. 1)

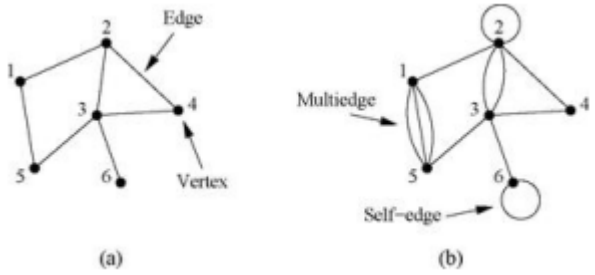

Figure 1: (a) a simple undirected graph, (b) an undirected graph with multiedges and loops. [2]

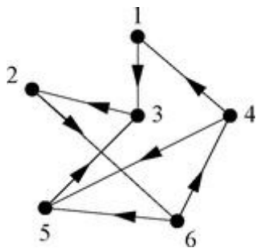

Figure 2: a directed graph with arrows representing the direction of edges [2]

Another distinguishing property is whether a network is directed or undirected. An *undirected* network contains edges that have no defined "direction" associated with it. So an edge with endpoints v_a and v_b is the same as an edge with endpoints v_b and v_a. A *directed* network has edges with an ordered pair of vertices as endpoints. (See Fig. 2)

2.2 Representation

As we've seen before, graphs can be visually drawn as a collection of vertices connected by edges (see Fig. 1). But how do we objectively quantify the relationships a system has between nodes and edges?

Often we can represent a graph numerically with what's known as an *adjacency matrix*. Consider an adjacency matrix $\mathbf{A}$. For a simple undirected graph, the entries a_{ij} of its adjacency matrix $\mathbf{A}$ are such that

$$\mathbf{A} = \begin{cases} 1 & \text{if an edge exists between } v_i \text{ and } v_j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For example, the adjacency matrices for Fig. 1(a) and Fig. 1(b) are

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{2}$$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 3 & 0 \\ 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{bmatrix} \tag{3}$$

For directed networks, their adjacency matrices are based on a similar idea as previously described. The main difference is that the row and column that each entry belongs to corresponds to the beginning and endpoint of each edge. Thus,

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if an edge exists } \textit{from } v_j \textit{ to } v_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Thus, the adjacency matrix for the directed network in Fig. 2 is

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5}$$

2.3 Measurements

There are different properties of graphs which we can numerically measure as well. For undirected netowkrs, the number of edges that each vertex v_i is adjacent to is called the *degree* of v_i and is denoted as k_i. Thus, the total number of links L of a network is

$$L = \frac{1}{2} \sum_{i=1}^{N} k_i \tag{6}$$

For directed networks, we need to make the distinction between edges that point towards a vertex versus edges that point away from a vertex. For each vertex v_i, there is a corresponding *in-degree* and *out-degree* value, k_i^{in} and k_i^{out}, that represent the number of edges pointing inwards to v_i and outwards to v_i, respectively.

$$k_i = k_i^{in} + k_i^{out} \tag{7}$$

The average degree of a network is defined as the sum of the degrees of each vertex divided by N. As such, for any graph G, its average degree is

$$avg(k) = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N} \tag{8}$$

For directed graphs the definition differs slightly to account for in- and out-degree distinction. The total number of links L is

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} \tag{9}$$

From Eqn. 8 and 9, it follows that

$$avg(k) = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \frac{1}{N} \sum_{i=1}^{N} = k_i^{out} = \frac{L}{N} \tag{10}$$

In addition to average degree, the degree distribution of a network is a very measurement in network science. The *degree distribution* of a network describes the probability p_k that a randomly selected node is of degree k, where $\sum_{i=0}^{\infty} p_k = 1$. An example of the degree distribution representation network is provided below.

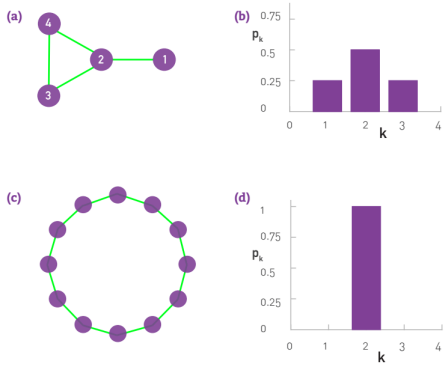

Figure 3: Two representations of the degree distribution of two small networks [3]

Additionally, the concept of *geodesic distance* (or just *distance*) is important when studying networks. This is analogous to the idea of paths in graph theory. Informally, a *path* is a sequence of n vertices and $n-1$ edges that begin at vertex v_1 and end at some vertex v_n [1]. In network theory, the geodesic distance generally refers the shortest possible path between two vertices. It is also possible for no geodesic distance to exist between two vertices. These vertices are disconnected from each other. Usually they belong to different components within a network or one or both nodes are isolates. In contrast, the *diameter* of a network is the largest distance in a network, denoted as d_{max}. We will see later how the distance, average degree, and diameter of a network are related.

