Causalidade de Granger entre grafos no domínio da frequência

Gustavo Pinto Vilela

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Ciência da Computação

Orientador: Prof. Dr. André Fujita

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da FAPESP (Processo 2012/12320-7) e do CNPq (Processo 132993/2011-2)

Novembro de 2016
São Paulo

Causalidade de Granger entre grafos no domínio da frequência

Esta é a versão original da tese elaborada pelo candidato (Gustavo Pinto Vilela), tal como submetida à Comissão Julgadora.

Aos meus pais.

Agradecimentos

Agradeço imensamente aos meus pais, Mário e Rosy, à minha irmã Caryna e aos meus amigos, em especial ao Marcos e à Nádia, por todo apoio, incentivo, paciência e compreensão durante o período do doutorado.

Ao meu orientador, Professor André Fujita, pelos conselhos, preocupação, discussões e as diversas contribuições à minha pesquisa.

A todos os amigos de laboratório, especialmente Fernando, Gabriela, Igor, Juan e Paulo, por toda a confiança, descontração, colaboração e produtivas discussões filosóficas. E à Suzana pela preocupação e ajuda com o desenvolvimento do meu trabalho.

Às instituições e seus funcionários que apoiaram o desenvolvimento desta tese. Ao Instituto de Matemática e Estatística e à Universidade de São Paulo por disponibilizar toda a infraestrutura necessária ao projeto. À FAPESP (Processo 2012/12320-7) e ao CNPq (Processo 132993/2011-2) pelo financiamento parcial da pesquisa.

Resumo

VILELA, G. P. **Causalidade de Granger entre grafos no domínio da frequência**. 2016. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

Diversos sistemas naturais, como a malha aeroviária, interações proteína-proteína, regulação genética, conectividade funcional do cérebro e relações sociais podem sem modeladas por grafos onde os vértices são as entidades sob estudo e as arestas representam quais pares de entidades se relacionam. Também é sabido que muitos desses sistemas são modulares, ou seja, podem ser particionados de alguma maneira em sub-sistemas que interagem ou se influenciam. No entanto, do ponto de vista estatístico-computacional, pouco se é conhecido sobre métodos de análise estatística em grafos. Por exemplo, como identificar que um grafo "causa" outro grafo? Dentro deste contexto, propomos um método de identificação de causalidade de Granger entre séries temporais de grafos no domínio da frequência. Este método se baseia tanto na análise espectral dos grafos aleatórios como também no método da Coerência Parcial Direcionada. Apresentamos o modelo, uma forma de estimação, um teste estatístico e resultados sobre o efetivo controle da taxa de falsos positivos, bem como seu poder estatístico em simulações de Monte Carlo. Finalmente, ilustramos uma aplicação do método em dados de eletrocorticografia coletados de um macaco sob estado de alerta e posteriormente em estado anestésico.

Palavras-chave: teoria dos grafos; causalidade de Granger; raio espectral; análise espectral; coerência parcial direcionada.

Abstract

VILELA, G. P. **Granger causality between graphs in frequency domain**. 2016. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

Several natural systems such as protein-protein interactions, genetic regulation, functional connectivity of the brain, and social relationships can be modeled as graphs where its vertices represent the entities under study and the edges represent which pair of entities are associated. It is known that much of these systems are modular, i.e., they can be clustered into sub-systems, which interact and influence each other. However, from a computational statistical viewpoint, little is known about statistical methods to analyze graphs. For example, how can one identify whether a graph "causes" another graph? In this context, we propose a method to identify Granger causality among time series of graphs in the frequency domain. This method is based on the idea of spectral analysis of random graphs and also on the Partial Directed Coherence. We present the model, a method to estimate the parameters of the model, and a statistical test. We demonstrate the usefulness of the method in intensive Monte Carlo simulations. Results show that the method effectively controls the type I error and also present high statistical power to identify Granger causality in five different random graph models. Finally, we illustrate an application of the method in an electrocorticography data collected from a macaque under anesthesia.

Keywords: graph theory; Granger causality; spectral radius; spectral analysis; partial directed coherence.

Sumário

Capítulo 1

Introdução

Recentemente muita atenção tem sido dada ao estudo das estruturas das redes do mundo real utilizando grafos aleatórios, por exemplo, redes sociais (Bollobás e Riordan , 2004), redes de interação proteína-proteína (Maslov e Sneppen, 2002), redes metabólicas (Jeong *et al.*, 2000), redes neuronais (Eguiluz *et al.*, 2005), redes da World Wide Web (Huberman e Adamic, 1999). Matematicamente, estas redes podem ser representadas por grafos, isto é, um conjunto de vértices e um conjunto de arestas, onde as arestas indicam quais pares de vértices estão associados. Os vértices podem representar as mais diversas variáveis como genes e proteínas das redes de regulação genética, aeroportos da malha aeroviária, pessoas das redes sociais, etc.

A maioria dos estudos das estruturas das redes são feitos, essencialmente, por meio de algumas medidas da estrutura do grafo, como diâmetro, número de arestas, integração funcional, motivos de redes e centralidades (Alon, 2007; Bullmore e Sporns, 2009; Mangan e Alon , 2003; Rubinov e Sporns, 2010; Zuo *et al.*, 2012). No entanto, essas análises dificilmente levam em conta a variação intrínseca das redes do mundo real, em que podem ocorrer flutuações.

Por exemplo, as redes de conectividade funcional do cérebro e as redes regulatórias de genes são variantes no tempo além de também serem diferentes de indivíduo para indivíduo. Isso faz com que, não somente metodologias computacionais, mas também estatísticas para grafos se tornem necessárias.

Uma das perguntas naturais no estudo de grandes redes é, como é a relação entre subsistemas ou sub-redes? Por exemplo, é sabido que a rede funcional do cérebro é modular, ou seja, particionada em sub-redes. Nesse tipo de rede cada vértice corresponde a uma região do cérebro e uma aresta conectando duas regiões indica que os níveis de atividade dessas regiões estão correlacionados. Como as sub-redes do cérebro interagem dependendo do estado do indivíduo? Como é o fluxo de informação de uma sub-rede à outra? Uma alternativa no estudo de interações e que está intimamente ligada ao conceito de fluxo de informação (Baccalá e Sameshima, 2001) é o conceito de causalidade de Granger (Granger, 1969).

Causalidade de Granger (Granger, 1969) foi primeiramente introduzida na área econométrica para analisar as relações e influências entre as séries temporais macroeconômicas. O estudo de séries temporais é interessante em razão de sua ordenação natural no tempo, permitindo estabelecer relações de causa e efeito entre as variáveis e fornecendo a direção do fluxo de informação. A causalidade de Granger é baseada na ideia de que o efeito nunca precede sua causa. Portanto, relações temporais como atrasos ou precedências podem conter informação sobre a causalidade.

Mais formalmente, uma série temporal x_t é dita como Granger causa de uma série temporal y_t, se o erro de predição de y_t no tempo t dada toda informação do passado $(x_{t-1}, x_{t-2}, \ldots, y_{t-1}, y_{t-2}, \ldots)$ é menor que considerando somente a informação nos passados de y_t $(y_{t-1}, y_{t-2}, \ldots)$. Portanto, valores passados de x_t permitem prever y_t. Note que esta relação não necessariamente precisa ser recíproca, isto é, podemos interpretar a direcionalidade como um fluxo de informação entre as séries temporais. No entanto, é necessário salientar que causalidade de Granger não implica em "causalidade efetiva". A causalidade de Granger é baseada somente na predição e informação numérica, enquanto a causalidade efetiva está profundamente relacionada com a influência de uma entidade sobre a outra. Por outro lado, devido à sua simplicidade, a causalidade de Granger pode ser útil para inferir ou sugerir causalidade efetiva. Comumente a causalidade de Granger é identificada no domínio do tempo pelo modelo vetor autoregressivo (Lütkepohl, 2011) ou no domínio da frequência pela coerência parcial direcionada (Baccalá e Sameshima, 2001). Apesar de ter sido primeiramente sugerido na área econométrica, existem diversas aplicações da causalidade de Granger em outras áreas como biologia molecular (Fujita *et al.*, 2007a,b, 2008, 2010a,b,c, 2012) e neurociência (Baccalá e Sameshima, 2001; Sato *et al.*, 2009, 2010, 2006; Takahashi *et al.*, 2007, 2010).

Um ponto interessante no estudo de grafos seria a possibilidade de identificar causalidade de Granger entre eles. Por exemplo, em neurociência, existe um grande número de evidências que sugerem que processos cognitivos complexos surgem a partir de uma comunicação orquestrada das áreas do cérebro, isto é, das redes de conectividade funcional do cérebro (Cassidy *et al.*, 2016). Mais recentemente, tem se dado muita atenção à ideia de que as regiões do cérebro são organizadas em sub-redes ou comunidades que, por sua vez, são interconectadas (Sporns, 2013). Assim, a identificação da conectividade entre sub-redes poderia auxiliar no entendimento das diferenças entre controles e doentes que não são detectadas por métodos tradicionais. Até onde pudemos constatar, não existe na literatura um método de identificação de causalidade de Granger entre grafos.

Isto ocorre basicamente porque redes ou grafos são difíceis de serem manipulados de um ponto de vista estatístico (grafos não são escalares). Assim, para construir uma metodologia para identificar causalidade de Granger entre grafos, uma ideia natural seria imaginar que grafos são gerados por um modelo matemático com um conjunto de parâmetros que são as variáveis aleatórias. Intuitivamente, dois vetores (séries temporais) de grafos têm uma relação

em termos de causalidade de Granger quando os parâmetros (nossas variáveis aleatórias) do modelo usado para gerar os grafos de um vetor Granger causam os parâmetros que geraram o outro vetor.

No entanto, dadas duas, ou mais, séries temporais de grafos, os modelos gerados são raramente conhecidos, e, consequentemente, os parâmetros não podem ser estimados. Assim, se torna necessário identificar uma característica do grafo que é altamente correlacionada com os parâmetros do grafo.

Em 2012, Takahashi *et al.* introduziram ideias sobre análises estatísticas em grafos como métodos de seleção de modelos, estimador de parâmetros e um teste estatístico, baseado na análise espectral (o espectro de um grafo é o conjunto de autovalores de sua matriz de adjacência) dos grafos. A intuição por trás do estudo do espectro do grafo é que propriedades estruturais, como o número de caminhos, diâmetro (é o maior caminho dentro do conjunto dos caminhos mínimos entre todos os pares de vértices), e cliques (um subconjunto dos vértices em que todos estão conectados entre si) estão contidos no espectro (Van Mieghem, 2010). Aplicações foram bem sucedidas nas áreas biológicas como biologia molecular (Takahashi *et al.*, 2012) e neurociência (Sato *et al.*, 2013, 2015). Baseado nas ideias iniciais de (Takahashi *et al.*, 2012), Fujita *et al.* (2015) identificaram que o raio espectral (maior autovalor da matriz de adjacência do grafo) é altamente correlacionado com os parâmetros de vários modelos geradores de grafos aleatórios e propuseram uma medida de correlação entre grafos baseada nesse valor.

Aqui propomos estimar a causalidade de Granger no domínio da frequência entre séries temporais de grafos usando o raio espectral. Nossos resultados mostram que o raio espectral está altamente correlacionado com os parâmetros do modelo que geram o grafo, e assim, pode ser usado como uma boa característica para identificar causalidade de Granger entre séries temporais de grafos no domínio da frequência. Por simulações, mostramos o alto poder estatístico como também o efetivo controle da taxa do erro tipo I (falsos positivos) da nossa proposta. Em seguida, ilustramos a aplicação do método de identificação de causalidade de Granger em dados de eletrocorticografia do cérebro de um macaco em estado de alerta e posteriormente sob efeitos de uma anestesia.

1.1 Objetivos

O objetivo da tese é a proposta de um método que permite identificar causalidade de Granger entre séries temporais de grafos através da análise espectral. Mostraremos também a adequação do método proposto a aplicação do método em dados reais.

Os objetivos específicos do trabalho podem ser divididos em:

1. apresentação do método para identificar causalidade de Granger entre séries temporais de grafos no domínio da frequência;

2. mostrar a adequação do uso do raio espectral;

3. mostrar o poder estatístico do método e controle sobre a taxa do erro tipo I através de modelos simulados;

4. aplicar o método proposto em dados reais de eletrocorticografia (ECoG).

1.2 Organização do texto

A sequencia da tese está organizada da seguinte forma.

O Capítulo 2 introduz alguns dos conceitos e métodos que servem de base para a formalização do método proposto. Introduzimos também o método para identificar causalidade de Granger entre grafos no domínio da frequência (Objetivo 1).

No Capítulo 3 mostramos a adequação da utilização do raio espectral como uma escolha representativa dos parâmetros da função geradora do grafo (Objetivo 2). Mostramos também, por meio de simulações de Monte Carlo, o poder estatístico do método proposto, bem como seu controle efetivo da taxa de falsos positivos (Objetivo 3).

O Capítulo 4 mostra a aplicação do método proposto em neurociência (Objetivo 4) utilizando dados reais de ECoG obtidos de um macaco em estado de alerta e posteriormente sob efeito de uma anestesia.

Por fim, no Capítulo 5 apresentamos nossas conclusões, comentamos brevemente algumas limitações do método proposto e indicamos possível direções futuras que podem ser exploradas.

Capítulo 2

Causalidade de Granger entre grafos

Neste capítulo são apresentados inicialmente alguns conceitos importantes como grafos e causalidade de Granger entre séries temporais. Em seguida, apresentaremos a nossa proposta que consiste em um método de identificação de causalidade de Granger entre séries temporais de grafos.

2.1 Grafos

Um grafo é um par de conjuntos $G = (V, E)$, onde V é um conjunto de n vértices $(v_1, v_2, \ldots, v_n)$ e E é um conjunto de m arestas, cada uma conectando dois vértices de V.

Qualquer grafo não dirigido (sem direção nas arestas) G com n vértices pode ser representado por sua matriz de adjacências $\mathbf{A}^G$ com $n \times n$ elementos $\mathbf{A}^G_{ij}$ $(i, j = 1, \ldots, n)$; seu valor é $\mathbf{A}^G_{i,j} = \mathbf{A}^G_{j,i} = 1$ se os vértices v_i e v_j são conectados e 0 caso contrário.

