Teste de hipóteses para grafos aleatórios com aplicação à neurociência

Andressa Cerqueira

Dissertação apresentada AO Instituto de Matemática e Estatística DA Universidade de São Paulo Para Obtenção do título DE Mestre em Ciências

Programa: Estatística

Orientador: Prof^a. Dr^a. Florencia G. Leonardi

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CNPq Esta dissertação foi produzida como parte das atividades do Centro de pesquisa, inovação e disufão em Neuromatemática financiado pela FAPESP (concessão #2013/ 07699-0)

São Paulo, 24 de Fevereiro de 2014

Teste de hipóteses para grafos aleatórios com aplicação à neurociência

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 24/02/2014. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof^a. Dr^a. Florencia G. Leonardi (orientadora) IME-USP
- Prof. Dr. Anatoli Iambartsev IME-USP
- Prof. Dra Claudia Domingues Vargas UFRJ

"Um dia de chuva é tão belo como um dia de sol.

Ambos existem; cada um como é."

Fernando Pessoa

Agradecimentos

Em primeiro lugar, eu gostaria de agradecer aos meus pais Rogério e Elizabette pelo apoio dado à mim durante todos os meus anos de graduação e mestrado. Também agradeço meu irmão Carlos pelos conselhos e ensinamentos dignos de um irmão mais velho.

Gostaria de agradecer à minha orientadora Florencia Leonardi por tornar possível a construção desta dissertação e por todo o suporte e amizade ao longo desses anos. Você é em quem me inspiro como pesquisadora.

Quero agradecer à Claudia Vargas e ao Daniel Fraiman pelas inúmeras tardes de discussões que sem dúvidas contribuíram para o meu aprendizado e amadurecimento.

Agradeço a todos os meus amigos do IME e aos amigos com quem compartilho o local de trabalho no NUMEC. Em especial, quero agradecer aos meus amigos Ana Paula e William por sempre estarem ao meu lado durante o mestrado.

Quero deixar registrado os meus agradecimentos ao Guilherme pelo companheirismo e paciência durante todos os anos de mestrado. Obrigada por estar presente mesmo estando distante fisicamente.

Por fim, quero agradecer ao Antonio Galves pelas oportunidades concedidas a mim durante todo o meu mestrado e por sua confiança em meu trabalho.

Resumo

Cerqueira, A. Teste de hipóteses para grafos aleatórios com aplicação à neurociência. Dissertação - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014.

Recentemente, a teoria de grafos aleatórios vem sendo aplicada para modelar interações neurais do cérebro. Enquanto as propriedades dos grafos aleatórios vem sendo vastamente estudadas na literatura, o desenvolvimento de métodos de inferência estatística para essa classe de objetos tem recebido menos atenção. Nesse trabalho propomos um teste de hipóteses não paramétrico para testar se duas amostras de grafos aleatórios provém da mesma distribuição de probabilidade. Nós provamos como computar de maneira eficiente a estatística do teste e estudamos o desempenho do teste em dados simulados de grafos. A principal motivação deste trabalho é a aplicação do teste proposto em dados de eletroencefalograma.

Palavras-chave: teste de hipóteses, grafos aleatórios, dados de eletroencefalograma.

Abstract

Cerqueira, A. Test of hypotheses on random graphs with application in neuroscience. Dissertação - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014.

The theory of random graphs has been successfully applied in recent years to model neural interactions in the brain. While the probabilistic properties of random graphs has been extensively studied in the literature, the development of statistical inference methods for this class of objects has received less attention. In this work we propose a non parametric test of hypotheses to decide if two samples of random graphs are originated from the same probability distribution. We show how to compute efficiently the test statistic and we study the performance of the test on simulated data. The main motivation of this work is to apply this test to analyze neural networks constructed from electroencephalographic data.

Keywords: test of hypotheses, random graphs, electroencephalographic data.

