Spectrum-Based Statistical Methods for Directed Graphs with Applications in Biological Data

Victor Chavauty Villela, Eduardo Silva Lira, and André Fujita(✉)

Instituto de Matemática e Estatística, Universidade de São Paulo (USP), São Paulo, Brazil
andrefujita@usp.br

Abstract. Graphs often model complex phenomena in diverse fields, such as social networks, connectivity among brain regions, or protein-protein interactions. However, standard computational methods are insufficient for empirical network analysis due to randomness. Thus, a natural solution would be the use of statistical approaches. A recent paper by Takahashi et al. suggested that the graph spectrum is a good fingerprint of the graph's structure. They developed several statistical methods based on this feature. These methods, however, rely on the distribution of the eigenvalues of the graph being real-valued, which is false when graphs are directed. In this paper, we extend their results to directed graphs by analyzing the distribution of complex eigenvalues instead. We show the strength of our methods by performing simulations on artificially generated groups of graphs and finally show a proof of concept using concrete biological data obtained by Project Tycho.

Keywords: Network Correlation · Graph Statistics · ECoG

1 Introduction

We often use graphs to model interactions between objects. Some examples include the functional connectivity of brain regions [4], social interactions [18], molecular interactions [2], and gene regulations [1].

Once we model these natural phenomena using graphs, it becomes of significant interest to discriminate graphs of two or more populations or make inferences [10]. For instance, suppose three patient groups were assigned different treatments for a neurochemical condition. By examining each patient's resting state magnetic resonance imaging (MRI) scans, can we discern whether there is a notable distinction among the administered drugs?

Traditional computation methods rely on the search for an isomorphism between graphs or sub-graphs, which are prone to failure when randomness is

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

M. S. Reis and R. C. de Melo-Minardi (Eds.): BSB 2023, LNBI 13954, pp. 46–57, 2023.
https://doi.org/10.1007/978-3-031-42715-2_5

applied to the graphs [9]. Because of this nature, these methods are unfit for usage in biological data, where intrinsic randomness is expected [10].

An alternative technique is to compare graph features, such as the number of nodes and edges, and, particularly, centrality measures, such as closeness and betweenness [8]. These centrality measures are estimated and then used as input in standard statistical methods. Although this is a step up from the previous techniques, centrality measures can under-represent variability between graphs. Take, for example, two graphs obtained from the Watts-Strogatz model. Even if distinct rewiring probabilities are used, they still present the same centrality measure since the number of edges does not change [10].

In 2017, Takahashi *et al.* [12] proposed that the graph spectrum is a good feature for describing the graph structure. They used the Kullback-Leibler and Jensen-Shannon divergences between spectral distributions to measure the distance between graphs. Using this concept, they constructed tools for 1. model selection; 2. a parameter estimator for random graph models; 3. a statistical test to compare two sets of graphs. More recently, these ideas have been used to create a concept of correlation [11]/causality [15] between graphs and spectrum-based clustering algorithms for complex networks [14].

One limitation of this work is that it is limited to undirected graphs whose eigenvalues are all real-valued. However, many empirical graphs are directed. A solution would be to symmetrize the graph. The problem is that we usually lose the directionality information, or it vastly influences the spectrum distribution.

In this paper, we extend the results of Takahashi et al. for directed graphs. Our ANOVA-like approach can distinguish between groups of directed graphs obtained from distinct populations. Also, we apply it to actual biological data for illustration.

2 Materials

2.1 Graphs

A graph G consists of a pair (N, E), where N is a set of nodes, and E is a set of edges connecting a pair of nodes of G.

We call a graph weighted if every edge between two nodes i and j is associated with a complex value $e_{i,j} \in \mathbb{C}$. In contrast, in non-weighted graphs, an edge between two nodes i and j will assume 1 if i and j are connected or 0 otherwise.

A graph is said to be undirected if, for every pair of nodes i and j, the edges $e_{i,j}$ and $e_{j,i}$ connecting i to j, and j to i respectively, are equal. That is: $e_{i,j} = e_{j,i}$. Otherwise, it is undirected.

Given a graph G with n nodes, we define its adjacency matrix $\mathbf{A}_G$ as the matrix $\mathbf{A}_G = (e_{i,j})_{i,j=1,\ldots,n}$, where $e_{i,j}$ is the value associated with the edge connecting node i and node j. Note that the adjacency matrix of an undirected graph is symmetric.