O espectro do grafo G é o conjunto de autovalores de sua matriz de adjacências $\mathbf{A}^G$. Assim, um grafo não dirigido com n vértices tem n autovalores reais $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Os autovalores são reais em decorrência da matriz de adjacências de um grafo não dirigido ser simétrica.

2.2 Causalidade de Granger

A ideia por trás da causalidade de Granger reside na suposição de que associações temporais podem conter informações que sugiram causalidade. Mais intuitivamente, é baseada na ideia de que o efeito jamais antecede a causa. Assim, se a série temporal x_t afeta a série temporal y_t, x_t deveria auxiliar na predição de y_t (Granger, 1969).

Para formalizar a ideia da causalidade de Granger (Lütkepohl, 2011), suponha que ζ_t é o conjunto contendo toda a informação relevante até (e incluindo) o tempo t. Seja $y_t(h|\zeta_t)$ o preditor ótimo de passo h do processo y_t na origem t, baseado na informação em ζ_t. O preditor do mínimo erro quadrático médio será denotado por $\Omega_t(h|\zeta_t)$. O processo x_t Granger-causa

y_t se $\Omega_t(h|\zeta_t) < \Omega_t(h|\zeta_t \backslash \{x_s|s \leq t\})$ para algum $h = 1, 2, \ldots$, onde $\zeta_t \backslash \{x_s|s \leq t\}$ é o conjunto de toda a informação exceto a informação no passado e presente do processo x_t. Em outras palavras, se y_t pode ser predito mais eficientemente quando a informação do processo x_t é levado em conta além de todo o restante da informação, então x_t é dito como Granger causa de y_t.

2.3 Métodos de identificação de causalidade de Granger

Existem diversos métodos para identificar causalidade de Granger em séries temporais tanto no domínio do tempo quanto no da frequência. Dentre eles, dois que desempenham um importante papel na identificação de causalidade de Granger em séries temporais multivariadas são o modelo de Vetores Autoregressivos (VAR) no domínio do tempo e o método da Coerência Parcial Direcionada (PDC - *Partial Directed Coherence*) no domínio da frequência.

2.3.1 Vetores Autoregressivos - VAR

Antes de definirmos o método para identificar causalidade de Granger no domínio da frequência, primeiro descreveremos o modelo vetor autoregressivo usado na identificação da causalidade de Granger no domínio do tempo (Lütkepohl, 2011).

Sejam:

- k o número de séries temporais;
- p a ordem do modelo (número de pontos no tempo no passado a ser analisado);
- T o comprimento da série temporal;
- $y_{i,t}$ a i-ésima série temporal, e
- $\varepsilon_{i,t}$ o vetor de variáveis aleatórias para a $i$ésima série temporal, com média zero e matriz de covariância

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{1,1}^2 & \sigma_{2,1} & \ldots & \sigma_{k,1} \\ \sigma_{1,2} & \sigma_{2,2}^2 & \ldots & \sigma_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,k} & \sigma_{2,k} & \ldots & \sigma_{k,k}^2 \end{pmatrix}.$$

Note que os resíduos ε_t são serialmente não correlacionados, mas podem ser contemporaneamente correlacionados. Em outras palavras, $\mathbf{\Sigma}$ pode não ser necessariamente uma matriz identidade.

Então, o sistema de equações de um modelo vetor autoregressivo k-dimensional de ordem p é dado por:

$$\begin{cases} y_{1,t} = v_1 + a^1_{1,1}y_{1,t-1} + \ldots + a^p_{1,1}y_{1,t-p} + \ldots + a^1_{k,1}y_{k,t-1} + \ldots + a^p_{k,1}y_{k,t-p} + \varepsilon_{1,t} \\ y_{2,t} = v_2 + a^1_{1,2}y_{1,t-1} + \ldots + a^p_{1,2}y_{1,t-p} + \ldots + a^1_{k,2}y_{k,t-1} + \ldots + a^p_{k,2}y_{k,t-p} + \varepsilon_{2,t} \\ \vdots \\ y_{k,t} = v_k + a^1_{1,k}y_{1,t-1} + \ldots + a^p_{1,k}y_{1,t-p} + \ldots + a^1_{k,k}y_{k,t-1} + \ldots + a^p_{k,k}y_{k,t-p} + \varepsilon_{k,t} \end{cases}$$

Para simplificar e facilitar a estimação dos coeficientes deste modelo, re-escreveremos o sistema de equações no formato matricial.

Sejam

$$\mathbf{Y} = \begin{pmatrix} y_{1,p+1} & y_{2,p+1} & \cdots & y_{k,p+1} \\ y_{1,p+2} & y_{2,p+2} & \cdots & y_{k,p+2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,T} & y_{2,T} & \cdots & y_{k,T} \end{pmatrix},$$

$$\mathbf{Z} = \begin{pmatrix} y_{1,p} & y_{1,p-1} & \cdots & y_{1,1} & \cdots & y_{k,p} & y_{k,p-1} & \cdots & y_{k,1} \\ y_{1,p+1} & y_{1,p} & \cdots & y_{1,2} & \cdots & y_{k,p+1} & y_{k,p} & \cdots & y_{k,2} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{1,T-1} & y_{1,T-2} & \cdots & y_{1,T-p} & \cdots & y_{k,T-1} & y_{k,T-2} & \cdots & y_{k,T-p} \end{pmatrix},$$

e

$$\mathbf{A} = \begin{pmatrix} a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ a^p_{1,1} & a^p_{1,2} & \cdots & a^p_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ a^1_{k,1} & a^1_{k,2} & \cdots & a^1_{k,k} \\ \vdots & \vdots & \ddots & \vdots \\ a^p_{k,1} & a^p_{k,2} & \cdots & a^p_{k,k} \end{pmatrix},$$

o modelo matricial pode ser escrito como

$$\mathbf{Y} = \mathbf{ZA} + \mathbf{u}. \tag{2.1}$$

Então, os coeficientes do modelo ($a^l_{i,j}$, com $i, j = 1, \ldots, k$ e $l = 1, \ldots, p$) podem ser estimados por método de mínimos quadrados ordinários como

$$\hat{\mathbf{A}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{2.2}$$

A matriz de resíduos $\mathbf{u}$ com dimensões $((T-p) \times k)$ pode ser estimada como

$$\hat{\mathbf{u}} = \mathbf{Y} - \mathbf{Z}\hat{\mathbf{A}}. \tag{2.3}$$

Uma condição necessária e suficiente para a série temporal $y_{i,t}$ ser não Granger causa

para o grafo $y_{j,t}$ é se, e somente se, $a^l_{i,j} = 0$.

Assim, a não causalidade de Granger pode ser identificada testando a significância dos elementos dos coeficientes autoregressivos ($\mathbf{A}$) dos modelos vetor autoregressivo.

O teste de hipótese para a significância da conectividade do $\mathbf{A}$ é $H_0 : \mathbf{CA} = 0$ versus $H_1 : \mathbf{CA} \neq 0$ onde $\mathbf{C}$ é uma matriz de contrastes dos parâmetros que queremos testar. Este teste pode ser obtido através da aplicação de um teste de Wald (Graybill, 1976).

Suponha que estejamos interessados em testar se y_i Granger causa y_j. Sejam $\hat{\Sigma} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{(T-p)-(kp)}$ a matriz de covariância de dimensão $(k \times k)$, $\mathbf{c}$ uma matriz de dimensões $(1 \times k)$ com o valor 1 na i-ésima posição, $\mathbf{0}$ uma matriz de zeros com dimensões $(1 \times k)$ e

$$\mathbf{C} = \begin{pmatrix} \mathbf{c} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{c} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{c} \end{pmatrix},$$

uma matriz de contrastes de dimensões $(p \times (kp))$.

Então, a estatística do teste de Wald é dado por $W = \frac{(\mathbf{CA}_j)'(\mathbf{C}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{C}')^{-1}(\mathbf{CA}_j)}{\Sigma_{j,j}}$.

A estatística do teste de Wald W segue uma distribuição χ^2 com posto ($\mathbf{C}$) graus de liberdade.

2.3.2 Coerência Parcial Direcionada - PDC

Apesar do conceito original de causalidade de Granger ser bastante útil no domínio do tempo, é uma abordagem que não permite discernir qual frequência do sinal é importante, no sentido em que permita prover interpretações interessantes dos dados. Um exemplo clássico é em neurociência, onde artefatos ou sinais não-neurológicos, como batimento do coração ou respiração, podem mascarar o sinal cerebral. A fim de resolver esses problemas, Sameshima e Baccalá propuseram um método chamado Coerência Parcial Direcionada (PDC - Partial Directed Coherence).

A técnica PDC (Baccalá e Sameshima, 2001) é uma adaptação do modelo VAR para o domínio da frequência que generaliza o conceito de coerência direta (Baccalá e Sameshima , 1998; Baccalá *et al.*, 1998; Saito e Harashima, 1981). O método do PDC é utilizado para inferir influências dirigidas em séries temporais multivariadas através da medida da força das interações entre elas. O PDC da série $y_{j,t}$ para $y_{i,t}$ na frequência ψ é a medida da razão entre a quantidade de informação partindo de $y_{j,t}$ para $y_{i,t}$ e o total do fluxo de informação partindo de $y_{j,t}$ para qualquer outra série. Ele é calculado da seguinte forma. Seja

$$a_{i,j}(\psi) = \delta_{i,j} - \sum_{l=1}^{p} a^l_{i,j} \exp(-2\pi/\psi\sqrt{-1}) \tag{2.4}$$

para $\delta_{i,j} = 1$ se $i = j$ e 0 caso contrário, então o PDC da série $y_{j,t}$ para a série $y_{i,t}$ é dado

por:

$$\pi_{i,j}(\psi) = \frac{a_{i,j}(\psi)\frac{1}{\sigma_i}}{\sqrt{\sum_{i=1}^{k} |a_{i,j}|^2 \frac{1}{\sigma_i^2}}} \tag{2.5}$$

Em razão da normalização, o valor de $\pi_{i,j}(\psi)$ varia entre 0 e 1. Um valor de $\pi_{i,j}(\psi)$ próximo de 0 indica a ausência de influência direta de $y_{j,t}$ para $y_{i,t}$ na frequência ψ, e um valor próximo de 1 indica que todo o espectro de potência de $y_{j,t}$ está indo para $y_{i,t}$. Se $\pi_{i,j}(\psi) = 0$ em toda frequência ψ então não há causalidade de Granger de $y_{j,t}$ para $y_{i,t}$. Portanto, testar a ausência de causalidade de Granger de $y_{j,t}$ para $y_{i,t}$ é o mesmo que testar $\pi_{i,j}(\psi) = 0$ para toda frequência ψ.

O teste de ausência de causalidade de Granger da série temporal $y_{j,t}$ para $y_{i,t}$ é baseado na ideia de testar se $\pi_{i,j}(\psi) = 0$ no contexto do PDC. A hipótese nula do teste é de ausência de causalidade de Granger e o teste pode ser definido pelas seguintes hipóteses:

$$H_0 : \pi_{i,j}(\psi) = 0 \quad \forall\psi$$
$$H_1 : \pi_{i,j}(\psi) \neq 0 \quad \text{para algum } \psi$$

Apesar de existir um teste estatístico analítico para o PDC (Takahashi *et al.*, 2007), comumente as séries temporais não são muito longas além também de não satisfazerem as hipóteses do teste. Assim, propomos a aplicação de um método bootstrap para estimar a distribuição da estatística $\pi_{i,j}(\psi)$. O algoritmo bootstrap pode ser descrito como (Baccalá *et al.*, 2006; Sato *et al.*, 2009):

Entrada: As séries temporais.

1. Ajuste um modelo VAR nas séries temporais.
2. Obtenha os coeficientes estimados do VAR como também os resíduos das séries temporais utilizando as equações 2.2 e 2.3, respectivamente.
3. Para cada frequência, calcule o PDC estimado usando a equação 2.4.
4. Para cada série temporal, re-amostre os resíduos obtidos no passo com reposição.
5. Para testar a influência da j-ésima série temporal na i-ésima série temporal, assuma o modelo onde os coeficientes do VAR $a^l_{i,j}$, com $l = 1, 2, \ldots, p$ (i.e., todos os coeficientes da série temporal da j-ésima causando a i-ésima) são zero. Os demais coeficientes do VAR continuam com os mesmos valores originalmente estimados pelo VAR do passo 4. A ideia por trás aqui consiste em assumir um modelo sob a hipótese nula, ou seja, na ausência de causalidade de Granger da j-ésima série para a i-ésima série, e assim gerar as amostras bootstraps sob a hipótese nula.
6. Utilizando os resíduos re-amostrados no passo 4, e o modelo desenhado sob a hipótese nula do passo 5, simule as séries temporais bootstrap (ou seja, as séries temporais sob a hipótese nula).
7. Obtenha a mediana do PDC para essa amostra bootstrap.
8. Volte ao passo 4 até que se obtenha o número desejado de amostras bootstraps (usualmente algo como 1.000).
9. Estime o p-valor como a fração das vezes que a mediana do PDC obtido passo 7 é igual ou maior que a mediana do PDC obtida nos dados originais.

Saída: o p-valor para o PDC da j-ésima série temporal para a i-ésima série temporal.

2.4 Proposta

Nossa proposta consiste em considerar que os parâmetros dos modelos dos grafos são nossas variáveis aleatórias. Assim, assumimos que não há causalidade de Granger entre séries temporais de grafos se, e somente se, não há causalidade de Granger entre os parâmetros dos modelos dos grafos.

Os parâmetros da distribuição desta variável aleatória será denotado pelos hiper-parâmetros dos modelos dos grafos. Seja Θ uma variável aleatória que será amostrada para gerar os pa-

râmetros dos grafos aleatórios. Os parâmetros que determinam a distribuição de Θ são os hiper-parâmetros dos grafos aleatórios.

Como exemplo ilustrativo, suponha G^1 e G^2 dois grafos aleatórios. Em um grafo aleatório de Erdös-Rényi G de n vértices, cada par de vértices está conectado por uma aresta com uma dada probabilidade p. Neste caso, a probabilidade p é o parâmetro do grafo G. Assim, os grafos aleatórios de Erdös-Rényi podem ser descritos como $G^1(p^1)$ e $G^2(p^2)$, onde p^1 e p^2 são parâmetros amostrados de Θ^1 e Θ^2, respectivamente. Dizemos que uma série temporal de grafos G^1 Granger causa outra série temporal de grafos G^2 se existe causalidade de Granger de Θ^1 para Θ^2.