3 Models

3.1 Random Graphs

The overall motivation of network science is to create and study models of real-world networks. These real-world networks differ from other systems such as a lattice. The most striking differences are with a network's lack of repetitive and orderly structure like that of a lattice. One of the earliest models is that of a *random graph*. It was first studied by Anatol Rapoport in 1951 but was brought to more well-known attention by Pál Erdős and Alfréd Rényi in 1959 [3]. They are historically recognized as the founders of random graph theory by merging techniques from the fields of combinatorics, graph theory, and probability. As a result, random graphs are generally called *Erdős-Rényi graphs*.

There are two ways of defining random networks. A graph that follows the *G(N, L)* model depends on a fixed N total number of nodes and L total number of links. In the $G(N, L)$ model, a collection of all possible configurations of graphs with N nodes and randomly placed L links is generated. As such, each unique graph has an equivalent chance of being selected from this collection. The graph selected from this collection is the graph that is constructed.

In contrast, the *G(N, p)* model depends on the N total number of nodes and a fixed probability p that an edge exists between each pair of nodes of a network. Starting with a collection of N isolates, a random number between 0 or 1 is generated for each pair of nodes in the system. Depending on whether the number falls above or below the probability p is what determines whether a link connects each pair of nodes.

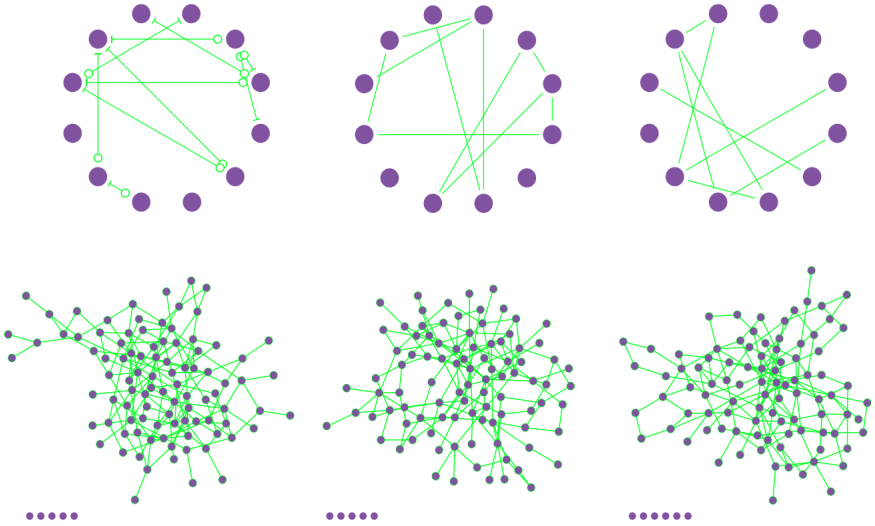

Figure 4: Three different representations of networks generated from the same $G(N, p)$ model with parameters $N = 12$ and $p = \frac{1}{6}$ [3]

On average, random networks are predicted to consist of some nodes that have more links than others. Since p_k is the probability that a randomly chosen node is of degree k, the degree distribution of a network can be exactly described by the binomial distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-k-1} \tag{11}$$

As $N \to \infty$, the degree distribution can be approximated by the Poisson distribution

$$p_k = e^{-avg(k)} \frac{avg(k)^k}{k!} \tag{12}$$

This behavior can be seen in Fig. 5

In more recent years, technology has enabled us to effectively construct network maps of real-world systems. This has led us to a surprising observation: real-world networks are not random. Barabasi highlights this by showing three examples of real-world networks and how their corresponding degree distributions deviate from their expected Poisson forms.