Sumário

Li	sta d	le Figuras	хi
Li	sta d	le Tabelas x	iii
1		rodução	1
	1.1	Objetivos	2
	1.2	Organização do trabalho	2
2	Mo	delos de grafos aleatórios	5
	2.1	Grafos	5
		2.1.1 Definições	6
	2.2	Modelos de grafos aleatórios	8
		2.2.1 Erdős-Rényi	8
		2.2.2 Condicionalmente Uniforme	9
		2.2.3 Distribuição livre de escala	10
		2.2.4 Distribuição do grau fixada	10
		2.2.5 Exponencial	11
		2.2.6 Mundo Pequeno	13
3	Tes	te de hipóteses para grafos aleatórios	15
	3.1	Teste de hipóteses	15
		3.1.1 Introdução	15
		3.1.2 Testes paramétricos e não paramétricos	16
	3.2	Teste de hipóteses não paramétrico para grafos	17
		3.2.1 Definições	17
		3.2.2 Teste de hipóteses para uma amostra única de grafos	17
		3.2.3 Teste de hipóteses para duas amostras de grafos	21
4	Sim	ulações	25
	4.1	Simulação de grafos aleatórios	25
			25
			26
	4.2		27
			 27
		4.2.2 Poder do teste de hipóteses	
		•	29

x SUMÁRIO

	4.3	Teste	de hipóteses para duas amostras de grafos	. 31
		4.3.1	Distribuição empírica de W	. 31
		4.3.2	Taxas de rejeição do teste	. 31
		4.3.3	Poder do teste de hipóteses	. 35
5	\mathbf{Apl}	licação		41
	5.1	Dados		41
	5.2	Const	rução dos grafos de interações	43
	5.3	Result	ados do teste de hipóteses para grafos e discussão	47
		5.3.1	Teste de hipóteses para as condições biológico e não biológico	. 48
		5.3.2	Teste de hipóteses para as fases visível e oclusão	. 52
6	Cor	nclusõe	es ·	57
\mathbf{A}	Apé	êndice		59
	A.1	Progra	amas em R	. 59
$\mathbf{R}_{m{\epsilon}}$	eferê	ncias I	Bibliográficas	63

Lista de Figuras

2.1	Representação dos grafos	6
2.2	Grafos construídos a partir da sequência de graus $(gr(1),gr(2),gr(3),gr(4))=(2,1,1,2)$) 7
2.3	Exemplos de grafos desconexos	7
2.4	Grafos construídos a partir da sequência de graus $(gr(1),gr(2),gr(3),gr(4),gr(5)) =$	
	(3,1,2,2,2)	11
4.1	Distribuição empírica da estatística W , sob H_0 , utilizando o modelo Exponencial	
	areta-triângulo para $\boldsymbol{\theta}_0 = (0.3, 0.08)$	29
4.2	Poder do teste de hipóteses para uma amostra de grafos do modelo Exponencial	
	aresta-triângulo	30
4.3	Amostra de quatro grafos do modelo Exponencial aresta-triângulo	32
4.4	Amostra de três grafos do modelo Exponencial aresta-2estrelas	36
4.5	Distribuição empírica da estatística W, sob a hipótese nula e alternativa, para duas	
	amostras de grafos Erdős-Rényi	37
4.6	Poder do teste obtido para amostras de grafos com 5 vértices do modelo de Erdős-	
	Rényi. A primeira amostra possui parâmetro fixado igual a 0.7 e a segunda amostra	
	possui o parâmetro q variável	38
4.7	Poder do teste obtido para amostras de grafos do modelo de Erdős-Rényi. A primeira	
	amostra foi gerada pelo modelo com parâmetro fixado e igual a 0.7 e o parâmetro q	
	da segunda amostra variável	38
4.8	Poder do teste obtido para amostras de grafos com 5 vértices do modelo de Erdős-	
	Rényi com uma aresta de diferença	39
5.1	Estímulo utilizado no experimento	42
5.2	Representação da posição dos eletrodos no escalpo do indivíduo	42
5.3	Correlações ao longo das janelas temporais entre os eletrodos Pz e Oz para a fase	
	visível e oclusão do movimento biológico	44
5.4	Correlações ao longo das janelas temporais entre os eletrodos Pz e Oz para a fase	
	visível do movimento biológico para diferentes repetições	45
5.5	Correlações entre os eletrodos F7 e T6 das janelas temporais da terceira repetição	
	do experimento para a fase da oclusão do movimento biológico	46
5.6	Comparação entre as regiões de rejeição do teste de hipóteses	49
5.7	Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da	
	janela 1 da fases visível para o movimento biológico e não biológico	50