The spectrum of a graph G is the set of eigenvalues of its adjacency matrix $\mathbf{A}_G$. If G is directed, its adjacency matrix is non-symmetrical. Therefore its eigenvalues are complex-valued.

2.2 Spectral Distribution

A random graph g is a family of graphs whose members are generated by some probability law. For example, we construct an Erdös-Rényi random graph by connecting two nodes with probability p.

We define the complex Dirac delta as the measure $\delta_{\mathbb{C}}$ satisfying for every compactly supported continuous function f:

$$\int_{\mathbb{C}} f(x)\delta_{\mathbb{C}}\{dx\} = f(0).$$

Alternatively, we construct the complex Dirac delta function as the product of the 1-dimensional Dirac delta in two variables (the real and the imaginary variables). That is:

$$\delta_{\mathbb{C}}(a + bi) = \delta(a)\delta(b).$$

Let g be a directed random graph generated by some probability law. Then its complex eigenvalues Δ form random vectors. Let brackets $\langle\rangle$ indicate expectations concerning the probability law. Then we define the spectral distribution of the directed random graph g as

$$\rho_g(\lambda) = \lim_{n\to\infty} \langle \frac{1}{n} \sum_{j=1}^{n} \delta_{\mathbb{C}}(\lambda - \frac{\lambda_j}{\sqrt{n}}) \rangle.$$

This distribution is highly correlated with distinct features of the graph. We can use it as a fingerprint of the random graph g [10].

2.3 Calculating the Graph Spectrum

Estimating the spectral distribution of a directed random graph is performed under a similar procedure as for the undirected case [10].

Since the spectral density ρ_g is unknown, we need an estimator $\hat{\rho}_g$. We initially compute the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the graph's adjacency matrix g and apply a multivariate kernel regression [6]. We divide the resulting 2-dimensional surface by the volume under the curve to ensure the final volume is one (probability function).

2.4 Statistical Distance

The spectrum distribution is the distribution of complex eigenvalues of a graph model. We are interested in using it as a fingerprint of the model so that by comparing the spectrum of two different random graph models, we can establish a certain distance between them. Similarly, we can compare the spectrum of a graph to the spectrum distribution of a random graph model and obtain a measure of how far apart the graph is from being generated from that specific model.

To compare these distributions, we will be using the Kullbeck-Leibler divergence [13]. The Kullbeck-Leibler divergence is a statistical distance measuring how a probability distribution differs from a second distribution. For two probability densities p and q, the Kullbeck-Leibler divergence is defined as

$$D(p,q) = \int_{\mathbb{C}} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

2.5 Random Graph Models

Graphs can often model very complex phenomena, and it is often impossible to establish how a graph was formed when dealing with biological data. Besides, it is difficult to establish whether two graphs are similar simply by analyzing their structures. Thus, one idea is to imagine these graphs resulting from a probabilistic model with a set of parameters.

Directed Models. Unfortunately, models for directed graphs are not as prevalent as the ones for undirected graphs. Thus, we propose the following general extension of any directed model.

Given a random model r with a parameter p, we extend this model as follows. Let p_1 and p_2 be two parameters for model r. Then

1. Generate a graph G_1 with parameter p_1 and construct its adjacency matrix.
2. Generate a graph G_1 with parameter p_2 and construct its adjacency matrix.
3. Generate a matrix M whose upper triangular is the same as of G_1 and whose lower triangular is the same as of G_2.
4. Generate a graph G with adjacency matrix M.

Note that the parameters p_1 and p_2 control the network's inner and external connections, respectively, which are represented on the upper and lower triangle of the graph's adjacency matrix. In the scenario in which $p_1 = p_2$, the resulting graph is still directed due to the random element of the graph generation process.

We will use this procedure to run our simulations.

3 Methods

Given k groups of graph samples, we are now interested in verifying whether they originated from the same population.

Naively, we could use a parametric approach by selecting a random graph model, estimating the parameters for each graph, and using traditional ANOVA with the estimated parameters as input. However, we must know the random graph model, which is very unlikely in most realistic scenarios. Other non-parametric methods, like the Kolmogorov-Smirnov test, require independence of the graphs, which is often not true when they result from a biological process. Therefore, we will use an ANOVA-like approach following the ideas described by Fujita *et al.* [9] called ANOGVA.

In other words, we will perform a variation of the ANOVA using the complex distribution of eigenvalues of the graphs.