What is the driving reason behind these differences? Some of it can be explained by nodes that are of unusually high degree. These high-degree nodes are known as *hubs*. Hubs essentially dominate the

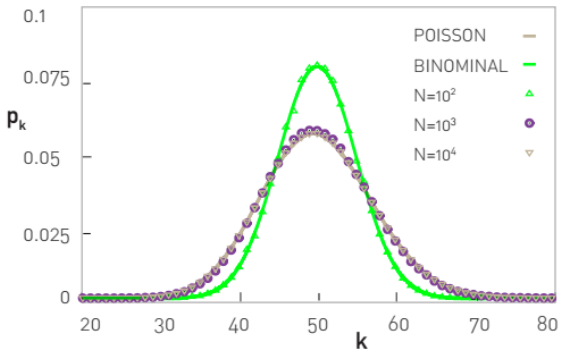

Figure 5: The degree distribution of three networks with $avg(k) = 50$: a small network ($N = 10^2$) which fits a binomial distribution (see Eq. 11) in the green line, and large networks ($N = 10^3$ and $N = 10^4$ which approach the Poisson distribution (see Eq. 12) in the gray line. [3]

characteristics of networks. As we'll see in the next section, discussing hubs leads to an extension of the random network model called the Watts-Strogatz model.

3.2 Real-World Networks and the Small World Phenomenon

The discrepancies between real-world systems and that of random-network models are revealed on closer inspection of our assumption that a Poisson distribution is the best fit. The term $1/k!$ in Eq. 12 results in greatly underestimating the number of hubs. Also the variance of nodes of average degree is much wider in real-world networks that that of random networks characterized by the Poisson distribution. The presence of hubs leads to an idea known as the *small world phenomenon*. This is more widely known as *six degrees of separation*.

The small world phenomenon stems from the Watts-Strogatz model which was proposed around 1998 [3] by Duncan Watts and Steven Strogatz. The Watts-Strogatz model makes two observations:

1. **Networks have a "small world" property**
   In simple terms, the small world aspect of networks is essentially the idea that given two randomly chosen nodes in a network, the distance between them is short. Of course, "short" is an ambiguous term. We shall discuss this in further detail below.

2. **Networks contain high clustering**
   A real network of N nodes and L links actually contains noticeably higher clustering of nodes and links than what is predicted by a random network of the same N and L values.

3.2.1 Small World Property

According to Barabasi [3], the small world property can be easily understood by looking at the relationship between a network's average degree and average number of neighbors a randomly chosen vertex has. The derivation of this property is beyond the scope of this paper but the relationship is summarized below.

Suppose we have a network of average degree $avg(k)$ and we randomly select a node v_i. On average, v_i will have $avg(k)$ neighboring vertices of distance $d = 1$ away from it. As we travel further away from v_i a relationship becomes clear. On average, v_i has $\left[avg(k)\right]^2$ neighbors of distance $d = 2$ away from it, and

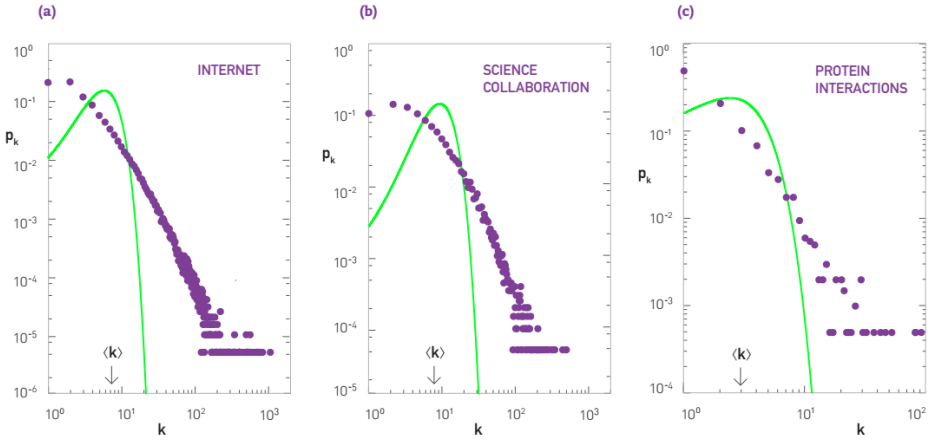

Figure 6: Plots of the degree distribution of three real networks (in purple) compared with their predicted Poisson degree distribution based on their measured average degree k. (a) Internet websites, (b) scientific collaborations based on PI's listed in published papers, (c) protein-protein interactions of biological systems [3]

$\big[avg(k)\big]^3$ neighbors of distance $d = 3$ away from it. Thus, the average number of neighbors any given node has increases to the power of distance d as we travel further away from a node. This means that although networks may be incredibly large, the shortest path between any two given vertices is surprisingly small. In fact, the average distance between two vertices in a network is described by

$$avg(d) = \frac{\ln N}{\ln\big(avg(k)\big)} \tag{13}$$

So our small world property is more objectively defined by saying the average diameter of a network is logarithmically proportional to the total number of vertices of that network.

