6
Statistical Methods in Graphs: Parameter Estimation, Model Selection, and Hypothesis Test

Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, João Ricardo Sato, Carlos Eduardo Ferreira, and André Fujita

6.1 Introduction

Graph theory dates back to 1735, when Leonhard Euler solved the Königsberg bridge problem [1]. Since then, this field has been contributing to several areas of knowledge, such as discrete mathematics, computer science, biology, chemistry, operations research, and social sciences.

In 1959, the study of probability in graphs gained attention with the works of Erdős and Rényi [2], and Gilbert [3] about random graphs. In 1999, in the beginning of the Information Age, a new family of graphs was studied and gained much importance in the 21st century. This family contains several graphs, such as the WWW graphs, the graphs of coauthors, social graphs, and biological graphs that share unexpected similarities. Examples of characteristics shared by some of those networks include the sparsity (the number of edges are usually linear on the number of vertices), the small-world structure (small distance between vertices and the presence of groups of densely connected vertices), and the power law degree distribution (the number of vertices with degree d is proportional to $d^{-\beta}$, for some exponent $\beta > 0$) [1].

Because of their nontrivial structure, which is neither totally random nor totally regular, these graphs are called complex networks. To study this type of network, several random graph models have been proposed, such as the Barabási–Albert [4] and the Watts–Strogatz models [5], which aim to generate graphs with power law degree distribution and small-world structure, respectively.

We can pose several questions about random graphs. For example, how predictable is the structure of a random graph? If we observe a graph, can we infer which random graph model generated it? How similar/different are two random graphs? To answer these questions, it is fundamental to construct statistical methods in graphs. However, the construction of statistical methods in graphs is not trivial due to the complexity to deal with sets of vertices and edges. To build the link between statistics and graph theory, Takahashi *et al.* [6] proposed the use of the spectrum of the graph (set of eigenvalues of the graph adjacency matrix) to

Mathematical Foundations and Applications of Graph Entropy, First Edition.
Edited by Matthias Dehmer, Frank Emmert-Streib, Zengqiang Chen, Xueliang Li, and Yongtang Shi.

define the spectral graph entropy. That measure of entropy is based on the differential entropy, which extends the classical Shannon measure [7] for continuous variables.

It is important to determine the differences between the graph spectral entropy and other existing graph entropy measures. While the graph spectral entropy is based on the differential entropy, other measures are usually based on the discrete Shannon entropy (e.g., we refer to the review written by Dehmer and Mowshowitz [8]). The discrete graph entropy usually aims to measure the complexity of a given graph. By contrast, the graph spectral entropy measures how uncertain/random is the structure of a random graph [6]. Each graph entropy measure is based on a different graph feature. Examples of features used by the discrete graph entropy measures include vertex centralities, distances between vertices, classes of topologically equivalent vertices, and graph decomposition into special subgraphs of complexity [8]. For random graphs, we are interested in features that are shared by all graphs generated by the same random graph model and that are different among graphs from distinct models. The graph spectrum, which is tightly associated with several graph structural properties, is considered an adequate characterization of a random graph [6]. Based on that feature, the graph spectral entropy was proposed as a measure of the randomness of a random graph.

On the basis of the graph spectral entropy, Takahashi *et al.* [6] introduced formal statistical methods in graphs for parameter estimation, model selection, and a hypothesis test to discriminate two populations of graphs. Applications of these methods were useful to uncover structural properties of the networks in molecular biology [6] and neuroscience [9]. The main contribution of this chapter is a review of statistical methods in graphs based on the graph spectral entropy. In the following sections, we provide an essential background regarding random graphs and then describe the details of the statistical framework in graphs. Monte Carlo simulations are also shown with the purpose of illustrating the performance of the methods.

6.2 Random Graphs

A graph $G = (V, E)$ is an ordered pair, where $V = \{1, 2, \ldots, n\}$ is a set of vertices and E is a set of edges connecting the elements of V. All graphs considered in this chapter are undirected, that is, each edge $e \in E$ is an unordered pair of vertices. To learn statistical methods in graphs, we must first understand the concept of probability distribution over graphs. The theory behind it is the theory of random graphs, which studies the intersection between graph theory and probability theory.

A *random graph* is a probability space $(\Omega, \mathcal{F}, P)$, where the sample space Ω is a nonempty set of graphs, the set of events $\mathcal{F}$ is a collection of subsets of the sample space (usually $\mathcal{F}$ is the power set of Ω), and P is a function that assigns a probability to each event. It is usual to describe a random graph by a sequence of steps to

construct it. An algorithm that describes the construction of a random graph is called a *random graph model*.

An example of random graph model is the Erdős–Rényi algorithm [2], in which the sample space Ω is the set of all graphs having n labeled vertices, and m edges (usually m is a function of n). Each graph of Ω can be generated by selecting m edges from the $\binom{n}{2}$ possible edges. Therefore, the set Ω has size

$$\binom{\binom{n}{2}}{m}.$$

Then, the probability to choose a graph from Ω is

$$\binom{\binom{n}{2}}{m}^{-1}.$$

A similar example of random graph model is the Gilbert model [3], in which two nodes in a graph with n vertices are connected with probability $0 \leq p \leq 1$. Then, the sample space of the Gilbert random graph is the set of all graphs having n labeled vertices. The probability to take a graph with m edges is $p^m(1-p)^{\binom{n}{2}-m}$ [10]. The Erdős–Rényi and Gilbert models are almost interchangeable when $m \sim p\binom{n}{2}$ [11].

Another simple example, in terms of construction, is the random geometric graph, in which n vertices are drawn randomly and uniformly on a unit square and a pair of vertices are connected by an edge if the distance between them are at most a given parameter r [12]. In some applications, the random geometric graph is considered more realistic than the Gilbert and Erdős–Rényi models [12]. However, it is not suited for many real networks, such as biological networks, which usually have a particular type of degree distribution.

A widely used random graph model that aims to construct graphs with a particular property of the degree distribution that is shared by many real networks is the Barabási–Albert algorithm [4]. In that model, we consider a small number (n_0) of initial vertices. Then, in each step, we add a new vertex and connect it to a fixed number ($\leq n_0$) of vertices that already exist in the network. The probability that the new vertex will be connected to a vertex i is proportional to $d(i)^{p_s}$, where $d(i)$ is the degree of i and p_s is a parameter called the scaling exponent parameter. That algorithm generates graphs in which the frequency of vertices with degree d is, asymptotically, proportional to d^{-3} [13]. Because of this power relationship between the frequency and degree, we say that the vertex degrees follow a *power law distribution*.

Another widely used model is the Watts–Strogatz [5] algorithm, which generates graphs with groups of densely connected vertices. The Watts–Strogatz model is as follows:

(1) Construct a ring lattice with n vertices and connect each vertex to the K nearest vertices ($K/2$ in each side of the ring).