Uma forma simples de identificar causalidade de Granger de Θ^1 para Θ^2, se os modelos dos grafos são conhecidos, consiste em estimar os parâmetros dos modelos dos grafos e então aplicar o PDC sobre eles. No entanto, os modelos dos grafos raramente são conhecidos na prática. Assim, o problema consiste em detectar causalidade de Granger utilizando apenas os grafos aleatórios observados (e não os parâmetros). Em outras palavras, é necessário identificar uma característica do grafo que seja altamente correlacionada com os parâmetros do modelo do grafo.

Da teoria espectral de grafos, o maior autovalor (λ_1) do grafo G é conhecido como o raio espectral ou índice espectral (por conveniência, denotaremos a partir daqui, o raio espectral λ_1 somente por λ). Para diversos grafos aleatórios, é sabido que o raio espectral é uma função dos parâmetros do grafo. Por exemplo, para o grafo aleatório de Erdös-Rényi, sejam n e p o número de vértices e a probabilidade de dois vértices estarem conectados por uma aresta, respectivamente, então, o raio espectral do grafo aleatório de Erdös-Rényi é np. Para outros exemplos, ver seção 3.1. Veja na Figura 2.1 que de fato, para pelo menos os modelos de grafos aleatórios de Erdös-Rényi, geométrico, regular, de Barabási-Albert e de Watts-Strogatz, o raio espectral e o parâmetro do grafo estão altamente correlacionados. A primeira linha da figura mostra a associação entre os parâmetros dos modelos dos grafos aleatórios. As linhas seguintes são as associações entre os raios espectrais estimados a partir dos grafos gerados usando os parâmetros da linha um. Assim, propomos usar o raio espectral para identificar causalidade de Granger entre grafos.

Dado um conjunto de séries temporais de grafos, nossa proposta para identificar causalidade de Granger entre séries temporais de grafos, consiste em: (i) transformá-las em um conjunto de séries temporais de escalares (utilizando o raio espectral para resumir os grafos); e (ii), sobre o conjunto resultante, aplicar um método de identificação de causalidade de Granger.

Sejam $\underset{\sim}{G}^1 = \{G_1^1, G_2^1, \ldots, G_T^1, \}$ e $\underset{\sim}{G}^2 = \{G_1^2, G_2^2, \ldots, G_T^2, \}$ duas séries temporais de comprimento T de grafos aleatórios e $\underset{\sim}{\lambda}^1 = \{\lambda_1^1, \lambda_2^1, \ldots, \lambda_T^1\}$ e $\underset{\sim}{\lambda}^2 = \{\lambda_1^2, \lambda_2^2, \ldots, \lambda_T^2\}$ os raios espectrais associados a $\underset{\sim}{G}^1$ e $\underset{\sim}{G}^2$, respectivamente. Assim, para identificar causalidade de Granger entre essas duas séries temporais de grafos, basta testar a causalidade de Granger entre $\underset{\sim}{\lambda}^1$ e $\underset{\sim}{\lambda}^2$ utilizando o PDC.

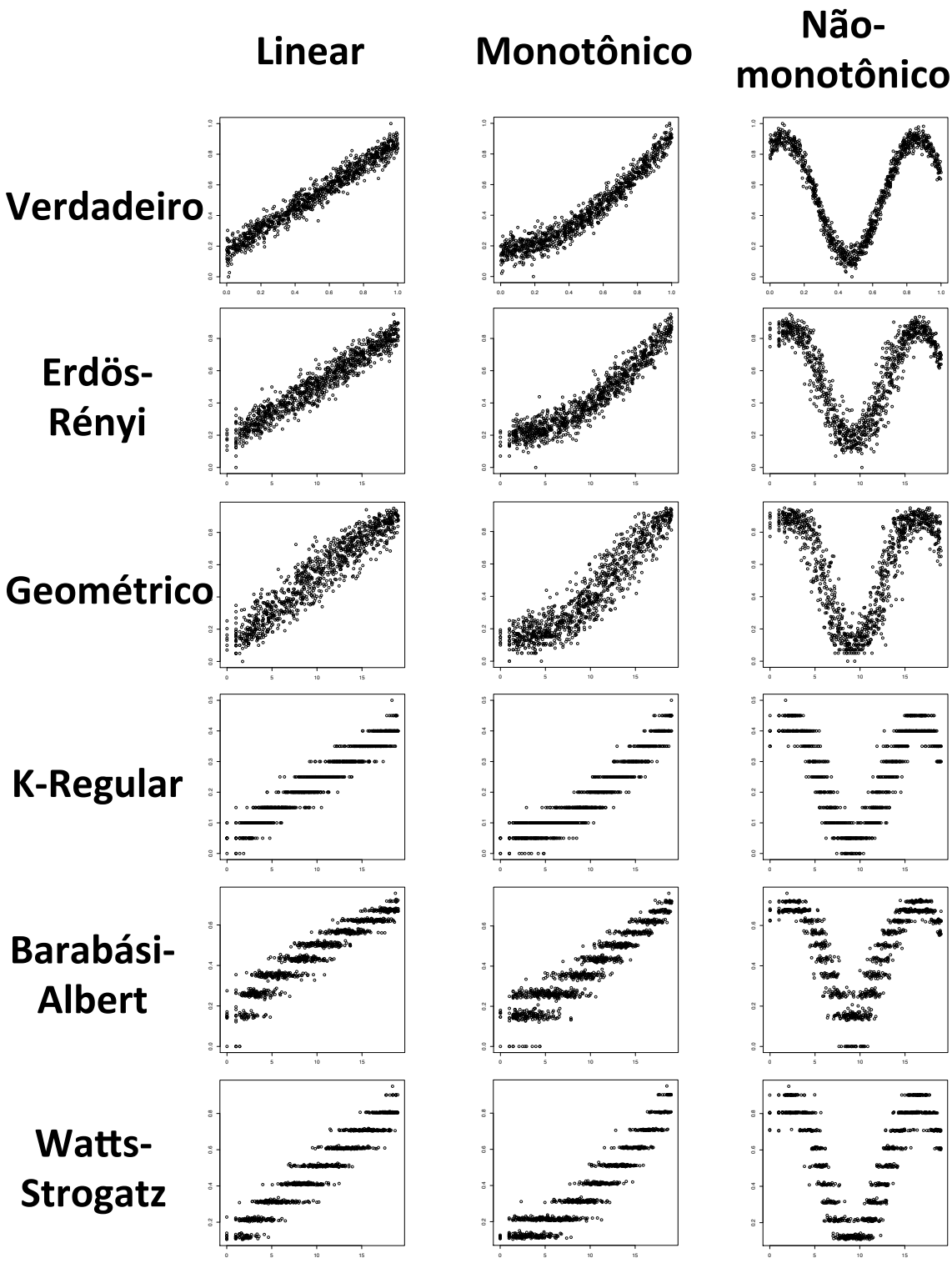


Figura 2.1: *Relação entre os raios espectrais. Simulamos dois vetores de parâmetros com associações lineares, monotônicas (exponencial) e não-monotônicas (senoide). Estes vetores foram usados como parâmetros dos modelos de grafos aleatórios. Após a construção dos grafos, o raio espectral para cada grafo foi calculado e o* scatterplot *construído. A primeira linha ilustra os* scatterplots *dos vetores originais dos parâmetros dos modelos dos grafos. As linhas seguintes mostram os* scatterplot *dos raios espectrais de diversos grafos como Erdös-Rényi, geométrico, regular e de Watts-Strogatz. Note que tanto para as relações linear, monotônico quanto não monotônico, o raio espectral parece recuperar bem a associação contida nos parâmetros do modelo dos grafos aleatórios.*

Capítulo 3

Simulações

A fim de avaliar o desempenho do modelo proposto em grafos, simulamos quatro tipos de grafos aleatórios (de Erdös-Rényi, geométrico, regular, de Barabási-Albert e de Watts-Strogatz) (seção 3.1) e comparamos nossa proposta de usar o raio espectral com outras características dos grafos como o número de arestas dos grafos e as medidas de centralidade descritos na seção 3.2. Executamos simulações de Monte Carlo em cinco cenários (seção 3.3) que representam diversas estruturas como ausência de causalidade de Granger, causalidade de Granger direta e indireta e laço. O desempenho do método é avaliado a partir do cálculo da área sob a curva ROC.

3.1 Modelos de grafos aleatórios

A seguir descreveremos alguns modelos geradores de grafos aleatórios. São eles: grafo aleatório de Erdös-Rényi (Erdös e Rényi, 1959), grafo aleatório geométrico (Penrose, 1999), grafo aleatório regular (Meringer, 1999), e grafo aleatório de Watts-Strogatz (Watts e Strogatz , 1998).

3.1.1 Modelo de grafo aleatório de Erdős-Rényi

Os grafos aleatórios propostos por Erdös-Rényi (Figura 3.1) são o modelo de grafos aleatórios mais estudados. Erdös e Rényi definiram um grafo aleatório como n vértices enumerados onde cada par de vértices (v_i, v_j) está conectado por uma aresta com uma dada probabilidade p (Erdös e Rényi, 1959).

O raio espectral de um grafo aleatório de Erdös-Rényi é np (Füredi e Komlós, 1981).

A função `R` usada para gerar um grafo aleatório de Erdös-Rényi é `erdos.renyi.game` (do pacote `igraph`[1]). O pacote `igraph` pode ser obtido na página da plataforma `R`[2].

[1] http://igraph.org/r/
[2] http://www.r-project.org

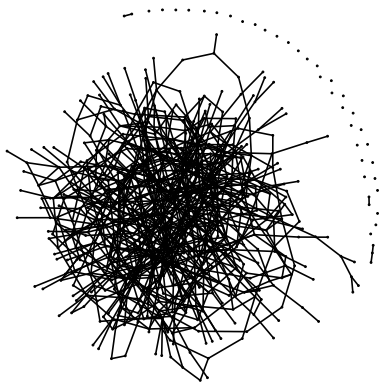


Figura 3.1: *Grafo gerado pelo modelo de Erdős-Rényi, com* $n = 600$ *vértices e* $p = 0.05$.

3.1.2 Modelo de grafo aleatório geométrico

Um grafo aleatório geométrico (Figura 3.2) é uma rede espacial. Um grafo não dirigido é construído distribuindo de forma aleatória n vértices em algum espaço topológico no R^d (por exemplo, no caso de um quadrado, $d = 2$) de acordo com uma distribuição de probabilidade específica (ex., uma distribuição uniforme) e conectando dois vértices por uma aresta se a distância entre eles (de acordo com uma métrica, por exemplo, a norma Euclidiana) é menor que um raio de vizinhança r. Então, grafos aleatórios geométricos apresentam um elemento espacial que está ausente nos outros modelos de grafos aleatórios (Penrose, 1999).

O raio espectral de um grafo geométrico converge quase que certamente para r^d (Bordenave , 2008).

A função `R` usada para gerar o grafo geométrico é `grg.game` (pacote `igraph`).

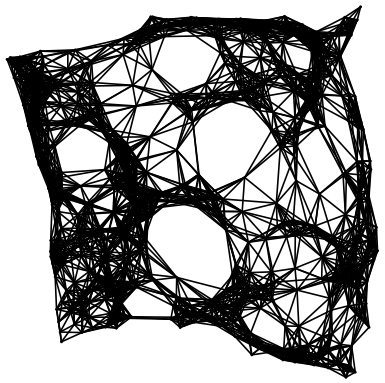


Figura 3.2: *Grafo gerado pelo modelo geométrico, com* $n = 600$ *vértices,* $d = 2$ *e* $r = 0.1$.

3.1.3 Modelo de grafo aleatório regular

Um grafo aleatório regular (Figura 3.3) é um grafo onde cada vértice tem o mesmo número de vértices adjacentes, isto é, todo vértice tem o mesmo grau. Um grafo aleatório regular com vértices de grau *deg* é chamado de grafo *deg*-regular ou grafo aleatório regular de grau *deg* (Meringer, 1999).

Grafos aleatórios regulares de grau no máximo dois são bem conhecidos: um grafo 0-regular consiste em vértices desconexos; um grafo 1-regular consiste de arestas desconexas; um grafo 2-regular consiste de ciclos desconexos; um grafo 3-regular é conhecido como o grafo cúbico.

O raio espectral de um grafo *deg*-regular é *deg* (Alon, 1986).

A função `R` usada para gerar o grafo aleatório regular é `k.regular.game` (pacote `igraph`).

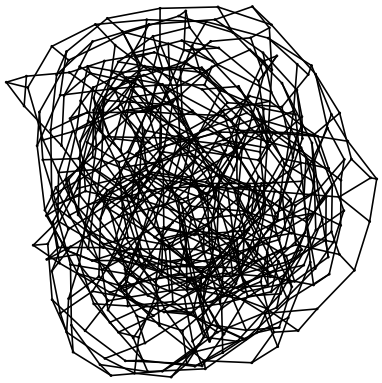


Figura 3.3: *Grafo gerado pelo modelo 3-regular, com* $n = 600$ *vértices.*

3.1.4 Modelo de grafo aleatório de Watts-Strogatz

O grafo aleatório de Watts-Strogatz (Figura 3.4) é um grafo aleatório que interpola entre um reticulado regular e um grafo aleatório de Erdös-Rényi (Watts e Strogatz, 1998). Este grafo apresenta propriedades de mundo pequeno (comprimento de médio dos caminhos bem pequeno, isto é, muitos dos vértices não são vizinhos um do outro mas podem ser alcançados a partir de qualquer vértice por um número pequeno de passos) e um coeficiente de agrupamento (o número de triângulos no grafo) maior que o grafo aleatório de Erdös-Rényi.

O algoritmo de construção de um grafo aleatório de Watts-Strogatz é como se segue:

Entrada: Sejam n, nei, e p_w o número de vértices, o número de vizinhos (grau médio), e a probabilidade de permutação das arestas, respectivamente.