Let $g_1, \ldots, g_k$ be k distinct graph populations. If these graphs come from the same population, their spectral distributions should be equal. Let ρ_{g_i} be the average graph spectrum for group i, $\rho_G = \frac{1}{k}\sum_{i=1}^{k} \rho_{g_i}$ be the overall graph spectrum average, and D be the Kullbeck-Leibler divergence.

Then, we test the following hypothesis:

$$H_0 : D(\rho_{g_1}, \rho_G) = D(\rho_{g_2}, \rho_G) = \ldots = D(\rho_{g_k}, \rho_G) = 0$$

H_1: At least one of the groups of graphs was generated in a different manner

Under the null hypothesis, we expect the statistic $\Delta = \sum_{i=1}^{k} D(\rho_{g_1}, \rho_G)$ to be small. Under the alternative hypothesis, we expect it to be large.

The distribution of Δ is unknown and highly dependent on the used random graph model. Therefore, to test for significance, we will use a bootstrap approach.

The following algorithm describes how we compute the bootstrap

Input: k groups of graphs, $g_1, \ldots, g_k$, and a number of max-iterations Max

Output: A p-value

Estimate $\hat{\rho}_{g_1}$ and $\hat{\rho}_G$;
Calculate $\hat{\Delta} = \sum_{i=1}^{k} D(\hat{\rho}_{g_1}, \hat{\rho}_G)$;
Set $\hat{\Delta}_l = []$;
for Max *iterations* **do**
    Construct k new groups $g_1^{'}, \ldots, g_k^{'}$ by resampling (without replacement) the original graph set;
    Estimate the average spectrum distribution $\hat{\rho}_{g_i^{'}}$ for each new graph $g_i^{'}$;
    Calculate the overall graph spectrum average $\hat{\rho}_G^{'}$;
    Calculate $\hat{\Delta}^{'} = \sum_{i=1}^{k} D(\hat{\rho}_{g_1}^{'}, \hat{\rho}_G^{'})$.;
    Append $\hat{\Delta}^{'}$ to $\hat{\Delta}_l$;
end
Let $p =$ **Cardinality**$(\hat{\Delta}^{'} \in \hat{\Delta}_l$: such that $\hat{\Delta}^{'} \geq \hat{\Delta}) \cdot \frac{1}{Max}$;
return p;

Algorithm 1: Anogva

Implementation. We implemented this method in R, extending the existing StatGraph package [17]. We constructed the multivariate kernel density estimator using the package 'ks.'

4 Simulations

To verify the power of the method described in this paper, we constructed a set of simulations to generate directed random graphs as defined in Sect. 2.5.

To evaluate the performance of ANOGVA, we need to verify the null (H_0) and alternative (H_1) hypotheses. Since we want to ensure that ANOGVA works in a wide range of random graph models, we generated the graphs using the Erdös-Rényi [7], Watts-Strogatz [19], and Barabási-Albert [3] models, as described in Sect. 2.4. We generated the graphs using the igraph package in R [5].

For each of the models, we performed the following simulation:

1. We generated three sets of graphs: G_1, G_2, and G_2, each containing ten graphs, for a total of 30 graphs.
2. All of the graphs were generated with $n = 800$ nodes and using specific parameters.
3. We then applied the ANOGVA algorithm using 500 bootstrap samples.
4. We ran this experiment 500 times, generating a p-value distribution.

Simulation (H_0)**:** All three groups should be generated using the same set of parameters under the null hypothesis.

Table 1 describes the parameters we used for each random graph model.

Table 1. Parameters used in the null hypothesis simulation

Model	Parameters
Erdös-Rényi	$p_1 = 0.1$ and $p_2 = 0.2$
Watts-Strogatz	$p_1 = 0.1$ and $p_2 = 0.3$ $neigh = 10, dim = 1$
Barabási-Albert	$p_1 = 1.0$ and $p_2 = 1.1$

Since we generated all groups using the same models and parameters, we can safely assure that they all come from the same population. In other words, they are under the null hypothesis. Under the null hypothesis, we expect that the distribution of p-values forms a uniform distribution in the $[0, 1]$ range.

Figure 1 shows the distribution of the p-values. As we can see, they form a uniform distribution, thus showing that our proposal controls the rate of false positives.

We remark that the power of the test increases with the number of graphs. Thus, a small number of graphs ($N = 30$) shows that even under a small sample size, the ANOGVA method performs well.