(2) Choose a vertex i and the edge e that connects i to its nearest neighbor in a clockwise manner.
(3) With probability p_r, replace the edge e by the edge that connects i to a vertex taken at random according to a uniform distribution over the entire ring.
(4) Repeat the steps 2 and 3 for each vertex in a clockwise manner.
(5) Choose a vertex i and the edge e that connects i to its second nearest neighbor in a clockwise sense, and repeat the steps (2)–(4).
(6) Repeat the process considering the third nearest neighbor and so on, until each edge of the original lattice has been considered.

Our last example is the k-regular model [14], which describes a probability space of all graphs of size n such that each vertex is connected to exactly k other vertices. Therefore, it is a particular type of the Erdős–Rényi random graph. We illustrate graphs generated by the Erdős–Rényi, Gilbert, Geometric, Barabási–Albert, Watts–Strogatz, and k-regular models in Figure 6.1. All these graphs were generated with $n = 500$ vertices and parameters indicated in the figure.

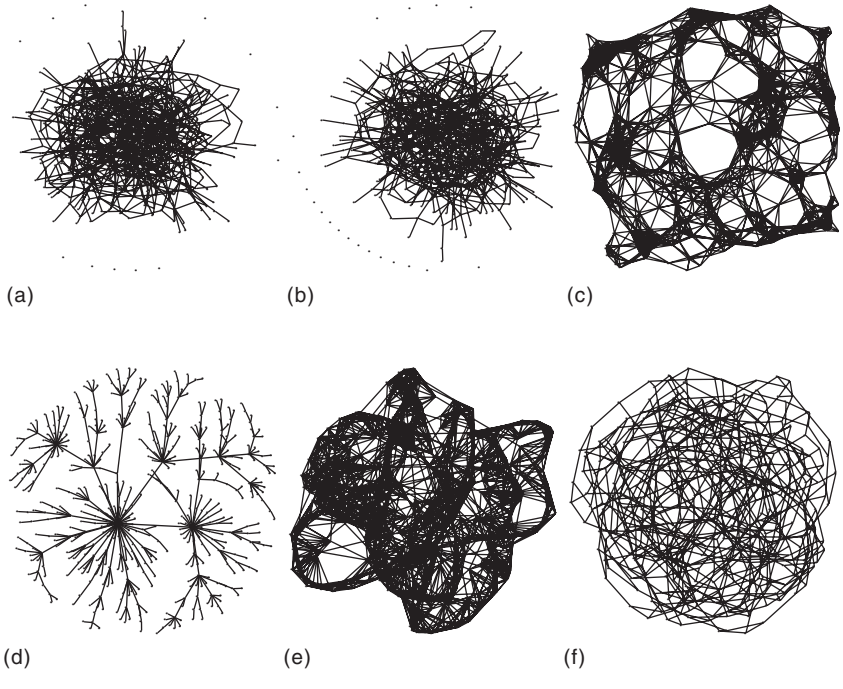

Figure 6.1 *Random graph models*. Graphs with $n = 500$ vertices generated by the Erdős–Rényi (a), Gilbert (b), Geometric (c), Barabási- Albert (d), Watts–Strogatz (e), and k-regular (f) random graph models. In (a), the number of edges is equal to $0.007N$, where $N = \binom{500}{2}$. In (b), the probability p of connecting two vertices is equal to 0.007. In (c), we have $r = 0.1$, where r is the radius used by the geometric model. In (d), we set the scale exponent p_s to 1. In (e), the probability of reconnecting a vertex p_r is equal to 0.07. In (f), the degree of each vertex is $k = 3$. (Takahashi *et al.* [6]. Used under CC BY 4.0 https://creativecommons.org/licenses/by/4.0/.)

Now, suppose that we take at random two graphs G_1 and G_2, each one of size n. If both graphs are from the same random graph, then it is reasonable to expect that in the limit as $n \to \infty$ they share some structural properties. By contrast, if G_1 and G_2 are from different random graphs, we may expect to find fundamental differences between their structural properties.

Given the graphs G_1 and G_2, can we measure the similarities between their structures? Is the probability of G_1 and G_2 being from the same random graph high? To answer these questions, we need a mathematical way to describe graph structural properties that are equal for graphs from the same random graph, but different for elements from distinct random graphs. Takahashi *et al.* [6] proposed that the spectrum of a graph is an adequate summarization of the graph structure for this problem. In the following section, we define the graph spectrum and other spectrum-based concepts that describe a set of graph structural properties.

6.3 Graph Spectrum

Let $G = (V, E)$ be an undirected graph and n the number of vertices (i.e., $n = |V|$). The spectrum of G is the set of eigenvalues of its adjacency matrix, which is denoted by A_G. As G is undirected, if two vertices i and j are connected by an edge, then $A_{Gij} = A_{Gji} = 1$, otherwise, $A_{Gij} = A_{Gji} = 0$ (i.e., if i and j are not connected). We have $A_G = A_G^{\mathrm{T}}$, and, therefore, all eigenvalues of the matrix A_G are real.

Let $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ be the spectrum of G such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The spectral graph theory studies graph spectrum properties and their association with the graph structure. We show some examples of this relationship below.

(1) Let $d(i)$ denote the number of edges in E connected to i (the degree of the vertex i). The eigenvalue λ_1 is at least

$$\frac{1}{n}\sum_{i=1}^{n} d(i),$$

and at most $\max_{i \in V} d(i)$ [15].

(2) The graph G is bipartite only if $\lambda_n = -\lambda_1$ [15].

(3) If G is connected, then the eigenvalue λ_1 is strictly larger than λ_2, and there exists a positive eigenvector of λ_1 [16].

(4) Each vertex in V is connected to exactly λ_1 vertices (i.e., G is λ_1-regular) only if the vector of 1s is an eigenvector of λ_1 [17].

(5) Let $C \subseteq V$ such that each pair of vertices in C are connected in G (i.e., C is a clique in G). Then, the size of C is at most $\lambda_1 + 1$ [18].

(6) Let k be the diameter of G. If G is connected, then A_G has at least $k + 1$ distinct eigenvalues [19].

The facts above illustrate some of the several relationships between the structure of a graph and its spectrum. Now, let us recall that we are particularly interested in random graphs. What if we take the spectrum of a random graph?

Given a set of n labeled vertices $V = \{1, 2, \dots, n\}$, let g be a random graph such that its sample space Ω has graphs on V. We define the *spectrum of* g as a random vector containing n random variables $\lambda_1, \lambda_2, \dots, \lambda_n$. Each function $\lambda_i : \Omega \to \mathbb{R}$ maps a graph in the sample space Ω to the ith largest eigenvalue of its adjacency matrix.

Let δ be the Dirac delta, which is the probability measure satisfying

(1) $\delta(x) = 0, x \in \mathbb{R}^*$,
(2) $\delta(0) = \infty$,
(3) $\int_{-\infty}^{+\infty} \delta(x)dx = 1$.