1. construa um reticulado circular com n vértices, em que cada vértice é conectado aos seus primeiros nei vizinhos ($\frac{nei}{2}$ de cada lado);

2. escolha um vértice e uma aresta que conecta este vértice a seu vizinho mais próximo em sentido horário. Com probabilidade p_w, reconecte esta aresta com um vértice escolhido de forma uniforme a partir de todos os vértices. Este processo é repetido em sentido horário, até que todos os vértices sejam considerados. Depois, faça o mesmo para as arestas que conectam vértices a seus segundo-vizinhos mais próximos. Como no passo anterior, cada aresta é reconectada com probabilidade p_w. Continue este processo até que até que todas as arestas sejam consideradas.

Saída: o grafo aleatório de Watts-Strogatz.

Até onde pudemos verificar, o raio espectral do grafo aleatório de Watts-Strogats não está analiticamente definido, mas existem evidências empíricas que mostram que o raio espectral é uma função de p_w e nei (Van Mieghem, 2010).

A função `R` usada para gerar o grafo aleatório de Watts-Strogatz `watts.strogatz.game` (pacote `igraph`).

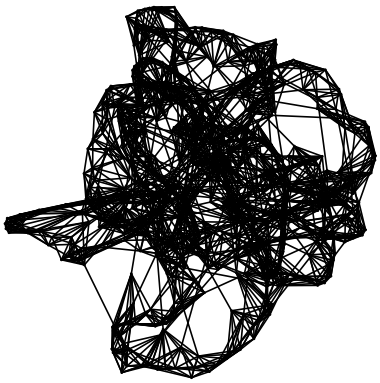


Figura 3.4: *Grafo gerado pelo modelo de Watts-Strogatz, com* $n = 600$ *vértices,* $nei = 5$ *e* $p_w = 0.03$.

3.1.5 Modelo de grafo aleatório de Barabási-Albert

Grafos aleatórios de Barabási-Albert (Figura 3.5) têm uma distribuição dos graus como um decaimento em potência devido à ligação preferencial (quanto mais o vértice está conectado, maior a chance de receber novas arestas) (Barabási e Albert, 1999). Barabási e Albert (1999) propuseram o seguinte algoritmo de construção do grafo:

Entrada: Sejam n, p_s, m_1 e n_0 o número de vértices, o expoente escalar, o número de arestas adicionadas a cada iteração e o número de vértices inicial, respectivamente.

1. comece com um número pequeno de vértices (n_0);
2. adicione um novo vértice com m_1 ($m_1 \leq n_0$) arestas que ligam este novo vértice à m_1 vértices diferentes já presentes no grafo. Na escolha do vértice v_i ao qual o novo vértice erá ligado, assuma que a probabilidade de ligação à v_i é proporcional ao grau de v_i e um expoente escalar p_s ($P(v_i) \sim grau(v_i)^{p_s}$, onde $grau(v_i)$ é o número de arestas adjacentes ao vértice v_i na iteração corrente) que indica a ordem da proporcionalidade ($p_s = 1$ linear; $p_s = 2$ quadrático e assim por diante).
3. volte ao passo 2 até que o número de vértices desejado seja atingido.

Saída: o grafo aleatório de Barabási-Albert

Seja k_0 o menor grau, o raio espectral de um grafo aleatório de Barabási-Albert é da ordem de $k_0^{1/2} n^{1/2(p_s-1)}$ (Dorogovtsev *et al.*, 2003).

A função `R` usada para gerar o grafo aleatório de Watts-Strogatz `barabasi.game` (pacote `igraph`).

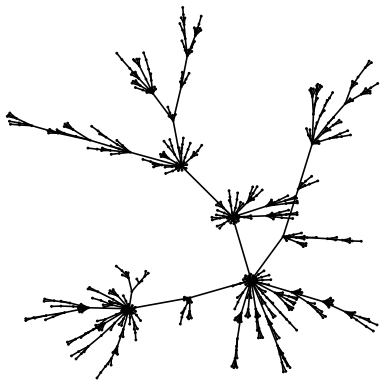


Figura 3.5: *Grafo gerado pelo modelo de Barabási-Albert, com $n = 600$ vértices, $m_1 = 1$, $n_0 = 1$ e $ps = 1.2$.*

3.2 Outras características dos grafos

Para avaliarmos o desempenho do nosso método proposto baseado em PDC no raio espectral, comparamos a nossa proposta com o PDC aplicado em outras medidas que podem ser extraídas de grafos, como por exemplo, o número de arestas, medidas de agrupamento e medidas de centralidade.

As medidas de centralidade de um vértice num grafo é a medida da "importância" relativa desse vértice no grafo. Naturalmente, a "importância" varia conforme o critério adotado.

Então adotamos quatro medidas de centralidade: proximidade, intermediação, autovetor e agrupamento. Como a centralidade é definida para cada vértice, utilizamos a média das centralidades como representativa do grafo todo.

3.2.1 Grau médio do vértice

O grau de um vértice de um grafo é o número de arestas incidentes no vértice. Aqui utilizamos como medida de um grafo o seu grau médio (ou seja, equivalente ao número de arestas).

$$C = \frac{m}{n} \tag{3.1}$$

lembrando que n e m são os números de vértices e arestas do grafo, respectivamente.

3.2.2 Centralidade de proximidade

A centralidade de proximidade de um vértice de um grafo conexo (um grafo em que todos os pares de vértices estão ligados por pelo menos um caminho) é o tamanho médio do caminho mais curto entre um vértice e todos os outros vértices no grafo. Assim, quanto mais central é o vértice, mais próximo ele é de todos os outros vértices.

Seja $d(v, v_j)$ o número de arestas do caminho mais curto entre os vértices v e v_j, a centralidade de proximidade definida por (Bavelas, 1950) é:

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

A fim de normalizar o valor da centralidade, o usual é dividir o valor obtido na equação 3.2.2 por $n - 1$, onde n é o número de vértices do grafo. Esta normalização permite a comparação entre vértices de grafos de tamanhos diferentes.

3.2.3 Centralidade de intermediação

A centralidade de intermediação quantifica o número de vezes que um vértice atua como uma "ponte" no caminho mais curto entre dois vértices. Esta centralidade foi introduzida para quantificar o controle de uma pessoa na comunicação entre outras pessoas numa rede social (Freeman, 1977).

A centralidade de intermediação de um vértice v de um grafo é computado como (Brandes , 2001):

1. Para cada par de vértices (v_i, v_j), identifique o caminho mais curto entre eles;
2. Para cada par de vértices (v_i, v_j), determine a fração de caminhos mais curtos que passam pelo vértice em questão (vértice v);
3. Some esta fração em todos os pares de vértices (v_i, v_j).

Em outras palavras, a centralidade de intermediação é dada por:

$$C(v) = \sum_{v_i \neq v \neq v_j \in V} \frac{\sigma_{v_i,v_j}(v)}{\sigma_{v_i,v_j}} \tag{3.2}$$

onde σ_{v_i,v_j} é o total do número de caminhos mais curtos do vértice v_i ao vértice v_j e $\sigma_{v_i,v_j}(v)$ é o número de caminhos que passam por v. A centralidade de intermediação pode ser normalizada pelos números de pares de vértices que não incluem o vértice v, ou seja, $(n-1)(n-2)/2$.

3.2.4 Centralidade de autovetor

A centralidade de autovetor de um vértice v é definida como:

$$C(v) = \frac{1}{\Lambda} \sum_{t \in M(v)} v_j = \frac{1}{\Lambda} \sum_{t \in G} a_{v,v_j} v_j \tag{3.3}$$

onde $M(v)$ é o conjunto de vizinhos de v e Λ é uma constante. A equação 3.3 pode ser reescrita como:

$$\mathbf{Ax} = \Lambda \mathbf{x}. \tag{3.4}$$

No geral, existem vários diferentes autovalores Λ que são soluções para os autovetores não-nulos. No entanto, para um autovetor não-negativo, pelo teorema de Frobenius, é sabido que apenas o maior autovalor resulta na medida de centralidade (Newman, 2008). A v-ésima componente do autovetor relacionado ao maior autovalor nos dá o nível da centralidade de autovetor. Para obter uma normalização, basta normalizar o autovetor de tal forma que a soma de todas as suas componentes seja um.

3.2.5 Coeficiente de agrupamento

O coeficiente de agrupamento é utilizado para identificar grupos de vértices densamente conectados. O coeficiente de agrupamento de um vértice v é o número de pares de vizinhos de v conectados entre si dividido pelo número de arestas que poderiam existir entre eles (Watts e Strogatz, 1998). Se $k(v)$ é o número de vizinhos de v, então o número máximo de arestas que poderiam existir entre os vizinhos de v é $k(v)(k(v)-1)/2$. Assim, podemos expressar o coeficiente de agrupamento de v por

$$C(v) = \sum_{v_j, v_k \in N(v)} \frac{\mathbf{A}_{jk}}{k(v)(k(v)-1),} \tag{3.5}$$

onde $N(v)$ denota a vizinhança de v e $\mathbf{A}_{jk}$ é a entrada correspondente à j-ésima linha e k-ésima coluna da matriz de adjacência.

Note que o coeficiente de agrupamento está entre 0 e 1, sendo que zero indica que não há conexões entre os vizinhos do vértice e 1 indica que todos os vértices da vizinhança estão conectados entre si.

3.3 Cenários

Aqui descreveremos cenários de inter-relação (em termos de causalidade de Granger) entre as séries temporais de grafos que foram simulados. Sejam $y_{i,t}$ com $i = 1, \ldots, 4$, as séries temporais de parâmetros geradores dos modelos dos grafos aleatórios, $\varepsilon_{i,t}$ os resíduos gerados a partir distribuições normais com média zero e $\Sigma = \mathbf{I}$. Os cenários estudados foram os seguintes:

3.3.1 Cenário 1

Com o objetivo de verificar o controle da taxa de falsos positivos, o cenário 1 descreve duas séries temporais independentes $y_{1,t}$ e $y_{2,t}$.isto é., sem causalidade de Granger entre elas, ou seja, sob a hipótese nula (Figura 3.6a).

$$\begin{cases} \mathbf{y}_{1,t} = 0.5 \times \mathbf{y}_{1,t-1} + \varepsilon_{1,t} \\ \mathbf{y}_{2,t} = 0.5 \times \mathbf{y}_{2,t-1} + \varepsilon_{2,t} \end{cases}$$

3.3.2 Cenário 2

Este cenário descreve uma causalidade de Granger direta da série temporal de $y_{1,t}$ para $y_{2,t}$ (Figura 3.6b).

$$\begin{cases} \mathbf{y}_{1,t} = 0.5 \times \mathbf{y}_{1,t-1} + \varepsilon_{1,t} \\ \mathbf{y}_{2,t} = 0.5 \times \mathbf{y}_{1,t-1} + \varepsilon_{2,t} \end{cases}$$

3.3.3 Cenário 3

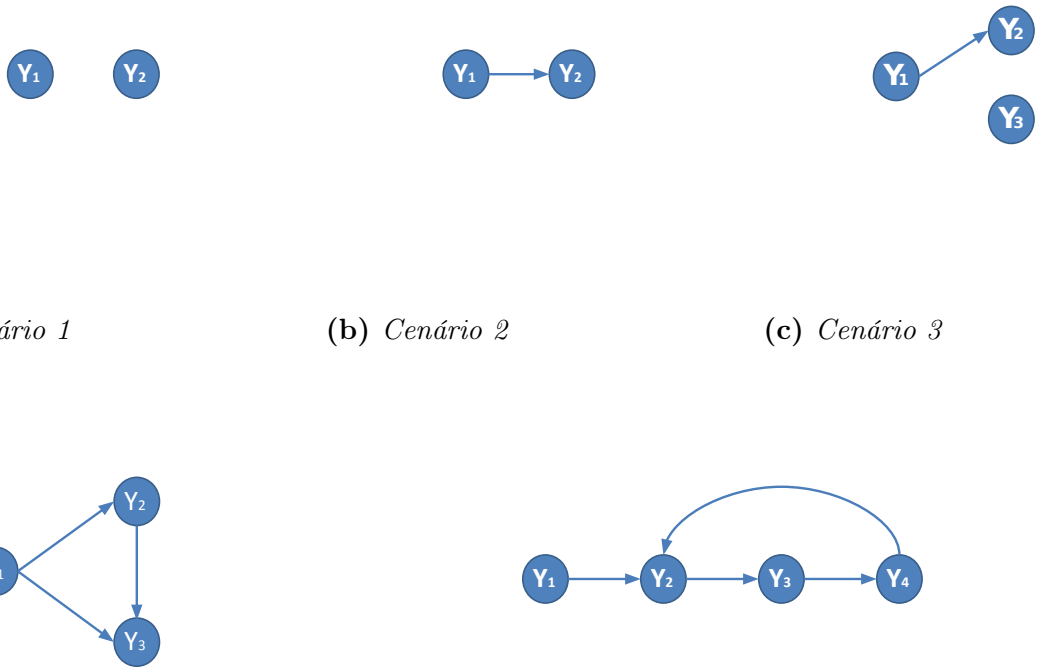
O cenário 3 representa três séries temporais com $y_{1,t}$ Granger causando $y_{2,t}$ (Figura 3.6c).

$$\begin{cases} \mathbf{y}_{1,t} = 0.5 \times \mathbf{y}_{1,t-1} + \varepsilon_{1,t} \\ \mathbf{y}_{2,t} = 0.5 \times \mathbf{y}_{1,t-1} + \varepsilon_{2,t} \\ \mathbf{y}_{3,t} = 0.5 \times \mathbf{y}_{3,t-1} + \varepsilon_{3,t} \end{cases}$$

3.3.4 Cenário 4

O quarto cenário descreve as seguintes causalidades de Granger: $y_{1,t} \to y_{2,t}$, $y_{2,t} \to y_{3,t}$ e $y_{1,t} \to y_{3,t}$ (Figura 3.6d).

$$\begin{cases} \mathbf{y}_{1,t} = \varepsilon_{1,t} \\ \mathbf{y}_{2,t} = 0.5 \times \mathbf{y}_{1,t-1} + \varepsilon_{2,t} \\ \mathbf{y}_{3,t} = 0.5 \times \mathbf{y}_{1,t-2} - 0.5 \times \mathbf{y}_{2,t-1} + \varepsilon_{3,t} \end{cases}$$



(a) *Cenário 1* **(b)** *Cenário 2* **(c)** *Cenário 3*

(d) *Cenário 4* **(e)** *Cenário 5*

Figura 3.6: *Representação gráfica dos cenários das simulações. Os círculos representam as séries temporais de grafos enquanto as arestas representam as direções da causalidade de Granger.*

3.3.5 Cenário 5

O quinto e último cenário descreve um *feedback* em termos de causalidade de Granger: $y_{1,t} \rightarrow y_{2,t}$, $y_{2,t} \rightarrow y_{3,t}$, $y_{3,t} \rightarrow y_{4,t}$ e $y_{4,t} \rightarrow y_{2,t}$ (Figura 3.6e).

$$\begin{cases} \mathbf{y}_{1,t} = \varepsilon_{1,t} \\ \mathbf{y}_{2,t} = 0.5 \times \mathbf{y}_{1,t-1} - 0.5 \times \mathbf{y}_{4,t-1} + \varepsilon_{2,t} \\ \mathbf{y}_{3,t} = -0.5 \times \mathbf{y}_{2,t-2} + \varepsilon_{3,t} \\ \mathbf{y}_{4,t} = 0.5 \times \mathbf{y}_{3,t-1} + \varepsilon_{4,t} \end{cases}$$

Após gerar as séries temporais como descritas nos cenários de 1 a 5, as séries temporais $(y_{i,t}, i = 1, \ldots, 4)$ foram linearmente normalizadas no intervalo entre zero e um e usadas como parâmetros dos modelos geradores de grafos. Para o modelo de grafo regular, consideramos a parte inteira do valor multiplicado por 10 (o parâmetro de um modelo de grafo regular é o grau de cada vértice). Cada cenário foi simulado 1.000 vezes para cada série temporal de comprimentos (número de grafos) $T = 50, 75, 100, 200, 300$. O tamanho dos grafos foi definido como $n = 30$. O número de reamostras bootstrap foi de 300.