5.8	Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da	
	janela 41 da fases visível para o movimento biológico e não biológico	50
5.9	Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da	
	janela 61 da fases oclusão para o movimento biológico e não biológico	51
5.10	Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da	
	janela 81 da fases oclusão para o movimento biológico e não biológico	52
5.11	Comparação entre as regiões de rejeição do teste de hipóteses	53
5.12	Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da	
	janela 1 das fases visível e oclusão para o movimento biológico	54
5.13	Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da	
	janela 1 das fases visível e oclusão para o movimento não biológico	54
5.14	Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da	
	janela 61 das fases visível e oclusão para o movimento biológico	55

Lista de Tabelas

2.1	Exemplos de estatísticas utilizadas para descrever grafos simples	12
3.1	Possíveis erros de decisão que podem ocorrer em um teste de hipóteses	15
4.1	Taxas de rejeição do teste proposto para uma amostra de grafos do modelo aresta-	0.1
4.0	triângulo	31
4.2	Taxas de rejeição do teste proposto para duas amostras de grafos do modelo de	00
	Erdős-Rényi	33
4.3	Taxas de rejeição dos múltiplos testes para duas amostras de grafos do modelo de	
	Erdős-Rényi	33
4.4	Taxas de rejeição dos teste proposto para duas amostras de grafos do modelo de	
	Erdős-Rényi e do modelo que difere em apenas uma aresta do Erdős-Rényi	34
4.5	Taxas de rejeição dos múltiplos testes para duas amostras de grafos do modelo de	
	Erdős-Rényi e do modelo que difere em apenas uma aresta do Erdős-Rényi	34
4.6	Taxas de rejeição do teste proposto para duas amostras de grafos do modelo Expo-	
	nencial aresta-2estrelas	35
5.1	Número de repetições do experimento	43
5.2	Janelas temporais em ms	43
5.3	Média do valor absoluto das correlações	47
5.4	Média do número de arestas das amostras de grafos	47
5.5	P-valor do teste de hipóteses para amostras de grafos das fases visível do movimento	
	biológico e não biológico e para amostras de grafos para a fase oclusão para o movi-	
	mento biológico e não biológico	48
5.6	P-valor do teste de hipóteses para amostras de grafos das fases visível e oclusão para	
	o movimento biológico e não biológico	53

Capítulo 1

Introdução

As redes que descrevem interações entre indivíduos ou componentes de um sistema vem sendo vastamente estudadas pela comunidade científica. O foco das pesquisas está relacionado à aplicações da teoria estatística em redes reais, sendo essas sociais, biológicas ou de informação. Essas redes de interações podem ser representadas por meio de grafos, onde os objetos de estudo e suas interações são representados pelos vértices e arestas do grafo, respectivamente.

Um grafo consiste em um conjunto (V, A) em que V é o conjunto de vértices e A é o conjunto de arestas formado por pares ordenados (grafo direcionado) ou não ordenados (grafos não direcionados) de vértices.

As redes de informação podem ser representadas e estudadas por meio de grafos, por exemplo, em sistemas de internet os computadores e roteadores podem ser descritos por vértices do grafo e os cabos ou conexões *wireless* por arestas. As redes presentes na internet também podem ser estudadas por meio da estrutura de grafos, onde as páginas e os hiperlinks representam os vértices e arestas dos grafos, respectivamente.