Now we verify the H_1 hypothesis.

Simulation (H_1)**:** To verify the power of the test, we need to generate groups from distinct populations.

Table 2 describes which parameters we used for each random graph model.

Since we generated all the groups using distinct parameters, this simulation satisfies the requirements for H_1, where we generated at least one of the groups (in this case, all of them) differently. Under the alternative hypothesis, we expect all the p-values to be small.

In all models, the resulting p-values were all equal to zero.

We can see that the ANOGVA method satisfies our expectations.

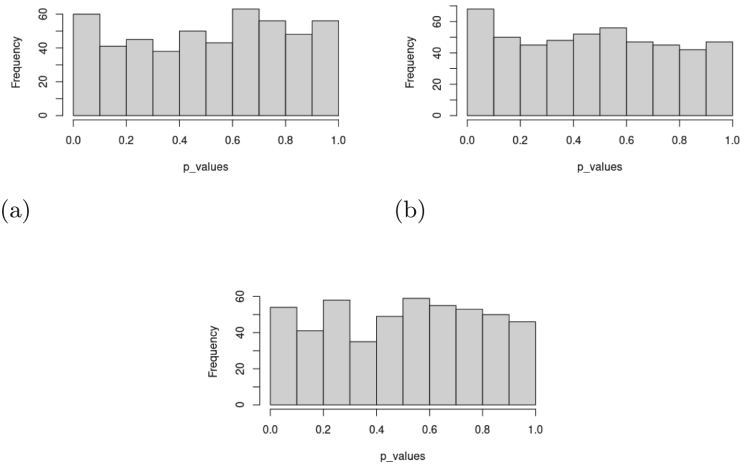

Fig. 1. (a) Distribution of p-values for the ANOGVA simulation under the null hypothesis using the Erdös-Rényi model. (b) Distribution of p-values for the ANOGVA simulation under the null hypothesis using the Watts-Strogatz model (c) Distribution of p-values for the ANOGVA simulation under the null hypothesis using the Barabási-Albert model. Notice that all of them are uniform distributions. Performing a Kolmogorov-Smirnov test comparing these values with the uniform distribution gives us p-values greater than 0.05.

Table 2. Parameters used in the alternative hypothesis simulations

Model	G_1	G_2	G_3
Erdős-Rényi	$p_1 = 0.1$, $p_2 = 0.3$	$p_1 = 0.2$, $p_2 = 0.4$	$p_1 = 0.3$, $p_3 = 0.2$
Watts-Strogatz	$p_1 = 0.1, p_2 = 0.3$, $neigh = 10, dim = 1$	$p_1 = 0.2$, $p_2 = 0.7$, $neigh = 10, dim = 1$	$p_1 = 0.3$, $p_3 = 0.3$, $neigh = 10, dim = 1$
Barabási-Albert	$p_1 = 1.1$, $p_2 = 1.3$	$p_1 = 1.1$, $p_2 = 1.8$	$p_1 = 1.7$, $p_2 = 1.8$

5 Applications to Biological Data

To illustrate ANOGVA, we applied it to a biological dataset. We used the data source titled 'Anesthesia Task' [20]. We obtained it from Project Tycho and downloaded it via their website at http://wiki.neurotycho.org The experiment aimed to compare neural activity between most of the lateral cortex measured with electrocorticographic signals (ECoG) in a macaque during five stages: awake with eyes opened, awake with eyes closed, anesthetized, recovering with eyes closed, and recovering with eyes open.

5.1 Data Source

Four experiments were conducted, each on a different monkey. In each experiment, a monkey was seated in a chair with restricted arms and head movement.

In particular, the following steps describe the experiment for the monkey we analyzed. Neural data was acquired through 128 ECoG electrodes measuring ECoG signals from most of the lateral cortex. Neural activity was recorded during all of the following stages. Initially, the monkey was awake and opened its eyes, sitting calmly in his chair for 10 min. Next, the eyes of the monkey were covered with an eye mask to avoid evoking a visual response. The monkey was left sitting in his chair for another 10 min. Recording of neural activity was stopped while anesthesia was intramuscularly injected into the monkey. By the point at which the monkey had stopped responding to manipulation of the monkey's hand or touching the nostril or philtrum with a cotton swab, neural activity recording was resumed for another 20 min. After the anesthetized condition, the monkey recovered from the anesthesia and was left alone for 55 min with its eyes still covered. Next, the eye mask was removed, and the monkey was left to sit calmly on his chair for another 10 min.