3.2.2 High Clustering

Although studying the degree of nodes is useful, it lacks describing the relationship between a node and its neighboring vertices. We can quantify this relationship with what's known as the *clustering coefficient*. For a vertex v_i with degree k_i, its clustering coefficient is

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{14}$$

where L_i is the number of links that exist amongst the set of neighbors of distance $d = 1$ away from v_i. Essentially, C_i describes the density of edges in the neighborhood of a vertex. As we have been able to study more real-world systems, we've found that real-world networks have a much higher average clustering coefficient than what is predicted by random network models.

As an extension of the random network model, the Watts-Strogatz model describes a network that falls between the orderly characteristics of a lattice and the complex nature of a purely random network. The Poisson degree distribution of random networks provides an upper bound description of the degree distribution of networks that have the small world property.

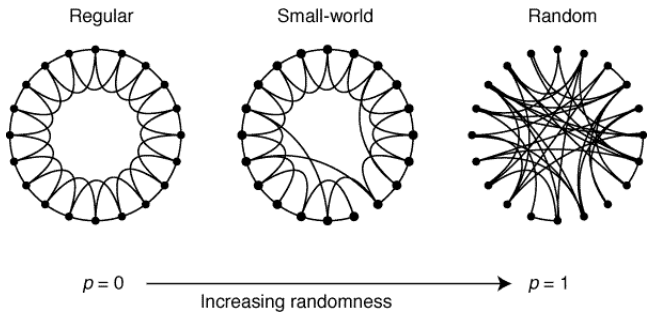

Figure 7: Diagram of how networks' properties change as randomness is increased [4]

3.3 Scale-Free Networks

As we saw previously, although random networks predict a degree distribution that follows binomial or Poisson distributions, networks in the real world do no behave in this manner. This led to the development of what are known as scale-free network models which are much better at characterizing the behavior of real network systems. Scale-free models follow what's known as a power law distribution:

$$p_k \approx k^{-\gamma} \tag{15}$$

The major differences between random and scale-free models are that random networks fail to capture the effect of hubs and clustering on an entire system. This is illustrated in Fig. 3.3 when looking at how a predicted degree distribution of a random network compares to that of an actual network.

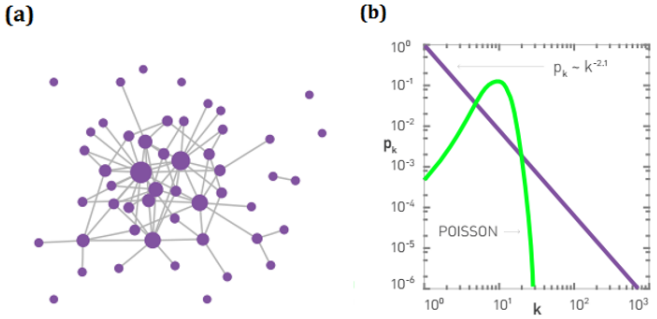

Figure 8: A scale-free network with $avg(k) = 3$ and $N = 50$. The predicted Poisson distribution is plotted in green, the network's actual distribution is plotted in purple. (a) diagram of the network where nodes are sized according to degree, (b) a log-log plot of the network's degree distribution [3]

It is important to understand the implications behind what it means for a network to be "scale-free." Consider the degree k of a randomly selected node. In random networks, the average degree of this chosen node falls within a range of $avg(k) \pm \sigma_k$. However, when selecting a node from a scale free network, the range of possible values of the degree of this node is so large that there is no measurable scale.