3.4 Avaliação

Para avaliar e comparar o poder do teste ao aplicar o método PDC no raio espectral ou nas outras medidas de grafos, construímos curvas ROC (do inglês *receiver-operating characteristic*) e calculamos a área sob esta curva. A curva ROC é um desenho bidimensional com um menos a especificidade (número de falsos positivos dividido pela soma do número de verdadeiros negativos mais o número de falsos positivos) no eixo x e a sensitividade (número de verdadeiros positivos dividido pela soma do número de verdadeiros positivos mais o número de falsos negativos) no eixo y. Uma curva acima da diagonal indica um alto poder, enquanto uma curva próxima da diagonal indica decisões aleatórias. No nosso caso, o p-valor nominal do teste está no eixo x e a proporção de hipóteses nulas rejeitadas no eixo y. A área sob a curva ROC indica o poder estatístico do teste. Quanto maior a área sob a curva ROC, maior o poder. As áreas sob as curvas ROC foram calculadas a fim de (i) verificar o controle da taxa do erro tipo I; (ii) avaliar o poder do teste estatístico; e (iii) comparar o desempenho do PDC no raio espectral frente as outras medidas de grafos.

3.5 Resultados e Discussões

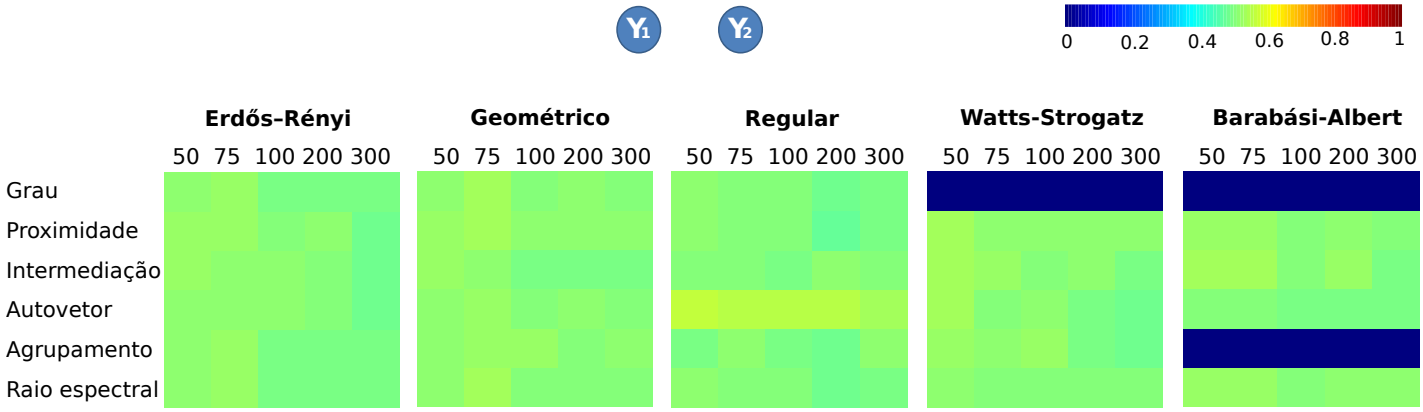
Analisando a Figura 3.7a (ausência de causalidade de Granger entre grafos), podemos verificar que, de fato, o teste estatístico baseado em bootstrap controla efetivamente a taxa do erro tipo I (taxa de falsos positivos). Note que as áreas abaixo das curvas ROC são próximas de 0.5, representando escolhas aleatórias.

A Figuras 3.7b-e ilustram o poder do teste em diferentes modelos de grafos aleatórios, comprimentos de série temporal e estruturas. No geral, as diversas medidas de grafos testadas aqui conseguem identificar causalidade de Granger no domínio da frequência. Contudo, são necessárias algumas observações. Tanto no modelo de grafo aleatório de Erdös-Rényi quanto no geométrico, todas as medidas foram capazes de identificar causalidade de Granger. No modelo de grafo aleatório regular, as medidas de centralidade de intermediação e autovetor apresentaram poder menor que as demais medidas. No modelo de grafo aleatório de Watts-Strogats, a medida de grau médio não foi capaz de detectar causalidade de Granger. Além disso, neste caso, apenas a raio espectral foi capaz de identificar causalidade de Granger de forma satisfatória. As demais medidas apresentaram baixo poder estatístico. Por último, no modelo de grafo aleatório de Barabási-Albert, as medidas de grau médio e coeficiente de agrupamento não foram capazes de detectar causalidade de Granger. A medida de centralidade proximidade apresentou alto poder estatístico, bem similar ao raio espectral.

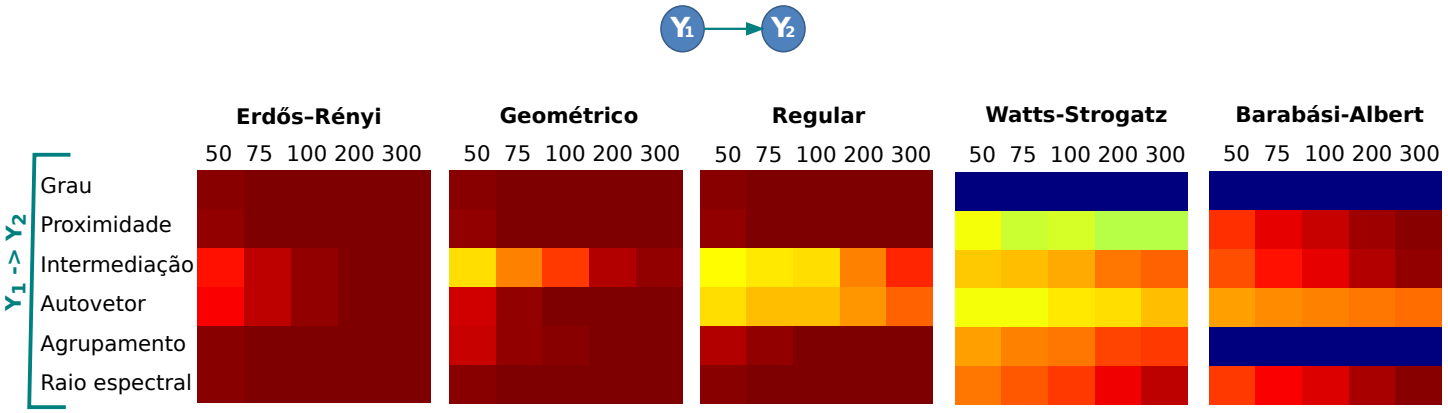
A dificuldade em se detectar causalidade de Granger no modelo de Watts-Strogats provavelmente está no fato da variável p_w que define o grau de permutação das arestas não estar associado com o número de arestas. Note que praticamente todas as medidas de grafos estão intimamente relacionadas com o número de arestas. As medidas que apresentaram

resultados nulos em termos de identificação de causalidade de Granger são devido a não-variação delas mesmo alterando o parâmetro do grafo. Por exemplo, nos modelos de grafos de Barabási-Albert e Watts-Strogats, o grau médio não varia. Note que no Barabási-Albert, a variável é o parâmetro p_s que indica a ordem da proporcionalidade, o que não tem relação com o número de arestas. No Watts-Strogats, o parâmetro p_w está relacionado somente a probabilidade de reconectar a aresta.

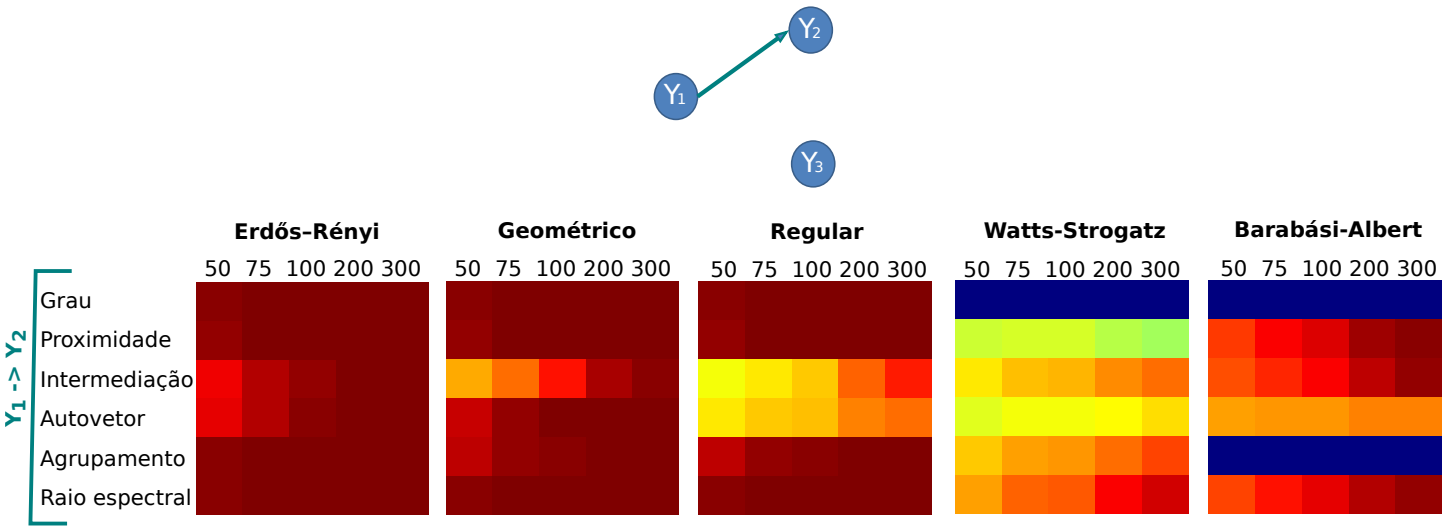
(a) *Cenário 1*



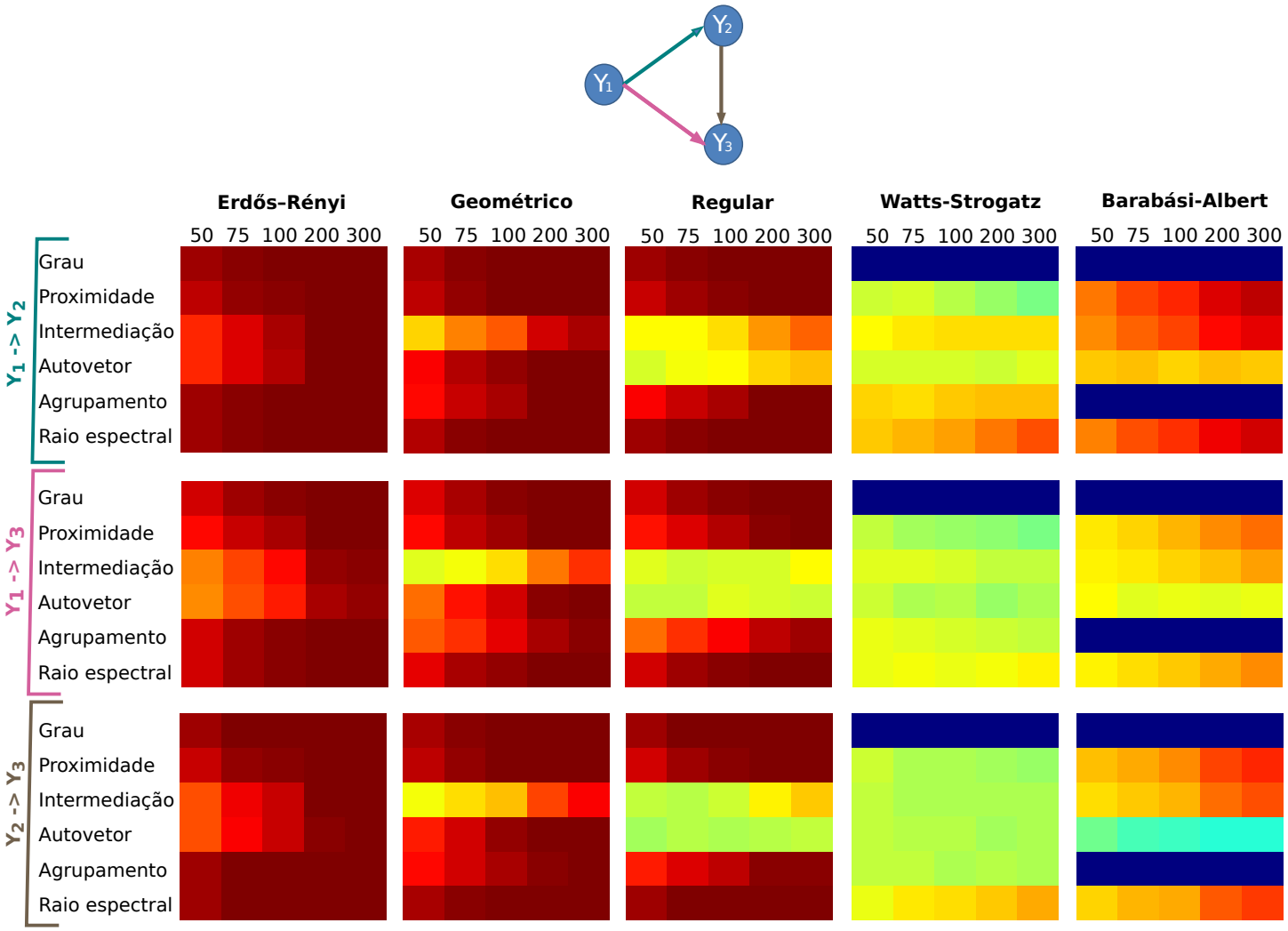
(b) *Cenário 2*



(c) *Cenário 3*



(d) *Cenário 4*


Y1
Y2
Y3
Erdős-Rényi
Geométrico
Regular
Watts-Strogatz
Barabási-Albert
50 75 100 200 300
Y1 -> Y2
Y1 -> Y3
Y2 -> Y3
Grau
Proximidade
Intermediação
Autovetor
Agrupamento
Raio espectral