5.2 Data Processing and Graph Generation

The initial data generated by the experiment consisted of 128-time series in 5 categories: conscious with open eyes, conscious with closed eyes, anesthetized, recovering with closed eyes, and recovering with open eyes.

Initially, the data was processed through several finite impulse response (FIR) filters to remove any effect caused by electrical interference. We divided the filtered data into several time windows, each lasting four seconds and generated the graph using generalized partial directed coherence (gPDC) [16].

The gPDC is a frequency domain approach to identify the direction of information flow (Granger causality) between multiple time series. We say that a time series X Granger cause another time series Y if knowledge of $X(t-1), \ldots, X(t-k)$ increases the prediction of $Y(t)$.

We carried out gPDC on the 128 frequencies of the filtered data. The result was five sets of 128 groups of graphs (one for each generated frequency). Each group consisted of several graphs, each representing a time window in its category. Each graph had 128 nodes (each corresponding to a different ECoG electrode). The graph was directed and weighted, where each edge between two nodes corresponded to the level of causality between the ECoG electrodes.

5.3 ANOGVA

We performed the following experiment to verify the power of the ANOGVA method. We selected a single-frequency domain. Given that frequency, we chose 100 graphs from each category. This procedure resulted in the following:

1. G_1: 100 graphs generated from when the monkey was awake with its eyes opened.
2. G_2: 100 graphs generated from when the monkey was awake with closed eyes.
3. G_3: 100 graphs generated from when the monkey was under anesthesia.

4. G_4: 100 graphs generated from when the monkey was recovering with closed eyes.
5. G_5: 100 graphs generated from when the monkey was recovering with closed eyes.

First Experiment: We first performed an ANOGVA test using the five groups. We used 1 000 bootstrap samples.

Second Experiment: We then performed the same experiment but compared it in a pairwise manner. Similarly to the previous experiment, we used 1 000 bootstrap samples.

Third Experiment: Since all graphs originate from the same monkey, there is a possibility that obtaining low p-values in the previous experiments is not a consequence of the difference between the distinct categories. To verify that the significance of the previous experiments was valid, we performed an ANOGVA test under the null hypothesis. In specific, we performed the following for each group G_i.

1. We split group G_i into two randomly sampled groups with no replacement, obtaining $G_{i,1}$ and $G_{i,2}$
2. We performed an ANOGVA test on these groups with 300 bootstrap samples.
3. We stored the calculated p-value.
4. We repeated this procedure 300 times, generating a distribution of p-values.

Suppose we explain low p-values because all graphs originate from the same monkey. In that case, performing ANOGVA using the described setup should give us mostly low p-values.

5.4 Results

First Experiment: For the first experiment, we obtained a p-value less than $\frac{1}{300}$. This shows that there is at least one sample of graphs that were generated differently.

Second Experiment: Table 3 shows the p-values obtained when comparing groups G_i and G_j. We note the low p-values, indicating that our method could differentiate between any two groups.

Third Experiments: Figure 2 shows the distribution of the p-values when comparing each group with itself. Any fear that previous low p-values might be because both groups originate from the same monkey can be eased by looking at the results of this experiment. We note a well-defined uniform distribution in each group, proving that the graphs from the same monkey are insufficient to justify a low p-value between groups.

Table 3. Results of second experiment:

	G_1	G_2	G_3	G_4	G_5
G_1		0.002	0	0	0
G_2	0.002		0	0	0.076
G_3	0	0		0	0
G_4	0	0	0		0
G_5	0	0.076	0	0	

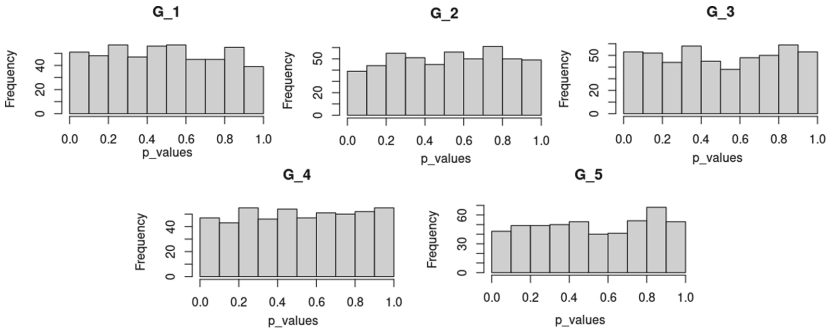

Fig. 2. Results of the third experiment.