Consider the nth moment of a degree distribution (from statistics). The first moment is the average degree of a network, and the second moment ($n = 2$) is what gives us the variance of the average degree of

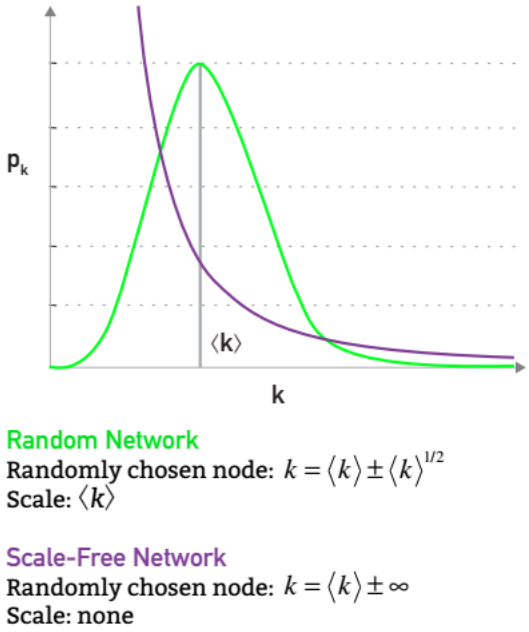

Figure 9: Degree distribution of a random (green) and scale-free network (purple) [3]

the network. As $N \to \infty$, the second moment of a Poisson distribution is always smaller than than its first. But in contrast, the second moment of scale-free networks can diverge. (See Fig. 9)

4 Conclusion

We have presented a brief overview of some essential models that have arose from the development of the network science field. Although the field itself is still relatively new, it has far-reaching applications to a surprisingly diverse range of fields.

Scale-free models are astoundingly universal. These systems can be found in nature, man-made constructs such as the internet. Fields such as connectomics in neuroscience aim to study brain networks on micro-, meso-, and macro-scale levels. The rise of social networks like Facebook and Instagram is now heavily studied by advertisers and social engineers alike. Even sometime as simple as optimizing airline routes reveals an underlying scale-free network. (see Fig. 10.)

It is outside the scope of this paper, but it is worth mentioning that there are many other models and topics not covered here that are relevant to our everyday lives. For instance, the centrality measurements focus on defining the "importance" of nodes based on their degree, weighted degree, and degree of neighboring vertices. Different algorithms exist to carry out these calculations and each has their own strengths and weaknesses in terms of how accurate, efficient, and useful they are. Probably the most well-known algortihm is *PageRank*, the trade named algorithm Google created for its web search engine.

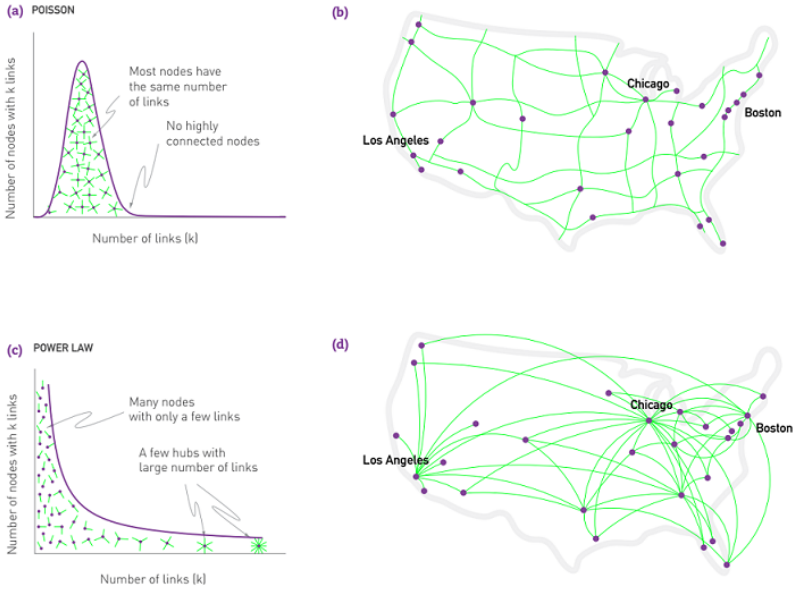

Figure 10: A visual map representation of what potential airline routes that follow a Poisson (random network) distribution versus a power law (scale-free network) distribution. [3]

References

[1] Kenneth H. Rosen. *Discrete Mathematics and Its Applications.* McGraw-Hill Companies Inc., New York, NY 10020, 2012.

[2] Mark Newman. *Networks: An Introduction.* Cambridge University Press, Cambridge, United Kingdom, 2010.

[3] Albert-Laszlo Barabasi. *Network Science.* Cambridge University Press, Cambridge, United Kingdom, 2016.

[4] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.