Let G be a graph in the sample space of g. The *empirical spectral density* of G is defined as [20]

$$\rho(\lambda, G) = \frac{1}{n} \sum_{i=1}^{n} \delta\left(\lambda - \frac{\lambda_i(G)}{\sqrt{n}}\right).$$

It is pertinent to ask whether we can deduce anything about the empirical spectral densities of graphs from g, particularly in the limit $n \to \infty$. Then, it is natural to take the limit of the expectation (denoted by $\langle\cdot\rangle$) of the empirical spectral density according to the probability law of g:

$$\rho(\lambda) = \lim_{n\to\infty} \left\langle \frac{1}{n} \sum_{i=1}^{n} \delta\left(\lambda - \frac{\lambda_i}{\sqrt{n}}\right) \right\rangle.$$

We refer to the expected empirical spectral density defined above as the *spectral density* of g. To illustrate the spectral distribution of different random graphs, we show in Figure 6.2 the empirical spectral density for graphs generated by the Erdős – Rényi, Gilbert, Geometric, Barabási – Albert, and Watts – Strogatz models.

Once we have observed the tight relationship between the spectrum and the graph structure and defined the spectral density, we can introduce measures that describe or compare random graph structural properties. Let g_1 and g_2 be two Gilbert random graphs with n vertices such that the probability p of connecting a pair of vertices is 0.5 in g_1 and 0.9 in g_2. If we take two graphs G_1 and G_2 from g_1 and g_2, respectively, can we deduce anything about their structures? Intuitively, as the probability to connect a pair of vertices is high in g_2 and intermediate in g_1, the structure of G_2 seems to be more predictable than that of G_1.

This notion of predictability of random variable outcomes is studied by the Information Theory, in which for each random variable, we associate an amount of uncertainty. To quantify how well an outcome can be predicted, Shannon introduced the entropy measure [7]. In the following section, we show this concept for a random graph.

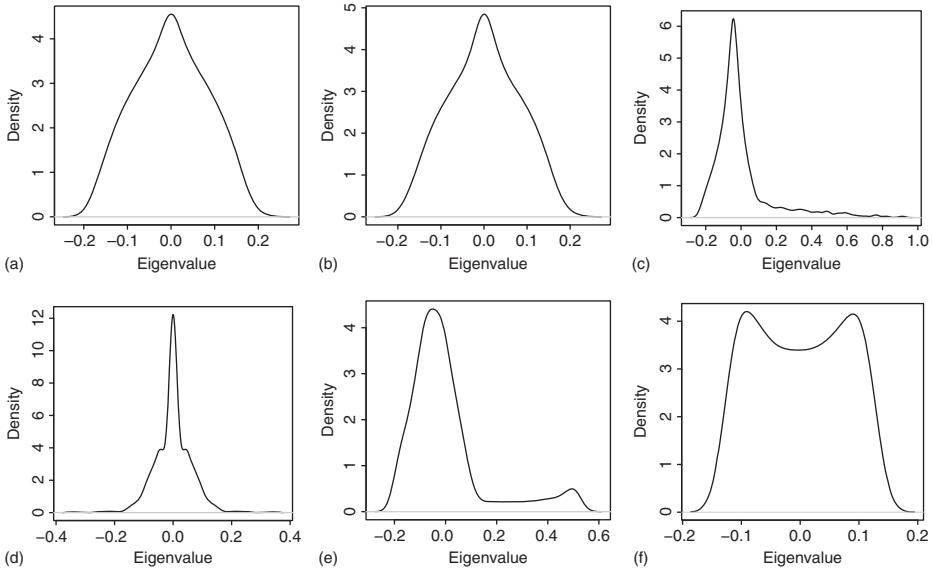

Figure 6.2 *Graph spectral density*. Spectral density estimators for graphs with 500 vertices generated by the Erdős–Rényi (a), Gilbert (b), Geometric (c), Barabási–Albert (d), Watts–Strogatz (e), and k-regular (f) random graph models. In (a), the number of edges is equal to $0.007N = 0.007 \binom{500}{2}$. In (b), the probability p of connecting two vertices is equal to 0.007. In (c), we have $r = 0.1$, where r is the radius used by the geometric model. In (d), we set the scale exponent p_s to 1. In (e), the probability of reconnecting a vertex p_r is equal to 0.07. In (f), the degree of each vertex is $k = 3$.

6.4 Graph Spectral Entropy

The entropy of a random graph quantifies the randomness of its structure. Let g be a random graph and ρ its spectral density. The spectral entropy of g is defined as

$$H(\rho) = -\int_{-\infty}^{+\infty} \rho(\lambda) \log \ \rho(\lambda) d\lambda, \tag{6.1}$$

where $0 \log \ 0 = 0$ [6].

We can approximate the spectral entropy of the Gilbert random graph to

$$H(\rho) \sim \frac{1}{2} \ln \ (4\pi^2 p(1-p)) - \frac{1}{2},$$

where p is the probability to connect a pair of vertices [6]. Then, the maximum spectral entropy of the Gilbert random graph is achieved when $p = 0.50$. This is consistent with the intuitive idea that when all possible outcomes have the same probability to occur, the ability to predict the system is poor. By contrast, when $p \to 0$ or $p \to 1$, the construction of the graph becomes deterministic, and the

amount of uncertainty associated with the graph structure achieves its minimum value.

To approximate the entropy of a random graph, one may use the random graph model to construct graphs, and then estimate the spectral density from that data set. For example, given a random graph model, we construct a set of graphs $\{G_1, G_2, \dots, G_N\}$ with n vertices using the model, and then for each G_j, $1 \leq j \leq N$, we apply a density function estimator. In the examples shown in this chapter, we have considered an estimator based on the Gaussian kernel. It can be interpreted as a smoothed version of a histogram. Given a graph G_j and its spectrum $\{\lambda_1^{(j)}, \lambda_2^{(j)}, \dots, \lambda_n^{(j)}\}$, each eigenvalue λ_i contributes to estimate the function in a point λ according to the difference between λ_i and λ. That contribution is weighted by the kernel (K) function and depends on a parameter known as bandwidth (h), which controls the size of the neighborhood around λ. Formally, the density function estimator in a point λ is

$$\hat{f}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{\lambda - \lambda_i}{h}\right),$$

where

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

For the examples given in this chapter, we have chosen the Silverman's criterion [21] to select the bandwidth h. To obtain an estimator for the random graph, we can apply the procedure described above for each observed graph $G_1, G_2, \dots, G_N$, and then take the average among all the estimators. We illustrate the empirical spectral entropy of the Erdős–Rényi, Gilbert, geometric, Barabási–Albert, Watts–Strogatz, and k-regular random graphs in Figure 6.3 by varying the parameter of each model as indicated in the images. For each model and each parameter, we have generated 50 graphs of size $n = 500$.

The spectral entropy tells us that the uncertainty of the Erdős–Rényi, the geometric, and the Gilbert models are associated with their parameters (m, r, and p, respectively) in a similar way. When the number of edges is $m = 0.5 \binom{n}{2}$, where n is the number of vertices, the entropy of the Erdős–Rényi random graph achieves its maximum value. Conversely, when the graph approximates to a full or empty graph (i.e., when $m \to \binom{n}{2}$ or $m \to 0$, respectively), the entropy achieves its lowest values. The k-regular random graph also achieves a lower entropy when the graph approximates to an empty graph (i.e., when $k \to 0$). If k increases, then the entropy will increase until k achieves an intermediate value.