Experiment Conclusion: We have shown that our methods can differentiate between the brain connectivity networks associated with all stages in the anesthesia experiment. These results promise that our methods can be used in future clinical trials.

6 Conclusion

To distinguish between populations of directed graphs, we explored measures based on the graph spectrum. We compared groups of graphs by calculating the Kullback-Leibler divergence between the graphs' spectra. This led to the development of ANOGVA, a non-parametric model for testing whether two or more groups of graphs share the same spectral distribution.

We demonstrated that our proposed method effectively distinguishes populations of graphs generated by different parameters, irrespective of the model used. Similarly, regular ANOVA on centrality measures can also distinguish various models. However, traditional ANOVA fails when centrality measures, like the number of edges in the Watts-Strogatz random model, remain unchanged. In our illustrative application with ECoG data, we successfully captured changes in the neural activity network of anesthetized monkeys. Unlike many classification methods, the proposed method can be used in clinical settings for diagnosing psychological conditions without the need for model training.

Our current approach is limited to single-edge graphs, in which a node i can only be connected to a node j via, at most, one edge. Multi-edge graphs, which permit several connections between two nodes, are not represented so simply by an adjacency matrix, and thus our method fails to apply.

Acknowledgements. This work has been supported by FAPESP grants 2018/21934-5, 2019/22845-9, and 2020/08343-8, CNPq grant 303855/2019-3 and 440245/2022-2, CAPES (finance code 001), Alexander von Humboldt Foundation, the Academy of Medical Sciences - Newton Fund, and Wellcome Leap.

References

1. Alon, U.: An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman & Hall/CRC Mathematical and Computational Biology (2006)
2. Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5**, 101–113 (2004)
3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509–512 (1999)
4. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. **10**, 186–198 (2009)
5. Csardi, G., Nepusz, T.: The igraph software package for complex network research. Interjournal Complex Syst. **1695** (2006)
6. Duong, T.: KS: kernel density estimation and kernel discriminant analysis for multivariate data in R. J. Stat. Softw. **21**(7), 1–16 (2007)
7. Erdős, P., Rényi, A.: On random graphs I. Publ. Math. Debrecen **6**, 290–297 (1959)
8. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977)
9. Fujita, A., Vidal, M.C., Takahashi, D.Y.: A statistical method to distinguish functional brain networks. Front. Neurosci. **11**, 66 (2017)
10. Fujita, A., Silva Lira, E., De Siqueira Santos, S.: A semi-parametric statistical test to compare complex networks. J. Complex Netw. **8** (2020)
11. Fujita, A., Takahashi, D.Y., Balardin, J.B., Vidal, M.C., Sato, J.R.: Correlation between graphs with an application to brain network analysis. Comput. Stat. Data Anal. **109**, 76–92 (2017)
12. Lees-miller, J., et al.: Correlation between graphs with an application to brain network analysis. Comput. Stat. Data Anal. **109**, 76–92 (2017)
13. MacKay, D.J.: Information Theory, Inference, and Learning Algorithms, 1st edn. Cambridge University Press, Cambridge (2003)
14. Ramos, T.C., Mourão-Miranda, J., Fujita, A.: Spectral density-based clustering algorithms for complex networks. Front. Neurosci. **17**, 926321 (2023)
15. Ribeiro, A., Vidal, M., Sato, J., Fujita, A.: Granger causality among graphs and application to functional brain connectivity in autism spectrum disorder. Entropy **23**, 1204 (2021)
16. Sameshima, K., Baccala, L.: Methods in brain connectivity inference through multivariate time series analysis (2016)
17. Santos, S.S., Fujita, A.: statGraph: statistical methods for graphs (2017). www.cran.r-project.org/package=statGraph
18. Scott, J.: Social Network Analysis. Sage, Newcastle upon Tyne (2012)

19. Watts, D., Strogatz, S.: Collective dynamics of "small-world' networks. Nature **393**, 440–442 (1998)
20. Yanagawa, T., Chao, Z.C., Hasegawa, N., Fujii, N.: Large-scale information flow in conscious and unconscious states: an ECoG study in monkeys. PLoS ONE **8**(11), e80845 (2013)