As redes sociais descrevem as interações presentes entre indivíduos, sendo elas as redes de citações em artigos científicos, redes de interação econômica ou comercial e interações de amizade. Os estudos envolvendo redes sociais têm recebido grande atenção da comunidade científica em estatística, pois o modelo mais utilizado para descrever esse tipo de rede é o modelo Exponencial para grafos. Esse modelo é vastantamente estudado no ponto de vista estatístico, tanto referente ao problema de estimação dos parâmetros quanto para a simulação do modelo.

As reações químicas presentes em uma célula, a predação entre animais e as interações neurais são exemplos de redes biológicas que podem ser descritas por meio de grafos (Newman, 2010). As redes neurais são o foco do estudo realizado neste trabalho.

Para compreender a estrurura de redes de informação, sociais ou biológicas, do ponto de vista estatístico, é possível inferir características dos grafos que descrevem essas redes, selecionar modelos probabilíticos que descrevam dados provenientes de redes, comparar grafos e predizer funções a partir dos grafos (Caimo e Friel, 2011).

A partir de dados provenientes de Eletroencefalograma (EEG) ou de ressonância magnética funcional (fMRI) é possível representar as interações cerebrais por meio de grafos. Primeiramente, defini-se o conjunto de vértices do grafo, que pode ser tomado como o conjunto de eletrodos para dados provenientes de EEG ou regiões do cérebro para dados de fMRI. A matriz de associação do grafo pode ser obtida através de medidas de associação entre pares de vértices do grafo e, a partir de um valor limiar dessa matriz, a matriz de adjacência do grafo é determinada (Bullmore e Sporns,

2 INTRODUÇÃO 1.2

2009).

No trabalho de Wang et al. (2010) foram estudados dados de fMRI de pacientes que sofreram derrame cerebral por meio de grafos. Esses estudos mostraram evidências de mudanças nas interações das redes cerebrais durante a recuperação dos pacientes que sofreram derrame cerebral, uma vez que os grafos construídos a partir de dados de fMRI desses pacientes apresentam diferentes características ao longo de sua recuperação.

Dessa forma, uma vez que os grafos referentes a um determinado conjunto de dados cerebrais (sendo eles provenientes de fMRI ou EEG) são obtidos há interesse em verificar se os grafos apresentam diferenças significativas, ou seja, se as interações cerebrais apresentam diferenças para diferentes estímulos ou situações.

Nos trabalhos Wang et al. (2010), Fraiman et al. (2014) e Wu et al. (2012) a partir dos grafos obtidos são utilizadas propriedades locais e globais desses grafos, como coeficiente de agrupamento e tamanho do caminho do grafo, para analisar as diferenças entre eletrodos, ou regiões do cérebro, para diferentes condições dos indivíduos em estudo.

Na literatura, um método utilizado para estudar as propriedades dos grafos obtidos à partir de dados reais é comparar a distribuição de uma propriedade do grafo predefinida com a distribuição dessa mesma propriedade para grafos com distribuição de Erdős-Rényi. Dito isso, o coeficiente de agrupamento e a menor distância entre dois vértices do grafo são propriedades vastamente estudas, pois essas duas propriedades são importantes para avaliar se os grafos em estudo apresentam caracterítica de redes de mundo pequeno. Essas propriedades são discutidas mais extensivamente no Capítulo 2.

1.1 Objetivos

O objetivo deste trabalho é propor um novo método para analisar amostras de grafos aleatórios. Para isso, propromos um teste de hipóteses não paramétrico para testar se duas amostras de grafos aleatórios são provenientes de uma mesma distribuição de probabilidade sobre grafos. Para uma única amostra de grafos deseja-se testar se essa amostra é proveniente da distribuição de probabilidade especificada pela hipótese nula.

Esta dissertação também tem como objetivo aplicar o teste de hipóteses proposto em dados de eletroencefalograma para comparar as interações cerebrais de indivíduos que são submetidos à diferentes estímulos visuais.