For the geometric random graph, it is intuitive that when the radius r is close to zero, a graph taken at random will probably have few edges, and when the parameter r is close to $\sqrt{2}$, the graph will probably approximate to a full graph. Therefore, in those scenarios, the amount of uncertainty is low. By contrast, if the radius r is intermediate, then the entropy of the geometric random graph achieves its highest values.

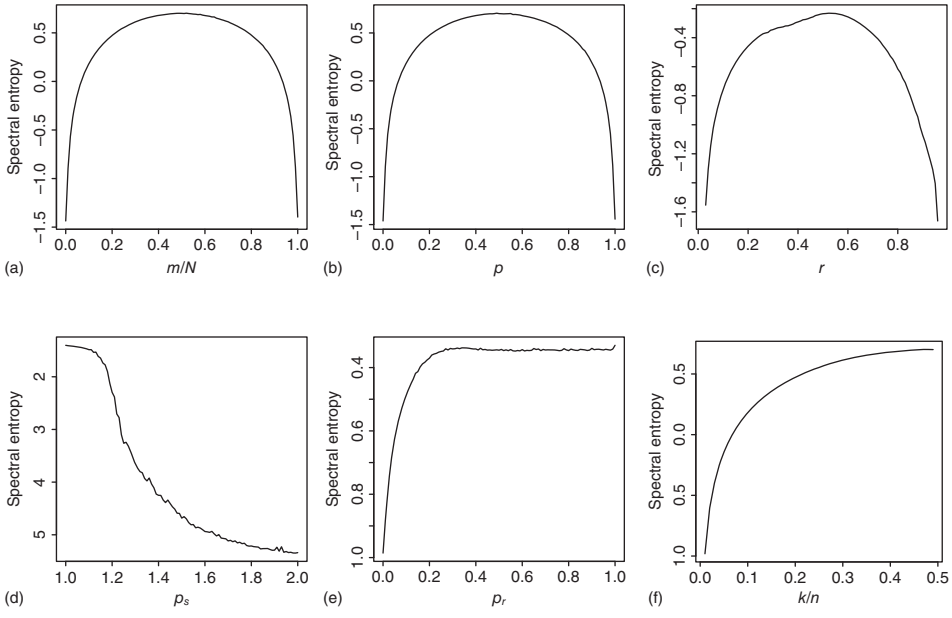

Figure 6.3 *Graph spectral entropy*. The empirical graph spectral entropy (*y*-axis) for the Erdős–Rényi (a), Gilbert (b), Geometric (c), Barabási–Albert (d), Watts–Strogatz (e), and *k*-regular (f) random graph models. In (a), (b), (c), and (e), the values on *x*-axis varies from 0 to 1. In (d), the values vary from 1 to 2. In (f), the values vary from 0 to 0.5. In (b), (c), (d), and (e), the *x*-axis corresponds to the parameters p, r, p_s, and p_r, respectively. In (a), the parameter m is obtained by multiplying the value on the *x*-axis by $N = \binom{n}{2}$. In (f), we multiply the value on the *x*-axis by n to obtain k. For each model, the empirical spectral entropy was obtained from 50 graphs of size $n = 500$.

For the Barabási–Albert and Watts–Strogatz models, the spectral entropy also satisfies the intuitive notion of uncertainty. As illustrated in Figure 6.3d, the Barabási–Albert empirical entropy is inversely proportional to the scale exponent (p_s). When p_s is low, the randomness of the graph construction is high, because the influence of the vertex degrees over the probability to connect two vertices is low. Conversely, when p_s is high, the vertex degrees greatly contribute to the insertion of edges, and, then, the amount of uncertainty is low. Finally, in Figure 6.3e, we note that in the Watts–Strogatz model, the spectral entropy increases with the increasing of the parameter p_r. Remember that the parameter p_r is the probability of replacing the last edge inserted, which connects a vertex i to another vertex that is near to it in the lattice, by an edge that connects the vertex i to another vertex chosen randomly. Thus, when $p_r = 1$, we have a graph constructed in a completely random way. Conversely, when $p_r = 0$, the graph construction is determined by the lattice structure, in which each vertex is connected to the K nearest vertices.

6.5 Kullback–Leibler Divergence

The Kullback–Leibler divergence (KL) measures the amount of information lost when a probability distribution is used to approximate other distribution. For graphs, the KL divergence can be used to discriminate random graphs, select a graph model, and estimate parameters that best describe the observed graph. Clearly, if two spectral densities are different, then the corresponding random graphs are different. However, different random graphs may have the same spectral density.

Let g_1 and g_2 be two random graphs with spectral densities ρ_1 and ρ_2, respectively. The *Kullback–Leiber divergence* is defined as follows. If the support of ρ_2 contains the support of ρ_1, then the KL *divergence* between ρ_1 and ρ_2 is

$$\mathrm{KL}(\rho_1|\rho_2) = \int_{-\infty}^{+\infty} \rho_1(\lambda) \log \frac{\rho_1(\lambda)}{\rho_2(\lambda)} d\lambda,$$

where $0 \log 0 = 0$ and ρ_2 is the *reference measure* [6]. If the support of ρ_2 does not contain the support of ρ_1, then $\mathrm{KL}(\rho_1|\rho_2) = +\infty$.

The KL divergence is nonnegative, and it is zero only if ρ_1 and ρ_2 are equal. Note that in many cases, $\mathrm{KL}(\rho_1|\rho_2)$ and $\mathrm{KL}(\rho_2|\rho_1)$ are different when $\rho_1 \neq \rho_2$, that is, KL is asymmetric.

The asymmetric property of the KL divergence is particularly useful when we want to find the reference measure that best describes the observed spectrum (we explain details about applications of the KL divergence in Section 6.7). However, if the goal is to discriminate graph structural properties between two random graphs, then we usually do not have a reference measure. In this case, it is more adequate to use a symmetric divergence between graphs, such as the Jensen–Shannon divergence, which is described in the next section.

6.6 Jensen–Shannon Divergence

The Jensen–Shannon (JS) divergence is a symmetric alternative to the Kullback–Leibler divergence. The JS *divergence* between two spectral densities ρ_1 and ρ_2 is

$$\mathrm{JS}(\rho_1, \rho_2) = \frac{1}{2}\mathrm{KL}(\rho_1|\rho_M) + \frac{1}{2}\mathrm{KL}(\rho_2|\rho_M),$$

where $\rho_M = \frac{1}{2}(\rho_1 + \rho_2)$ [6].