(e) *Cenário 5*

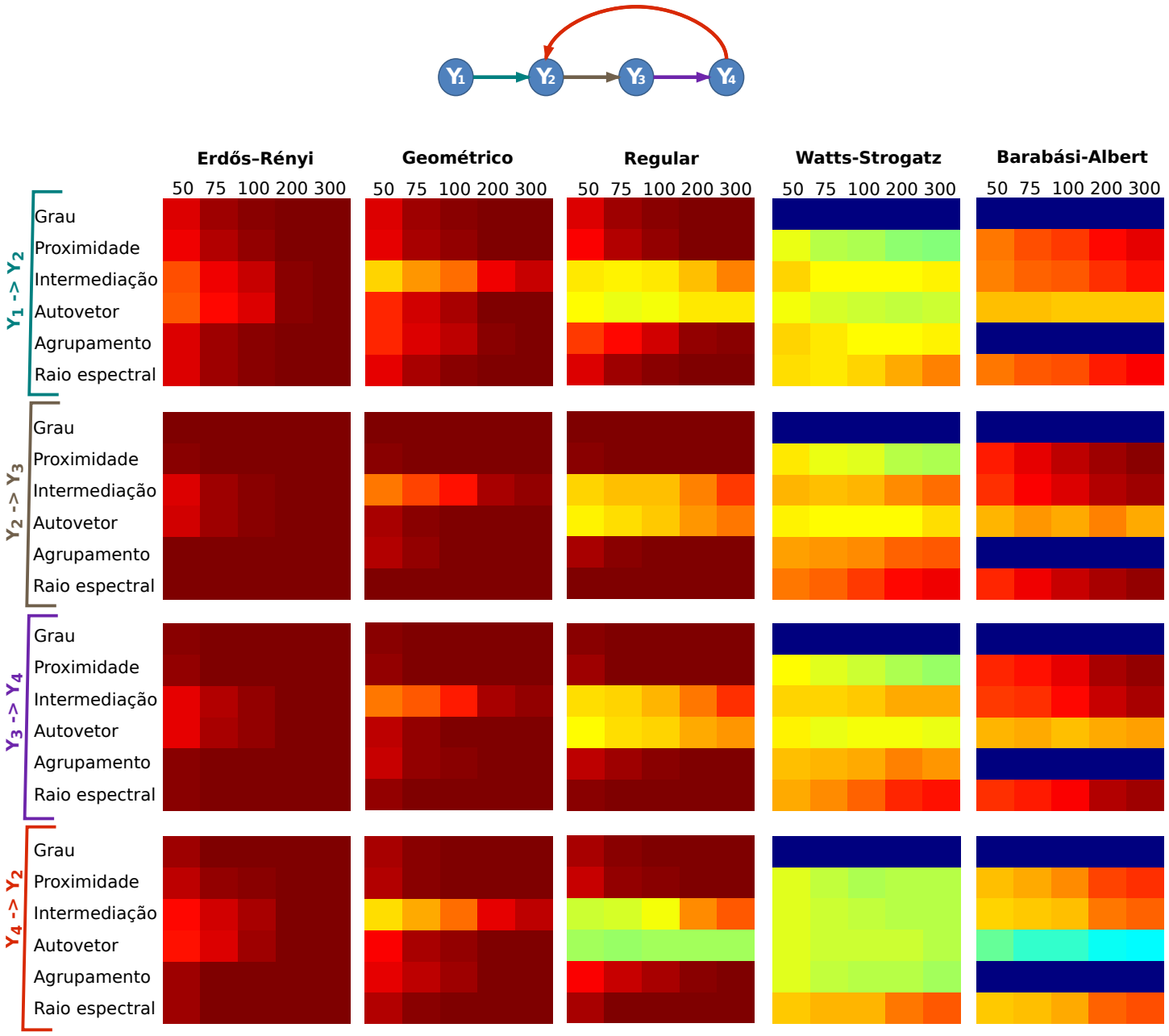


Figura 3.7: *Áreas debaixo das curvas ROC obtidas para cada cenário simulado. Em cada painel, exibimos um* heatmap *por modelo de grafo (Erdős-Rényi, Geométrico, Regular, Watts-Strogatz e Barabási-Albert), em que as linhas correspondem aos métodos utilizados para calcular a causalidade entre grafos (grau, centralidade de proximidade, centralidade de intermediação, centralidade de autovetor, coeficiente de agrupamento e raio espectral), as colunas correspondem aos tamanhos de séries temporais (50, 75, 100, 200, 300) e cada posição indica uma área debaixo da curva ROC, que é representada por uma cor. Áreas entre 0 e 0.5 são representadas por cores frias, enquanto áreas entre 0.5 e 1 são representadas por cores quentes, como indicamos à direita, na parte superior da imagem. Em (a) mostramos as áreas debaixo da curva ROC obtidas no cenário 1 (hipótese nula), onde testamos se y_1 Granger causa y_2. No painel (b) indicamos os resultados obtidos no cenário 2, onde testamos se y_1 Granger causa y_2. Em (c), são exibidos os* heatmaps *referentes aos testes realizados no cenário 3, para verificar se y_1 Granger causa y_2. No painel (d) mostramos as áreas debaixo das curvas ROC obtidas no cenário 4, onde testamos se y_1 Granger causa y_2 (primeiro agrupamento de* heatmaps*), se y_1 Granger causa y_3 (segundo agrupamento de* heatmaps*) e se y_2 Granger causa y_3 (terceiro agrupamento de* heatmaps*). Por fim em (c), são exibidos 4 agrupamentos de* heatmaps*, que correspondem (do primeiro ao último) às áreas de debaixo da curva ROC dos testes y_1 Granger causa y_2, y_2 Granger causa y_3, y_3 Granger causa y_4 e y_4 Granger causa y_2.*

Capítulo 4

Aplicação em dados reais

Neste capítulo ilustramos uma aplicação do método de identificação de causalidade de Granger estre grafos em dados reais. Escolhemos aplicar o método em neurociência em um estudo da conectividade funcional entre sub-regiões do cérebro.

Os dados consistem de séries temporais de eletrocorticografia (ECoG) de macaco. Redes funcionais do cérebro são construídas e agrupadas em sub-redes, onde cada sub-rede representa um grafo. Depois, nosso método é aplicado a fim de identificar fluxo de informação entre essas sub-redes.

4.1 Dados

Dados de eletrocorticografia (ECoG) (ver Apêndice, seção 6.1) de um macaco sob efeitos de anestesia foram obtidos na página do Neurotycho[1]. Cento e vinte e oito eletrodos de ECoG espaçados numa distância de 5 mm foram implantados no hemisfério esquerdo do macaco, cobrindo as áreas dos lobos frontal, parietal, temporal, e occipital.

O macaco encontra-se sentado calmamente com a cabeça e braços amarrados. Estes dados foram obtidos numa frequência de 1 kHz, a princípio em estado de alerta e posteriormente em um estado sob anestesia (Yanagawa *et al.*, 2013).

As características do experimento são as seguintes:

- Nome do macaco: Chibi

- Data do experimento: 13 de agosto de 2012

- Agente anestésico: Ketamina

- Número de sessões: 3

A série temporal está dividida da seguinte forma:

[1] http://neurotycho.org

Seção 1:

- 0.93 [s]: Acordado, olhos abertos - início
- 682.95 [s]: Acordado, olhos abertos - fim
- 2653.35 [s]: Acordado, olhos fechados - início
- 3352.28 [s]: Acordado, olhos fechados - fim

Seção 2:

- 21.82 [s]: Injeção da anestesia
- 676.11 [s]: Anestesiado - início
- 1250.12 [s]: Anestesiado - fim
- 2931.97 [s]: Recuperação, olhos fechados - início
- 3532.57 [s]: Recuperação, olhos fechados - fim

Seção 3:

- 84.95 [s]: Recuperação, olhos abertos - início
- 983.15 [s]: Recuperação, olhos abertos - fim

4.2 Pré-processamento

Em termos de pré-processamento, fizemos o *downsampling* dos dados de 1kHz de ECoG para 200Hz utilizando a função `decimate` do Matlab[2] e dividimos em janelas de dois segundos cada (400 pontos no tempo em cada janela). Para cada janela de dois segundos, uma rede de conectividade funcional do cérebro foi construída estimando uma correlação parcial de Spearman entre os 128 canais. A média dos z-valores para cada par de eletrodos foi calculado utilizando a média de todas as janelas no momento em que o macaco estava em alerta. O p-valor correspondente à média dos z-valores foi estimado e considerado como a matriz de dissimilaridade média do cérebro ao longo de toda a série temporal (para maiores informações sobre a correlação parcial de Spearman, veja Apêndice seção 6.2).

Vale ressaltar aqui que uma alternativa natural ao invés dessa abordagem seria estimar a correlação parcial usando-se a série temporal inteira. No entanto, o tamanho da série do macaco em estado de alerta é da ordem de 670.000 pontos no tempo (3.352 seg $\times$ 200Hz), o que torna seu cálculo inviável mesmo após o *downsampling* para 200Hz.

[2]https://www.mathworks.com/downloads/

Um menos a matriz de dissimilaridade foi considerada como a matriz de adjacência ponderada da rede funcional do cérebro. Sub-redes funcionais do cérebro foram obtidas aplicando o algoritmo de *clusterização* espectral (Ng *et al.*, 2002) na rede funcional média do cérebro. Note que o agrupamento do cérebro do macaco em estado de alerta foi usado como referência. O número de sub-redes foi obtida pelo uso da estatística *slope* (Fujita *et al.*, 2014). Para maiores detalhes da *clusterização* espectral e da estatística slope, ver Apêndice seções 6.3 e 6.4, respectivamente. Os rótulos de cada ROI (região de interesse) obtidos a partir da *clusterização* espectral aplicada nos dados do macaco em estado de alerta foram também usados nos dados do estado anestesiado.

O raio espectral (λ) foi calculado para cada sub-rede e cada janela de tempo, criando uma série temporal de escalares. Finalmente, o método proposto para identificar causalidade de Granger no domínio da frequência foi aplicado entre as sub-redes (grafos) tanto no macaco em estado de alerta quanto no estado anestesiado. Os p-valores corrigidos pelo método de Bonferroni para múltiplos testes (Dunn, 1959, 1961) foi aplicado a fim de identificar causalidades de Granger significativas entre as sub-redes. Para maiores detalhes sobre a correção de Bonferroni, ver Apêndice, seção 6.5

4.3 Resultados e Discussões

Pelo critério da estatística *slope*, o número de agrupamentos selecionado é cinco. No entanto, para cinco grupos a *clusterização* resultou em um grupo contendo apenas um único elemento. Assim, optamos por usar quatro grupos.

A Figura 4.1 representa os 128 eletrodos de ECoG posicionados no hemisfério esquerdo do cérebro do macaco. As cores representam as sub-redes obtidas pelo algoritmo de *clusterização* espectral aplicado na matriz de correlação parcial de Spearman.

Interessantemente, os sinais dos eletrodos foram agrupados de forma anatomicamente contínua mesmo não sendo fornecido nenhum tipo de informação *a priori* no algoritmo de *clusterização* espectral. O agrupamento resultante está consistente com a hipótese de que o algoritmo de *clusterização* espectral agrupa áreas cerebrais com atividades similares no mesmo grupo.

Para cada uma das sub-redes (cada cor da Figura 4.1), calculamos o raio espectral em cada janela de tempo, obtendo assim, uma série temporal de raios espectrais (grafos). Aplicamos então o algoritmo da coerência parcial direcionada nessas séries temporais. As interações de causalidade de Granger entre sub-redes obtidas antes e depois da injeção do anestésico estão descritas na Figura 4.2. As setas indicam a direção da causalidade de Granger no domínio da frequência em um corte de p-valor < 0.05, após correção para múltiplos testes por Bonferroni (Dunn, 1959, 1961). Note que, no caso do macaco anestesiado, existe uma perda de conectividade do lobo occipital com o lobo parietal temporal indicando que provavelmente o macaco não está usando a área visual, consistente com o macaco estar anestesiado.