1.2 Organização do trabalho

Inicialmente, no Capítulo 2 são apresentados alguns conceitos relacionados à teoria de grafos. Além disso, os principais modelos de grafos aleatórios e os estudos realizados para esses modelos são discutidos nesse Capítulo.

No Capítulo 3 é proposto um teste de hipóteses não paramétrico para uma amostra de grafos para testar se a distribuição da qual a amostra de grafos foi extraída é a mesma distribuição definida na hipótese estatística. O teste de hipóteses para verificar se duas amostras de grafos foram originadas da mesma distribuição também é proposto nesta dissertação. Ainda nesse Capítulo, alguns resultados assintóticos referentes à estatística do teste são apresentados e provados.

Para verificar o desempenho dos testes de hipóteses propostos nesta dissertação são utilizados grafos simulados dos modelos de Erdős-Rényi e Exponencial e os resultados do poder do teste e das taxas de rejeição do teste são apresentados no Capítulo 4.

No Capítulo 5 o teste de hipóteses é utilizado para duas amostras de grafos obtidas a partir de dados de eletroencefalograma de indivíduos que foram submetidos à diferentes estímulos visuais. O teste de hipóteses é utilizado para testar se a distribuição das amostras de grafos obtidas para cada estímulo são iguais, ou seja, se as interações presentes entre os eletrodos são as mesmas para diferentes estímulos.

Por fim, no Capítulo 6 são feitas as considerações finais a respeiro dos resultados obtidos nesta dissertação e os possíveis trabalhos futuros a serem desenvolvidos.

4 INTRODUÇÃO 1.2

Capítulo 2

Modelos de grafos aleatórios

Este capítulo introduz a definição de grafo e os conceitos de grau, caminho, caminho geodésico e coeficiente de agrupamento. Na Seção 2.2 é feita uma revisão dos principais modelos de grafos aleatórios abordados e estudados na literatura.

2.1 Grafos

Um grafo consiste em um par (V,A) em que V é o conjunto de vértices e A é o conjunto de arestas que conectam pares de vértices. Quando dois vértices são conectados por uma aresta, ou seja, para $i,j \in V$ temos que $(i,j) \in A$, eles são ditos vértices adjacentes.

As arestas de um grafo podem ser direcionadas ou não direcionadas. No primeiro caso, o conjunto A é um conjunto de pares ordenados, ou seja, se existe uma aresta de i para j não implica a existência de uma aresta de j para i, assim $(i,j) \neq (j,i)$. No segundo caso, o conjunto A é um conjunto de pares não ordenados, ou seja, (i,j) = (j,i). À aresta que conecta um vértice a ele mesmo damos o nome de laço.

Quando $A \subseteq V \times V$ é um conjunto de pares não ordenados sem a existência de laços dizemos que o grafo é simples. Por outro lado, se A é um conjunto de pares ordenados temos um dígrafo.

Os grafos que possuem vértices conectados por mais de uma aresta, chamadas arestas múltiplas, são ditos multigrafos. Neste caso, o conjunto A é chamado de multiconjunto de pares não ordenados. Quando A é um multiconjunto de pares ordenados os grafos são chamados de multidígrafos.

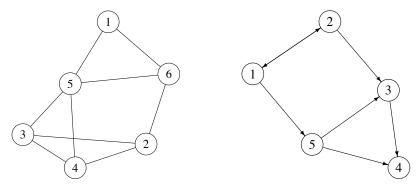
A representação gráfica destes tipos de grafos pode ser observada na Figura 2.1. Os grafos presentes ao longo dessa dissertação foram construídos utilizando o programa R (R Development Core Team , 2008) e os pacotes gRbase (Dethlefsen e Højsgaard, 2005) e igraph (Csardi e Nepusz, 2006).

Nesta dissertação, todos os grafos possuem o conjunto de vértices V finito, onde |V|=N e $\mathbb{G}(V)$ é a coleção de todos os grafos simples com o conjunto de vértices V.