We can interpret the Jensen–Shannon divergence as a measure of the structural differences between two random graphs. The square root of the Jensen–Shannon is a metric, that is, $\mathrm{JS}(\rho_1, \rho_2) \geq 0$, $\mathrm{JS}(\rho_1, \rho_2) = 0$ only if $\rho_1 = \rho_2$, $\mathrm{JS}(\rho_1, \rho_2) = \mathrm{JS}(\rho_2, \rho_1)$ and $\sqrt{\mathrm{JS}(\rho_1, \rho_3)} \leq \sqrt{\mathrm{JS}(\rho_1, \rho_2)} + \sqrt{\mathrm{JS}(\rho_2, \rho_3)}$ for any spectral density ρ_3.

6.7 Model Selection and Parameter Estimation

Once we have defined random graphs and divergences between random graphs, several questions may be pertinent. Given a random graph g and a random graph model $M(\theta)$ with a parameter θ, can we measure how well $M(\theta)$ describes g? If we take a set of random graph models $S = \{M_1, M_2, \dots, M_P\}$, can we find which of them does best describe g?

As we have discussed in the previous sections, the spectral density describes several structural properties of random graphs. Then, we may consider to address the first question by measuring the dissimilarity between the spectral densities of the random graph g and model $M(\theta)$ (i.e., the spectral density of the random graph generated by the model M with a given parameter θ). To measure that dissimilarity, we can, for example, choose the random graph model spectral density as the reference measure, and then use the KL divergence to measure how much information is lost when the model is used to estimate the random graph spectral density.

For addressing the second question, it is reasonable to select the model that minimizes the dissimilarity between the random graph spectral density and the model spectral density. We can break this approach into two steps. First, for a fixed model, we estimate the parameter and then, given the parameter estimator for each model, finally select a model to describe the random graph.

Formally, we describe the first step as follows. Let ρ_g be the spectral density of the random graph g. Given a random graph model M, let θ be a real vector containing values for each parameter of M. If we consider all possible choices for θ, then the model M generates a parametric family of spectral densities $\{\rho_\theta\}$. Assuming that there exists a value of θ that minimizes $\mathrm{KL}(\rho_g|\rho_\theta)$, which is denoted by θ^*, we have

$$\theta^* = \arg\min_\theta \mathrm{KL}(\rho_g|\rho_\theta).$$

However, in real applications, the spectral density ρ_g is unknown. Therefore, in practice, an estimator $\hat{\rho_g}$ of ρ_g is used, as described in Section 6.4. Then, an estimator $\hat{\theta}$ of θ^* is

$$\hat{\theta} = \arg\min_\theta \mathrm{KL}(\hat{\rho_g}|\rho_\theta). \tag{6.2}$$

The second step for the model selection consists in using Eq. 6.2 to estimate the parameters for each model and then choosing the model that minimizes the KL divergence. Let $\{\rho_{\theta 1}\}, \{\rho_{\theta 2}\}, \dots, \{\rho_{\theta P}\}$ be parametric families of spectral densities, $\hat{\theta}_i$ for $i = 1, 2, \dots, P$ be the estimates of θ_i obtained by Eq. 6.2, and $\#(\theta_i)$ be the dimension of θ_i. Then, the best candidate model j is selected by

$$j = \arg\min_i 2\mathrm{KL}(\hat{\rho_g}|\rho_{\hat{\theta} i}) + 2\#(\hat{\theta}_i). \tag{6.3}$$

Table 6.1 Parameter estimation.

Model	Size			
	20	50	100	500
ER ($m = 0.5N$)	$0.503N \pm 0.013N$	$0.500N \pm 0.002N$	$0.500N \pm 0$	$0.499N \pm 0.003N$
GI ($p = 0.5$)	0.506 ± 0.039	0.501 ± 0.014	0.501 ± 0.008	0.499 ± 0.003
GE ($r = 0.5$)	0.493 ± 0.061	0.506 ± 0.037	0.502 ± 0.022	0.500 ± 0.010
BA ($p_s = 1$)	1.128 ± 0.309	1.044 ± 0.125	1.026 ± 0.047	1.020 ± 0.025
WS ($p_r = 0.07$)	0.129 ± 0.155	0.069 ± 0.011	0.071 ± 0.008	0.070 ± 0.003
KR ($k = 0.25n$)	$0.264n \pm 0.013n$	$0.245n \pm 0.005n$	$0.250n \pm 0$	$0.249n \pm 0.004n$

Average and standard deviations of the parameters estimated by the Kullback–Leibler divergence for the Erdős–Rényi (ER), Gilbert (GI), Geometric (GE), Barabási–Albert (BA), Watts–Strogatz (WS), and k-regular (KR) models. For each random graph model, we have applied the parameter estimation to 1000 graphs of sizes 20, 50, 100, and 500. The reference measure of a model was estimated by randomly generating 100 graphs from the model and then by taking the average spectral density estimator. The true parameters used to generate the graphs are shown with parentheses (the number of edges m, the probability of connecting two vertices p, the radius r, the scale exponent p_s, the probability of reconnecting a vertex p_r, and the degree k). In general, the parameter m is a function of $N = \binom{n}{2}$ and the parameter k is a function of n, where n is the number of vertices.

Equation 6.3 is based on the Akaike information criterion (AIC), which includes, in addition to the KL divergence, a penalty for the number of estimated parameters. That penalty aims to avoid overfitting, once adjusting the model is easier when the number of parameters is high. In particular, all models considered in this chapter have the same number of parameters, and therefore, we can select the model by minimizing the KL divergence without penalty.

To illustrate the performance of the parameter estimation and model selection approaches, we have generated 1000 graphs of sizes $n = 20, 50, 100, 500$, using the Edős–Rényi (ER), Gilbert (GI), Geometric (GE), Barabási–Albert (BA), Watts–Strogatz (WS), and k-regular (KR) random graph models with parameters $m = 0.5N$, $p = 0.5$, $r = 0.5$, $p_s = 1$, $p_r = 0.07$, and $k = 0.25n$. In Table 6.1, we can see that the average estimator is very close to the true parameter for all models when the graph is sufficiently large.

The results of our illustrative model selection experiments are shown in Table 6.2. We can observe that when the true model is GE, BA, and WS, the number of right choices increases with the size of the graph. As expected, the ER graphs approximate to GI graphs asymptotically, and therefore, when the true model is ER or GI, the model selection approach cannot discriminate them. Since the k-regular graph is a particular type of ER graph, it is natural that some graphs generated by the ER/GI model will approximate to a regular graph. As we can see in Table 6.2 about 14% of the ER and GI graphs have been classified as KR. Similarly, some large KR graphs are classified as ER/GI.

Table 6.2 Model selection simulation.