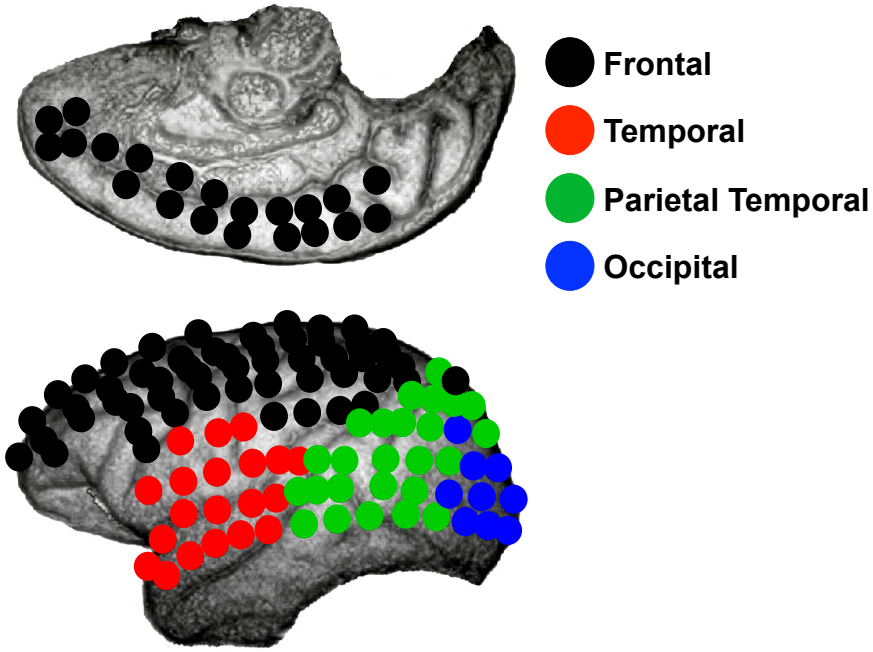


Figura 4.1: *Hemisfério esquerdo do cérebro do macaco agrupado antes da injeção do anestésico (estado de alerta). Cada círculo representa um eletrodo. As cores representam os grupos aos quais cada eletrodo foi inserido. Preto: área frontal. Vermelho: área temporal. Verde: área parietal temporal. Azul: área occipital.*

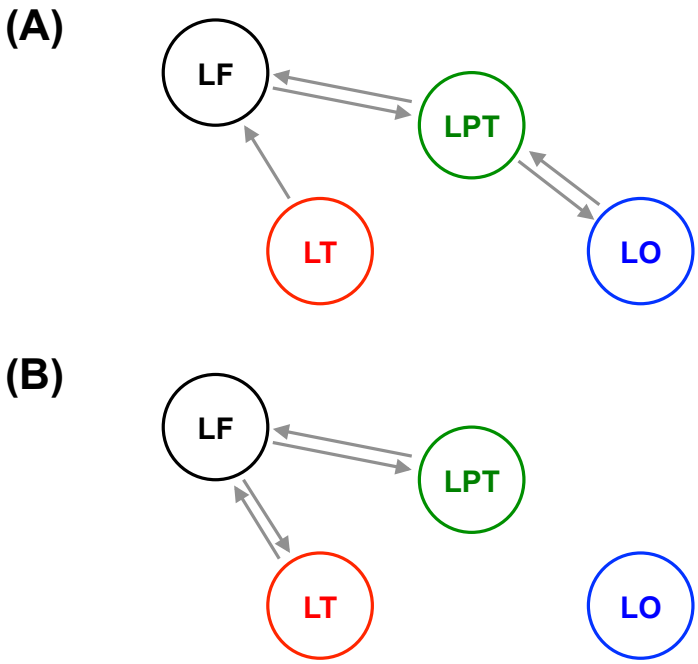


Figura 4.2: *Rede de redes. A direção das setas representa a direção da causalidade de Granger entre grafos/sub-redes. O painel (A) é a rede obtida nos dados de ECoG antes da aplicação da anestesia - sessão 1 (ketamina). O painel B é a rede obtida após a injeção do anestésico, ou seja, o macaco encontra-se sedado - sessão 2. LF: lobo frontal; LT: lobo temporal; LPT: lobo parietal temporal; LO: lobo occipital.*

Capítulo 5

Conclusões

A principal contribuição deste trabalho é um método de identificação de causalidade de Granger entre grafos no domínio da frequência. Este método mistura ideias de causalidade de Granger e análise espectral de grafos, provenientes das áreas da estatística e ciência da computação, respectivamente. O método apresentou alto poder estatístico em nossas simulações e também um controle efetivo da taxa do erro tipo I.

No entanto, vale discutir algumas limitações da proposta como também alguns trabalhos futuros relacionados ao tema. Nosso teste estatístico é baseado numa abordagem bootstrap, já que desconhecemos a distribuição assintótica do estimador sob a hipótese nula. A principal vantagem da abordagem bootstrap é que podemos obter uma estimativa do p-valor exato mesmo para dados finitos independente da distribuição dos dados. A desvantagem é que o procedimento é computacionalmente custoso, já que requer tempo linearmente proporcional ao número de reamostragens. Contudo, com os avanços no poder computacional e também da computação paralela, acreditamos que este não seja um gargalo real.

O método proposto é baseado no raio espectral, que parece conter a informação dos parâmetros do modelo do grafo gerador. Mas vale ressaltar que isto é verdade somente para alguns modelos de grafos aleatórios, como por exemplo, os grafos aleatórios de Erdös-Rényi, geométrico e regular. Para muitos outros modelos de grafos aleatórios, o raio espectral não é conhecido analiticamente. Além disso, até onde pudemos verificar, não existe resultado teórico a respeito do raio espectral para um caso mais geral de modelo de grafo aleatório.

Como trabalhos futuros, seria interessante mostrar de forma teórica a consistência do método ao menos para o modelo de grafo aleatório de Erdös-Rényi, que é um dos modelos mais bem estudados. Um outro projeto seria desenvolver uma metodologia para identificar causalidade de Granger mesmo para grafos dirigidos. Note que os grafos adotados nesta tese são todos não-dirigidos, ou seja, o maior autovalor é sempre real (pois a matriz de adjacências é simétrica). Infelizmente, pouco se é sabido sobre o espectro dos grafos dirigidos.

Por fim, acreditamos que esta proposta de PDC baseado no raio espectral seja promissora e esperamos que abra outras oportunidades de pesquisa na área de métodos estatísticos em grafos.

Capítulo 6

Apêndice

6.1 Eletrocorticografia - ECoG

Eletrocorticografia ou electroencefalografia cortical é um tipo de monitoramento electrofisiológico que usa eletrodos implantados diretamente na superfície do cérebro para realizar a leitura da atividade do córtex cerebral (Yang *et al.*, 2014), diferentemente da usual electroencefalografia (EEG) em que os eletrodos são posicionados do lado de fora do crânio. A grande vantagem do ECoG frente ao EEG é a pureza do sinal. O sinal de ECoG apresenta ordens de grandeza maiores de razão sinal / ruído. No entanto, a desvantagem é que os eletrodos do ECoG precisam ser implantados, ou seja, é um método invasivo.

6.2 Correlação parcial de Spearman

Sejam $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ um conjunto de n observações de duas variáveis aleatórias X e Y.

O teste de dependência entre X e Y é descrita como:
H_0: X e Y são não dependentes
versus
H_1: X e Y são dependentes.

Uma forma de calcular o coeficiente de correlação de Spearman (ρ) (Spearman, 1904) entre X e Y consiste em converter os valores de x_i e y_i $(i = 1, \ldots, n)$ para postos, calcular a diferença d_i entre os postos de x_i e y_i e finalmente calcular o coeficiente de correlação de Spearman como:

$$\rho(X, Y) = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{6.1}$$

O coeficiente de correlação de Spearman sob a hipótese nula pode ser assintoticamente aproximado por um distribuição t de Student com $n - 2$ graus de liberdade:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}. \tag{6.2}$$

O coeficiente de correlação de Spearman assume valores entre -1 e 1, onde $\rho(X,Y) = 1$ no caso de perfeita relação monotonicamente crescente e $\rho(X,Y) = -1$ no caso de perfeita relação monotonicamente decrescente. No caso de variáveis aleatórias monotonicamente independentes, $\rho(X,Y) = 0$. No caso de relações de dependências monotônicas imperfeitas, $-1 < \rho(X,Y) < 1$ (de Siqueira Santos *et al.*, 2013).

Agora, suponha que $\mathbf{V}$ seja o conjunto de todas as variáveis (inclusive contendo X e Y) de cardinalidade $|V|$ e queremos calcular a correlação de Spearman entre X e Y parcializado por todas as outras variáveis $\mathbf{V}\{X,Y\}$. Neste caso, basta criar uma matriz quadrada $|V|\times|V|$ contendo as correlações entre todas as variáveis duas a duas e invertê-la.

A função `R` usada para estimar o coeficiente de correlação parcial de Spearman foi implementada da seguinte forma:

```
## Entrada:
## x: matriz contendo as séries temporais nas colunas
##
## Saída
## res$correl: matriz de dimensões ncol(x) X ncol(x) contendo os
##             coeficientes de correlação parcial de Spearman
## res$pvalue: matriz de dimensões ncol(x) X ncol(x) contendo os
##             p-valores para a correlação parcial de Spearman
## res$tvalue: matriz de dimensões ncol(x) X ncol(x) contendo os
##             t-valores para a correlação parcial de Spearman
partial.correlation <- function(x, type="spearman") {
    R <- cor(x, method=c(type))
    Rinv <- qr.solve(R)
    D <- diag(1/sqrt(diag(Rinv)))
    P <- -D %*% Rinv %*% D
    diag(P) <- 1
    tvalue <- matrix(0, ncol(x), ncol(x))
    k <- ncol(x) - 2
    for (i in 1:(ncol(x)-1)) {
        for (j in (i+1):ncol(x)) {
            tvalue[i,j] <- ((sqrt(nrow(x)-k-2)*P[i,j])/(sqrt(1-P[i,j]^2))
            tvalue[j,i] <- tvalue[i,j]
        }
    }
    pvalue <- matrix(0, ncol(x), ncol(x))
```

```
    for (i in 1:(ncol(x)-1)) {
        for (j in (i+1):ncol(x)) {
            pvalue[i,j] <- 2*(1-pt(abs(tvalue[i,j]), (nrow(x)-k-2)))
            pvalue[j,i] <- pvalue[i,j]
        }
    }

    res <- list()
    res$correl <- P
    res$pvalue <- pvalue
    res$tvalue <- tvalue
    return(res)
}
```

6.3 Clusterização espectral

Uma das formas de se agrupar os vértices de um grafo de tal forma que vértices do mesmo grupo estejam mais conectados do que vértices de grupos diferentes é utilizando um algoritmo de *clusterização* espectral. Existem diversas variantes do algoritmo de *clusterização* espectral. Para um bom tutorial, ver (Von Luxburg, 2007).

O algoritmo de *clusterização* espectral usado aqui é descrito como (Ng *et al.*, 2002):

Entrada: Seja $\mathbf{A}^G$ a matriz de adjacências do grafo G e k o número de grupos.

1. Seja $\mathbf{D}$ uma matriz diagonal com os graus $d_1, d_2, \ldots, d_n$ dos vértices $v_1, v_2, \ldots, v_n$ respectivamente, na diagonal.
2. Compute a matriz Laplaciana $\mathbf{L} = \mathbf{D} - \mathbf{A}^G$.
3. Calcule os k autovetores $\mathbf{u}_1, \ldots, \mathbf{u}_k$ associados aos k menores autovalores de $\mathbf{L}$.
4. Seja $\mathbf{U} \in R^{n \times n}$ a matriz contendo os autovetores $\mathbf{u}_1, \ldots, \mathbf{u}_k$ como vetores colunas
5. Para $i = 1, \ldots, n$, seja $\mathbf{w}_i \in R^k$ o vetor correspondente à i-ésima linha de $\mathbf{U}$.
6. Agrupe os pontos $(\mathbf{w}_i)_{i=1,\ldots,n}$ com o algoritmo de k-medóides nas sub-redes $C_1, \ldots, C_k$.

Saída: Sub-redes $C_1, \ldots, C_k$.

É necessário ressaltar que o algoritmo da *clusterização* espectral original descrito por (Ng *et al.*, 2002) usa o algoritmo k-médias no passo 6. Aqui utilizamos o algoritmo k-medóides no lugar do usual k-médias porque o k-medóides é mais robusto a pontos aberran-

tes (*outliers*) e também apresenta soluções determinísticas dependendo da implementação. Nós utilizamos a implementação da função `pam` do `R` (Reynolds *et al.*, 2006) que apresenta soluções determinísticas.

6.4 Estatística slope

Seja $X = \{x_1, x_2, \ldots, x_n\}$ os dados com n elementos e seja $d(x_i, x_j)$ a distância entre x_i e x_j. Em nosso caso, a distância é dada pelo p-valor do coeficiente de correlação de Spearman entre as séries temporais x_i e x_j. Suponha que queiramos classificar os elementos de X em um dos k grupos $C^1, C^2, \ldots, C^k$.

Defina $d(x_i, B) = \frac{1}{|B|}\sum_{x\in B} d(x_i, x)$, a dissimilaridade média de x_i a todos os elementos de B, onde $|B|$ é o número de elementos de B. Denote por A o grupo ao qual x_i foi classificado pelo algoritmo de *clusterização* e C algum outro grupo diferente de A. Defina $a_i = d(x_i, A)$ e $b_i = \min_{C\neq A} d(x_i, C)$.

As quantidades a_i e b_i representam as dissimilaridades "dentro" do grupo e a menor "entre" os demais grupos, respectivamente. Então a estatística da silhueta do elemento x_i é dada por (Rousseeuw, 1987):

$$s_i = \begin{cases} \dfrac{b_i - a_i}{\max\{b_i, a_i\}}, & \text{se } |\mathrm{A}|{>}1 \\ 0, & \text{se } |\mathrm{A}|{=}1. \end{cases}$$

Para cada número de grupos $k = 2, 3, \ldots, n$, compute a estatística da silhueta como $s(k) = \frac{1}{n}\sum_{i=1}^{n} s_i$.

A parte interessante na estatística da silhueta é sua interpretação. Note que $-1 \leq s_i \leq 1$ e consequentemente existem três casos a serem analisados. Primeiro, quando $s_i \approx 1$, isto implica que a dissimilaridade "dentro" é muito menor que a dissimilaridade "entre" grupos, i.e., $a_i \ll b_i$. Isso significa que o elemento x_i está agrupado de forma adequada, ou seja, no seu grupo certo. O segundo caso ocorre quando $s_i \approx 0$. Neste caso, temos que $a_i \approx b_i$ e assim, não sabemos se o elemento x_i deveria ser classificado no grupo em que está ou no segundo grupo mais próximo, já que x_i encontra-se igualmente longe de ambos. O terceiro e último caso é quando $s_i \approx -1$. Neste caso, $a_i \gg b_i$ e assim, o elemento x_i foi classificado de forma errada, ou seja, deveria ter sido classificado como pertencente ao segundo melhor grupo do que no que foi classificado atualmente. Dada essas interpretações, pode-se notar que a estatística da silhueta mede o quão bem um elemento x_i foi classificado. A silhueta para o grafo inteiro $s(k)$ denota o quão bem todos os vértices do grafo foram em media, bem agrupados (Rousseeuw, 1987).