True model		Predicted model					
	n	ER	GI	GE	BA	WS	KR
ER	20	835	36	0	0	0	129
	50	564	282	0	0	0	154
	100	841	28	0	0	0	131
	500	378	480	0	0	0	142
GI	20	762	135	0	0	0	103
	50	725	213	0	0	0	62
	100	587	337	0	0	0	76
	500	346	510	0	0	0	144
GE	20	0	0	1000	0	0	0
	20	0	0	1000	0	0	0
	100	0	0	1000	0	0	0
	500	0	0	1000	0	0	0
BA	20	571	134	197	98	0	0
	50	313	70	2	615	0	0
	100	6	40	0	954	0	0
	500	0	0	0	1000	0	0
WS	20	75	2	0	0	923	0
	50	0	0	0	0	1000	0
	100	0	0	0	0	1000	0
	500	0	0	0	0	1000	0
KR	20	7	0	0	0	0	993
	50	0	0	0	0	0	1000
	100	0	0	0	0	0	1000
	500	56	72	0	0	0	872

The rows show the true models used to generate the graphs. Each position shows the number of graphs that were classified as the corresponding column. The model selection was performed on 1000 graphs of sizes $n = 20, 50, 100, 500$ for the Erdős–Rényi (ER), Gilbert (GI), Geometric (GE), Barabási–Albert (BA), Watts–Strogatz (WS), and the k-regular (KR) random graphs. To estimate the spectral density of each random graph, we have taken the average among 100 graphs generated by the model.

6.8 Hypothesis Test between Graph Collections

Let T_1 and T_2 be two collections of graphs. Assuming that all graphs from T_1 were generated by the same random graph model M_1 with parameter θ_1 and the graphs from T_2 were generated by a model M_2 with parameter θ_2, could we check if $M_1 = M_2$ and $\theta_1 = \theta_2$?

To answer this question, a natural approach is to measure the dissimilarity between the two graph collections. As we have shown in the previous section, comparing spectral densities is a reasonable approach for model selection and parameter estimation. However, differently from the previous problem, in the

hypothesis test between graph collections, it is not clear which distribution is the reference measure. Therefore, as the Kullback–Leibler divergence is an asymmetric measure, it is not suited for the hypothesis test.

Remember that the Jensen–Shannon divergence is the symmetric version of the Kullback–Leibler divergence and then it is a natural candidate for the statistic of the hypothesis test between collections of graphs, which we describe formally as follows. Let g_1 and g_2 be two random graphs with spectral densities ρ_1 and ρ_2, respectively. We want to test the following hypotheses:

$$H_0 : \mathrm{JS}(\rho_1\ \rho_2) = 0$$
$$H_1 : \mathrm{JS}(\rho_1\ \rho_2) > 0$$

Given a sample from g_1, denoted by T_1, and a sample from g_2, denoted by T_2, the test statistic is $\mathrm{JS}(\hat{\rho_1}, \hat{\rho_2})$, where $\hat{\rho}_1$ and $\hat{\rho}_2$ are the average spectral density estimates obtained from T_1 and T_2, respectively.

After choosing the test statistic and the null and alternative hypotheses, our goal is to obtain the p-value, which is the probability that the test statistic will be at least as extreme as the value observed in the data, assuming that the null hypothesis is true. Thus, to obtain a p-value for the hypothesis test, we need to obtain the statistic distribution under the null hypothesis. This is usually done from an asymptotic distribution or from random resamplings of the data. In the first case, an analytic formula, which is unknown for many statistics and usually depends on assumptions about the population distribution that cannot be verified, is necessary. In the second case, the random resamplings are usually constructed by the Monte Carlo method, which is based on pseudorandom numbers generated by the computer. In general, for that approach, we assume only that the data are independent and identically distributed.

Takahashi *et al.* [6] proposed a test for the Jensen–Shannon divergence between spectral densities that relies on random resamplings of the data with replacement. That approach is known as the bootstrap procedure. The bootstrap was proposed by Efron [22] to estimate the sample distribution of a statistic from random resamplings of the data. This approach is used in several applications, such as standard error estimation for a statistic, confidence intervals for population parameters, and hypothesis tests.

The idea behind the bootstrap is that the observed sample is usually the best population approximation. When the data are independent and identically distributed, the bootstrap approach can be implemented by resampling with replacement of the original data set.

Let n_1 and n_2 be the number of graphs in T_1 and T_2, respectively; B the number of desired bootstrap replications; and $T = T_1 \cup T_2$. The bootstrap procedure for the hypothesis test between T_1 and T_2 is described as follows:

(1) Calculate $\mathrm{JS}(\hat{\rho_1}, \hat{\rho_2})$.
(2) Resample with replacement n_1 graphs from T and construct a new set (bootstrap sample) $\tilde{T}_1$. Obtain the $\tilde{T}_1$ average spectral density estimator, which is denoted by $\hat{\rho}_{\tilde{T}_1}$.

(3) Resample with replacement n_2 graphs from T and construct a new set (bootstrap sample) $\tilde{T}_2$. Obtain the $\tilde{T}_2$ average spectral density estimator, which is denoted by $\hat{\rho}_{\tilde{T}_2}$.
(4) Calculate $\mathrm{JS}(\hat{\rho}_{\tilde{T}_1}, \hat{\rho}_{\tilde{T}_2})$.
(5) Repeat steps 2–4 for B times.
(6) The p-value is the proportion of bootstrap replications such that $\mathrm{JS}(\hat{\rho}_{\tilde{T}_1}, \hat{\rho}_{\tilde{T}_2}) \geq \mathrm{JS}(\hat{\rho}_1, \hat{\rho}_2)$.

Note that $\tilde{T}_1$ and $\tilde{T}_2$ are constructed by taking at random, with replacement, graphs from the set T, which has elements from both T_1 and T_2. Therefore, the new data set containing $\tilde{T}_1$ and $\tilde{T}_2$ simulates the null hypothesis that the graphs in both sets are from the same population.

To show that the bootstrap procedure for the Jensen–Shannon divergence rejects/accepts the null hypothesis as expected, we have generated, for each scenario, 1000 data sets of graphs using the Erdős–Rényi (ER), Gilbert (GI), Geometric (GE), Barabási–Albert (BA), Watts–Strogatz (WS), and k-regular (KR) random graph models. To evaluate the proportion of rejections under the null hypothesis that both collections of graphs being tested are from the same population, we have constructed each data set by generating, with the same random procedure, two collections of 50 graphs. For the ER, GI, GE, BA, WS, and KR models, we have used, respectively, the parameters $m = 0.5N$, $p = 0.5$, $r = 0.1$, $p_s = 0.5$, $p_r = 0.07$, $k = 0.05n$, where $n = 100$ is the number of vertices, and $N = \binom{100}{2}$.

We have also generated data sets under the alternative hypothesis (H_1 : the collections being tested were generated by different random processes) to illustrate the empirical statistical power of the Jensen–Shannon test. In our experiment under the alternative hypothesis, each data set has two collections of 50 graphs, all of them generated by the same random graph model, but using different parameter values for each collection. The graphs from the ER, GI, GE, BA, WS, and KR models were constructed using, respectively, the parameters $m = 0.4N$ versus $0.41N$, $p = 0.4$ versus 0.41, $r = 0.25$ versus 0.26, $p_s = 0.5$ versus 0.6, $p_r = 0.2$ versus 0.21, $k = 0.05n$ versus $0.06n$, where $n = 100$ is the number of vertices, and $N = \binom{100}{2}$.

For both experiments (under the null and alternative hypotheses), we have evaluated the results by constructing an ROC curve described as follows. The ROC curve is constructed over a two-dimensional plot, in which the x-axis corresponds to the significance level α of the test and the y-axis corresponds to the proportion of rejected null hypotheses. Then, the y-axis represents the empirical statistical power of the test. If we reject the null hypothesis randomly, we expect that the proportion of rejected null hypotheses will be equal/close to the significance level, and then the ROC curve will lie on the diagonal. Conversely, if we reject the null hypothesis with high power, then the ROC curve will be drawn above the diagonal.

Under the null hypothesis, we want the proportion of rejected null hypothesis (false positive rate) to be controlled by the significance level α. We then expect that the ROC curve will lie on the diagonal. In Figure 6.4, we show the ROC curves for the experiments under the null hypothesis. The dashed and continuous lines

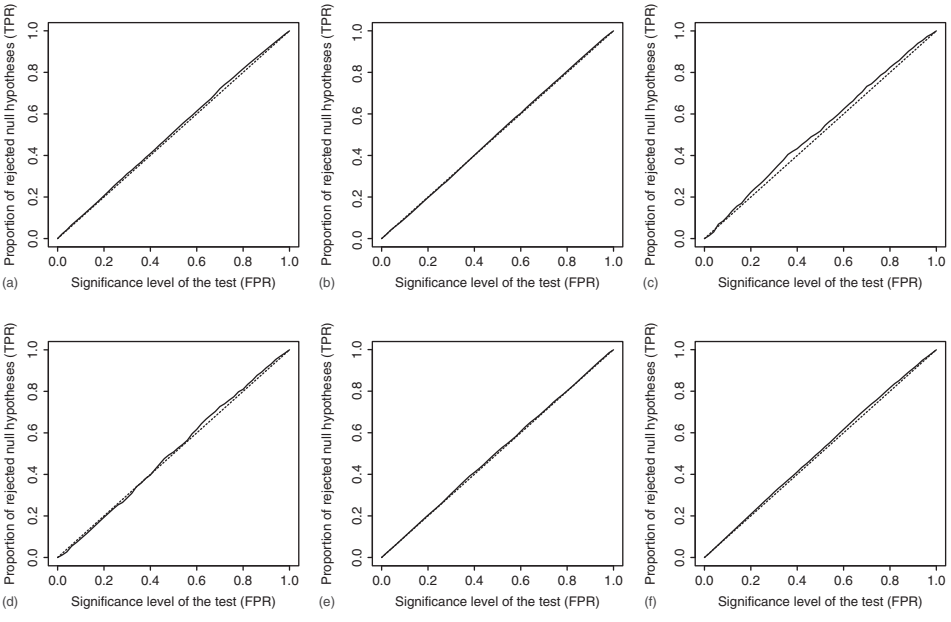

Figure 6.4 *ROC curves under H_0*. The dashed lines show the expected outcome. The continuous lines indicate the observed ROC curve (significance level vs. proportion of rejected null hypothesis) for the Jensen–Shannon test between two collections of 50 graphs of size $n = 100$ generated by the same random process (null hypothesis). In each scenario, the Jensen–Shannon test was applied to 1000 data sets generated by the Erdős–Rényi (a), Gilbert (b), Geometric (c), Barabási–Albert (d), Watts–Strogatz (e), and k-regular random graph models. For each model, all graphs were constructed with the same parameters. In (a), the number of edges is equal to $0.5N = 0.5\binom{100}{2}$. In (b), the probability of connecting two vertices p is equal to 0.5. In (c), we have $r = 0.1$, where r is the radius used by the geometric model. In (d), we set the scale exponent p_s to 0.5. In (e), the probability of reconnecting a vertex p_r is equal to 0.07. In (f), the vertex degrees are equal to $k = 0.05\ n = 5$.

show, respectively, the expected and observed results. We can then observe that the ROC curves approximate the expected results for all graph models.

Under the alternative hypothesis, it is desirable that the ROC curve is as far as possible from the diagonal. We can see in Figure 6.5 that the ROC curve is above the diagonal for all models. In particular, the empirical power of the ER, GI, WS, and KR models achieve its maximum value for all significance levels.

6.9 Final Considerations

In the previous sections, we have explained some concepts behind statistical methods in graphs. Then, it is pertinent to ask which types of real problems

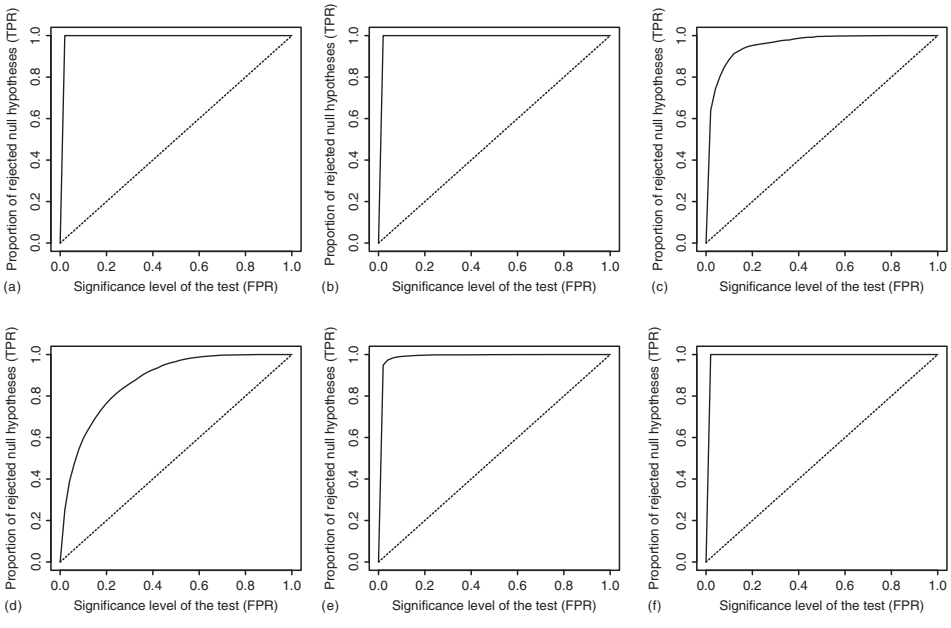

Figure 6.5 *ROC curves under* H_1. The dashed lines show the poorest possible outcome. The continuous lines indicate the observed ROC curve (significance level vs. proportion of rejected null hypothesis) for the Jensen–Shannon test between two collections of 50 graphs of size $n = 100$ generated by different random processes (alternative hypothesis). In each scenario, the Jensen–Shannon test was applied to 1000 data sets generated by the Erdős–Rényi (a), Gilbert (b), Geometric (c), Barabási–Albert (d), and Watts–Strogatz (e) random graph models. For each model, two different parameters are used to generate the two collections of graphs. In (a), the number of edges is equal to $0.4N$ and $0.41N$, where $N = \binom{100}{2}$. In (b), the probability of connecting two vertices p is equal to 0.4 and 0.41. In (c), we have $r = 0.25$ for one collection and $r = 0.26$ for the other, where r is the radius used by the geometric model. In (d), we set the scale exponent p_s to 0.5 and 0.6. In (e), the probability of reconnecting a vertex p_r is equal to 0.2 and 0.21. In (f), the vertex degrees are equal to $k = 0.05$ $n = 5$ and $k = 0.06$ $n = 6$.

can be addressed using those methods. In the following sections, we illustrate interesting applications in biological problems.