Brevemente, seja $s(k)$ a estatística da silhueta para k grupos e p um inteiro positivo. Então, o número de grupos k estimado pela estatística *slope* é dado por

$$\hat{k} = \underset{k\in\{2,\ldots,n-1\}}{\arg\max} -[s(k+1) - s(k)]s(k)^p \tag{6.3}$$

O parâmetro p serve para interpolar o critério onde o *gap* $s(k+1) - s(k)$ é mais importante (p pequeno) e um critério onde o valor da silhueta tenha maior peso (p grande). Para nossas análises, fixamos $p = 1$. A vantagem da estatística da silhueta frente a outros métodos é que ela é robusta a um grupo dominante.

A função `slope` usada neste trabalho está implementado `R` e foi obtida na página do autor[1].

6.5 Correção de Bonferroni

Sejam $H_1, H_2, \ldots, H_m$ uma família de hipóteses e $p_1, p_2, \ldots, p_m$ os correspondentes p-valores. A correção de Bonferroni consiste em, dada uma taxa de falsos positivos α, rejeitar as hipóteses nulas para todo $p_i \leq \frac{\alpha}{m}$ (Dunn, 1959, 1961).

[1] http://www.ime.usp.br/ fujita/software.html

Referências Bibliográficas

Alon (1986) Noga Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96. ISSN 0209-9683. doi: 10.1007/BF02579166. URL http://dx.doi.org/10.1007/BF02579166. Citado na pág. 15

Alon (2007) Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461. Citado na pág. 1

Baccalá e Sameshima (2001) Luiz A Baccalá e Koichi Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6): 463–474. Citado na pág. 1, 2, 8

Baccalá *et al.* **(2006)** Luiz Antonio Baccalá, Daniel Y Takahashi e Koichi Sameshima. 16 computer intensive testing for the influence between time series. *Handbook of time series analysis: Recent theoretical developments and applications*. Citado na pág. 9

Baccalá e Sameshima (1998) L.A. Baccalá e K. Sameshima. Directed coherence: a tool for exploring functional interactions among brain structures. Em M.A.L. Nicolelis, editor, *Methods for neural ensemble recordings*, páginas 179–192. CRC Boca Raton, Fla. Citado na pág. 8

Baccalá *et al.* **(1998)** L.A. Baccalá, K. Sameshima, G. Ballester, A.C. Valle e Timo-Iaria C. Studying the interaction between brain structures via directed coherence and granger causality. *Appl Signal Process*, 5:40,48. Citado na pág. 8

Barabási e Albert (1999) Albert-László Barabási e Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512. Citado na pág. 16

Bavelas (1950) Alex Bavelas. Communication patterns in task-oriented groups. *Journal of the acoustical society of America*. Citado na pág. 18

Bollobás e Riordan (2004) Béla Bollobás e Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34. Citado na pág. 1

Bordenave (2008) Charles Bordenave. Eigenvalues of euclidean random matrices. *Random Structures & Algorithms*, 33(4):515–532. Citado na pág. 14

Brandes (2001) Ulrik Brandes. A faster algorithm for betweenness centrality*. *Journal of mathematical sociology*, 25(2):163–177. Citado na pág. 18

Bullmore e Sporns (2009) Ed Bullmore e Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198. Citado na pág. 1

Cassidy *et al.* **(2016)** Clifford M Cassidy, Jared X Van Snellenberg, Caridad Benavides, Mark Slifstein, Zhishun Wang, Holly Moore, Anissa Abi-Dargham e Guillermo Horga. Dynamic connectivity between brain networks supports working memory: Relationships to dopamine release and schizophrenia. *The Journal of Neuroscience*, 36(15):4377–4388. Citado na pág. 2

de Siqueira Santos *et al.* **(2013)** Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata e André Fujita. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in bioinformatics*, página bbt051. Citado na pág. 36

Dorogovtsev *et al.* **(2003)** S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes e A. N. Samukhin. Spectra of complex networks. *Phys. Rev. E*, 68:046109. doi: 10.1103/PhysRevE.68.046109. URL http://link.aps.org/doi/10.1103/PhysRevE.68.046109. Citado na pág. 17

Dunn (1959) Olive Jean Dunn. Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics*, páginas 192–197. Citado na pág. 29, 39

Dunn (1961) Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64. Citado na pág. 29, 39

Eguiluz *et al.* **(2005)** Victor M Eguiluz, Dante R Chialvo, Guillermo A Cecchi, Marwan Baliki e A Vania Apkarian. Scale-free brain functional networks. *Physical review letters*, 94(1):018102. Citado na pág. 1

Erdös e Rényi (1959) Paul Erdös e Alfréd Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297. Citado na pág. 13

Freeman (1977) Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, páginas 35–41. Citado na pág. 18

Fujita *et al.* **(2007a)** André Fujita, João R Sato, Humberto M Garay-Malpartida, Rui Yamaguchi, Satoru Miyano, Mari C Sogayar e Carlos E Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1(1):1. Citado na pág. 2

Fujita *et al.* **(2007b)** André Fujita, João Ricardo Sato, Humberto Miguel Garay-Malpartida, Pedro Alberto Morettin, Mari Cleide Sogayar e Carlos Eduardo Ferreira. Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method. *Bioinformatics*, 23(13):1623–1630. Citado na pág. 2

Fujita *et al.* **(2008)** André Fujita, João Ricardo Sato, Humberto Miguel Garay-Malpartida, Mari Cleide Sogayar, Carlos Eduardo Ferreira e Satoru Miyano. Modeling nonlinear gene regulatory networks from time series gene expression data. *Journal of bioinformatics and computational biology*, 6(05):961–979. Citado na pág. 2

Fujita *et al.* **(2010a)** André Fujita, Kaname Kojima, Alexandre G Patriota, João R Sato, Patricia Severino e Satoru Miyano. A fast and robust statistical test based on likelihood ratio with bartlett correction to identify granger causality between gene sets. *Bioinformatics*, 26(18):2349–2351. Citado na pág. 2

Fujita *et al.* **(2010b)** André Fujita, Joao Ricardo Sato, Kaname Kojima, Luciana Rodrigues Gomes, Masao Nagasaki, Mari Cleide Sogayar e Satoru Miyano. Identification of granger causality between gene sets. *Journal of Bioinformatics and Computational Biology*, 8(04): 679–701. Citado na pág. 2

Fujita *et al.* **(2010c)** André Fujita, Patricia Severino, João Ricardo Sato e Satoru Miyano. Granger causality in systems biology: modeling gene networks in time series microarray data using vector autoregressive models. Em *Brazilian Symposium on Bioinformatics*, páginas 13–24. Springer. Citado na pág. 2

Fujita *et al.* **(2012)** André Fujita, Patricia Severino, Kaname Kojima, João Ricardo Sato, Alexandre Galvão Patriota e Satoru Miyano. Functional clustering of time series gene expression data by granger causality. *BMC systems biology*, 6(1):1. Citado na pág. 2

Fujita *et al.* **(2014)** André Fujita, Daniel Y Takahashi e Alexandre G Patriota. A non-parametric method to estimate the number of clusters. *Computational Statistics & Data Analysis*, 73:27–39. Citado na pág. 29

Fujita *et al.* **(2015)** André Fujita, Daniel Yasumasa Takahashi, Joana Bisol Balardin e João Ricardo Sato. Correlation between graphs with an application to brain networks analysis. *arXiv preprint arXiv:1512.06830*. URL http://arxiv.org/abs/1512.06830. Citado na pág. 3

Füredi e Komlós (1981) Zoltán Füredi e János Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241. Citado na pág. 13

Granger (1969) Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, páginas 424–438. Citado na pág. 1, 2, 5

Graybill (1976) Franklin A Franklin A Graybill. *Theory and application of the linear model*. Number 04; QA279, G7. Citado na pág. 8

Huberman e Adamic (1999) Bernardo A Huberman e Lada A Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131. Citado na pág. 1

Jeong *et al.* **(2000)** Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai e A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654. Citado na pág. 1

Lütkepohl (2011) Helmut Lütkepohl. *Vector autoregressive models*. Springer. Citado na pág. 2, 5, 6

Mangan e Alon (2003) Shmoolik Mangan e Uri Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100 (21):11980–11985. Citado na pág. 1

Maslov e Sneppen (2002) Sergei Maslov e Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913. Citado na pág. 1

Meringer (1999) Markus Meringer. Fast generation of regular graphs and construction of cages. *Journal of Graph Theory*, 30(2):137–146. ISSN 1097-0118. doi: 10.1002/(SICI)1097-0118(199902)30:2<137::AID-JGT7>3.0.CO;2-G. URL http://dx.doi.org/10.1002/(SICI)1097-0118(199902)30:2<137::AID-JGT7>3.0.CO;2-G. Citado na pág. 13, 14

Newman (2008) Mark EJ Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12. Citado na pág. 19

Ng *et al.* **(2002)** Andrew Y Ng, Michael I Jordan, Yair Weiss *et al.* On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856. Citado na pág. 29, 37

Penrose (1999) Mathew D Penrose. On k-connectivity for a geometric random graph. *Random Structures & Algorithms*, 15(2):145–164. Citado na pág. 13, 14

Reynolds *et al.* **(2006)** Alan P Reynolds, Graeme Richards, Beatriz de la Iglesia e Victor J Rayward-Smith. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504. Citado na pág. 38

Rousseeuw (1987) Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65. Citado na pág. 38

Rubinov e Sporns (2010) Mikail Rubinov e Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069. Citado na pág. 1

Saito e Harashima (1981) Y. Saito e H. Harashima. Tracking of information within multichannel eeg record - causal analysis in egg. Em N. Yamaguchi e K. Fujisawa, editors, *Recent Advances in EGG and EMG Data Processing*, páginas 133–146. Elsevier, Amsterdam. Citado na pág. 8

Sameshima e Baccalá (1999) Koichi Sameshima e Luiz Antonio Baccalá. Using partial directed coherence to describe neuronal ensemble interactions. *Journal of neuroscience methods*, 94(1):93–103. Citado na pág. 8

Sato *et al.* **(2009)** João R Sato, Daniel Y Takahashi, Silvia M Arcuri, Koichi Sameshima, Pedro A Morettin e Luiz A Baccalá. Frequency domain connectivity identification: an application of partial directed coherence in fmri. *Human brain mapping*, 30(2):452–461. Citado na pág. 2, 9

Sato *et al.* **(2010)** João R Sato, André Fujita, Elisson F Cardoso, Carlos E Thomaz, Michael J Brammer e Edson Amaro. Analyzing the connectivity between regions of interest: an approach based on cluster granger causality for fmri data analysis. *Neuroimage*, 52(4): 1444–1455. Citado na pág. 2

Sato *et al.* **(2006)** Joao Ricardo Sato, Edson Amaro Junior, Daniel Yasumasa Takahashi, Marcelo de Maria Felix, Michael John Brammer e Pedro Alberto Morettin. A method to produce evolving functional connectivity maps during the course of an fmri experiment using wavelet-based time-varying granger causality. *Neuroimage*, 31(1):187–196. Citado na pág. 2

Sato *et al.* **(2013)** João Ricardo Sato, Daniel Yasumasa Takahashi, Marcelo Queiroz Hoexter, Katlin Brauer Massirer e André Fujita. Measuring network's entropy in adhd: A new approach to investigate neuropsychiatric disorders. *NeuroImage*, 77:44–51. Citado na pág. 3

Sato *et al.* **(2015)** Joao Ricardo Sato, Maciel Vidal, Suzana de Siqueira Santos, Katlin Brauer Massirer e Andre Fujita. Complex network measures in autism spectrum disorders. Citado na pág. 3

Spearman (1904) Charles Spearman. "general intelligence,"objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292. Citado na pág. 35

Sporns (2013) Olaf Sporns. Network attributes for segregation and integration in the human brain. *Current opinion in neurobiology*, 23(2):162–171. Citado na pág. 2

Takahashi *et al.* **(2007)** Daniel Y Takahashi, Luiz Antonio Baccalá e Koichi Sameshima. Connectivity inference between neural structures via partial directed coherence. *Journal of Applied Statistics*, 34(10):1259–1273. Citado na pág. 2, 9

Takahashi *et al.* **(2010)** Daniel Y Takahashi, Luiz A Baccalá e Koichi Sameshima. Information theoretic interpretation of frequency domain connectivity measures. *Biological cybernetics*, 103(6):463–469. Citado na pág. 2

Takahashi *et al.* **(2012)** Daniel Yasumasa Takahashi, Joao Ricardo Sato, Carlos Eduardo Ferreira e André Fujita. Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PloS one*, 7(12):e49949. Citado na pág. 3

Van Mieghem (2010) Piet Van Mieghem. *Graph spectra for complex networks*. Cambridge University Press. Citado na pág. 3, 16

Von Luxburg (2007) Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416. Citado na pág. 37

Watts e Strogatz (1998) D.J. Watts e S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442. Citado na pág. 13, 15, 19

Yanagawa *et al.* **(2013)** Toru Yanagawa, Zenas C Chao, Naomi Hasegawa e Naotaka Fujii. Large-scale information flow in conscious and unconscious states: an ecog study in monkeys. *PloS one*, 8(11):e80845. Citado na pág. 27

Yang *et al.* **(2014)** Tong Yang, Shahin Hakimian e Theodore H Schwartz. Intraoperative electrocorticography (ecog): indications, techniques, and utility in epilepsy surgery. *Epileptic Disorders*, 16(3):271–279. Citado na pág. 35

Zuo *et al.* **(2012)** Xi-Nian Zuo, Ross Ehmke, Maarten Mennes, Davide Imperati, F Xavier Castellanos, Olaf Sporns e Michael P Milham. Network centrality in the human functional connectome. *Cerebral cortex*, 22(8):1862–1875. Citado na pág. 1