6.9.1 Model Selection for Protein–Protein Networks

In Section 6.7, we have described approaches for estimating parameters and selecting random graph models. Takahashi *et al.* [6] have applied the model selection approach based on the Kullback–Leibler divergence to protein–protein networks of eight species. In accordance with the literature, all graphs were classified as scale-free.

6.9.2
Hypothesis Test between the Spectral Densities of Functional Brain Networks

In addition to the model selection approach, Takahashi *et al.* [6] applied the method described in Section 6.8, which is based on the Jensen–Shannon divergence between the graph spectral densities, to test if two collections of functional brain networks have similar structures. The functional brain networks were inferred from fMRI (functional magnetic resonance imaging) data of two groups of people: the first containing 479 individuals with typical development and the second containing 159 individuals diagnosed with ADHD (attention-deficit hyperactivity disorder). The Jensen–Shannon test was then performed between the groups of patients with typical development and those diagnosed with ADHD. The resulting p-value was lower than 0.05, suggesting differences between the functional brain networks of typical development and ADHD individuals.

6.9.3
Entropy of Brain Networks

As we discussed in Section 6.4, the graph spectral entropy can be interpreted as a measure of the amount of uncertainty associated with the graph structure. Given a graph that represents the functional connectivity of the brain, can the entropy reveal anything about the brain patterns of connectivity?

Sato *et al.* [9] measured the entropy of the functional brain network in the data set described in Section 6.9.2 containing individuals with typical development and individuals diagnosed with ADHD. After inferring the functional brain connections, the authors clustered the regions of the brains into four groups, each one representing a functional connectivity (subnetwork) of the brain.

For each cluster, Sato *et al.* [9] have tested if the entropy of the corresponding subgraph is different between individuals with typical development and individuals with ADHD. The results were consistent with the literature, identifying brain regions associated with ADHD. Furthermore, in that study, the entropy measure has identified clusters that other commonly used measures of complex networks (e.g., average degree, average clustering coefficient, and average shortest path length) have not. Thus, the results suggested that the entropy could reveal abnormalities in the functional connectivity of the brain.

6.10
Conclusions

The development of statistical methods in graphs is crucial to better understand the mechanisms underlying biological systems such as functional brain networks and gene regulatory networks. This chapter has shown a statistical framework composed of methods to (i) measure the network entropy; (ii) estimate the parameters of the graphs; (iii) model selection; and (iv) test whether two populations of

graphs are generated by the same random graph model. Experiments were conducted using Monte Carlo simulation data to illustrate the strengths of the statistical approaches. These approaches are flexible and allow generalizations to other families of graphs that are not limited to the ones illustrated in this chapter.

6.11 Acknowledgments

Suzana de Siqueira Santos was supported by grants from FAPESP (2012/25417-9; 2014/09576-5). Daniel Yasumasa Takahashi was supported by grants from FAPESP (2014/09576-5; 2013/07699-0) and CNPq (473063/2013-1). João Ricardo Sato was supported by grants from FAPESP (2013/10498-6; 2014/09576-5). Carlos Eduardo Ferreira was supported by a grant from FAPESP (2013/07699-0). André Fujita was supported by grants from FAPESP (2014/09576-5; 2013/03447-6), CNPq (304020/2013-3; 473063/2013-1), NAP-eScience – PRP – USP, and CAPES.

References

1. Chung, F. and Lu, L. (2006) *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics)*, American Mathematical Society, Boston, MA.
2. Erdös, P. and Rényi, A. (1959) On random graphs. *Publ. Math. Debrecen*, **6**, 290 – 297.
3. Gilbert, E.N. (1959) Random graphs. *Ann. Math. Stat.*, **30** (4), 1141 – 1144, doi: 10.1214/aoms/1177706098.
4. Barabási, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286** (5439), 509 – 512, doi: 10.1126/science.286.5439.509.
5. Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393** (6684), 440 – 442, doi: 10.1038/30918.
6. Takahashi, D.Y., Sato, J.R., Ferreira, C.E., and Fujita, A. (2012) Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PLoS ONE*, **7** (12), e49 949, doi: 10.1371/journal.pone.0049949.
7. Shannon, C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27** (3), 379 – 423, doi: 10.1002/j.1538-7305.1948.tb01338.x.
8. Dehmer, M. and Mowshowitz, A. (2011) A history of graph entropy measures. *Inf. Sci.*, **181** (1), 57 – 78, doi: 10.1016/j.ins.2010.08.041.
9. Sato, J.R., Takahashi, D.Y., Hoexter, M.Q., Massirer, K.B., and Fujita, A. (2013) Measuring network's entropy in ADHD: a new approach to investigate neuropsychiatric disorders. *NeuroImage*, **77**, 44 – 51, doi: 10.1016/j.neuroimage.2013.03.035.
10. Bollobás, B. and Chung, F.R.K. (1991) *Probabilistic Combinatorics and its Applications*, American Mathematical Society.
11. Bollobás, B. and Riordan, O.M. (2002) Mathematical results on scale-free random graphs, in *Handbook of Graphs and Networks* (eds S. Bornholdt and H.G. Schuster), Wiley-VCH Verlag GmbH & Co. KGaA, pp. 1 – 34.
12. Penrose, M. (2003) Random geometric graphs, *Oxford Studies in Probability*, vol. **5**, Oxford University Press.
13. Bollobás, B., Riordan, O., Spencer, J., and Tusnády, G. (2001) The degree sequence of a scale-free random graph process. *Random Struct. Algorithms*, **18** (3), 279 – 290, doi: 10.1002/rsa.1009.

14. Bollobás, B. (2001) *Random Graphs*, Cambridge University Press.
15. Cvetković, D.M., Doob, M., and Sachs, H. (1980) *Spectra of Graphs: Theory and Application*, Academic Press.
16. Bapat, R.B. and Raghavan, T.E.S. (1997) *Nonnegative Matrices and Applications*, Cambridge University Press, Cambridge.
17. Spielman, D. (2012) Spectral graph theory, in *Combinatorial Scientific Computing*, 1st edn (eds U. Naumann and O. Schenk), Chapman and Hall/CRC, Boca Raton, FL, pp. 495–517.
18. Wilf, H.S. (1967) The Eigenvalues of a Graph and Its Chromatic Number. *J. London Math. Soc.*, **42** (1), 330–332.
19. Brouwer, A.E. and Haemers, W.H. (2011) *Spectra of Graphs*, Springer Science & Business Media.
20. Rogers, T. (2010) New results on the spectral density of random matrices. PhD thesis, King's College London.
21. Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, Boca Raton, FL.
22. Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7** (1), 1–26, doi: 10.1214/aos/1176344552.