Um grafo g pode ser representado por sua matriz de adjacência \mathbf{X} , que indica o conjunto de arestas existentes entre pares de vértices. No caso de grafos simples essa matriz é determinada por

$$X_{ij} = \begin{cases} 1, & \text{se os v\'ertices } i \in j \text{ s\~ao adjacentes} \\ 0, & c.c. \end{cases}$$

onde $X_{ij} = X_{ji}$ e $X_{ii} = 0$, para $1 \le i < j \le N$.



- (a) Grafo simples com 6 vértices e 9 arestas.
- (b) Dígrafo com 5 vértices e 6 arestas direcionadas.

Figura 2.1: Representação dos grafos.

Observe que para grafos sem laços os elementos da diagonal da matriz de adjacência \mathbf{X} são todos iguais a zero. Além disso, quando A é um conjunto de pares não ordenados \mathbf{X} é simétrica.

Exemplo 2.1.1. Considerando o grafo representado na Figura 2.1(a), sua matriz de adjacência é dada por:

$$\boldsymbol{X} = \left(\begin{array}{cccccc} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$

2.1.1 Definições

Dado um grafo simples, o grau de um vértice é definido como o número de vértices adjacentes a esse vértice, ou seja, o número de arestas conectadas a este vértice. Analiticamente, temos que

$$gr(i) = \sum_{i < j} x_{ij} \tag{2.1}$$

Por exemplo, considerando o grafo da Figura 2.1(a), a sua sequência de graus é (2,3,3,3,4,3). A sequência dos graus de um grafo é uma informação importante a respeito do grafo, porém apenas com essa sequência, algumas vezes, não é possível identificar a estrutura do grafo. Para exemplificar essa afirmação tome como sequência de graus de um grafo com 4 vértices a sequência (2,1,1,2). Observe pela Figura 2.2 que dois grafos possuem a sequência de graus dada. Na Seção 2.2 será introduzido o modelo de grafos que baseia-se nas sequências de graus dos vértices.

Um caminho em um grafo é uma sequência de vértices de modo que os pares de vértices consecutivos dessa sequência estão conectados por uma aresta. Assim, o tamanho de um caminho é igual ao número de vértices pertencentes ao caminho.

O menor caminho possível entre dois vértices é chamado de caminho geodésico. A partir do conceito de caminho pode-se definir a distância geodésica l_{ij} entre os vértices i e j como sendo o tamanho do menor caminho entre eles. A distância média do vértice i com relação a todos os

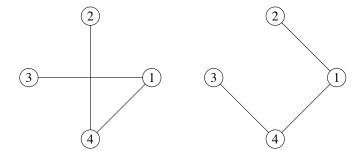


Figura 2.2: Grafos construídos a partir da sequência de graus (gr(1), gr(2), gr(3), gr(4)) = (2, 1, 1, 2).

vértices do grafo é dada pela Equação 2.2.

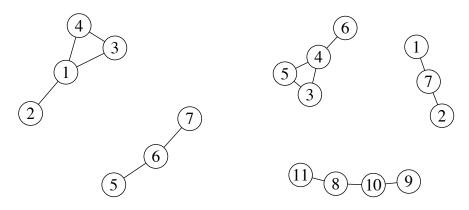
$$l_i = \frac{1}{N} \sum_{i=1}^{N} l_{ij} \tag{2.2}$$

A distância geodésica média entre todos os vértices do grafo é dada pela Equação 2.3 como a média da distância média de cada vértice do grafo.

$$l = \frac{1}{N} \sum_{i=1}^{N} l_i = \frac{1}{N^2} \sum_{1 \le i \le j \le N} l_{ij}$$
(2.3)

O diâmetro de um grafo é definido como a maior distância geodésica entre quaisquer dois pares de vértices do grafo.

O conjunto de vértices de alguns grafos pode ser dividido em subconjuntos de modo que não exista nenhum caminho entre pares de vértices de subconjuntos distintos e que exista pelo menos um caminho entre pares de vértices pertencentes ao mesmo subconjunto. Esses tipos de grafos são chamados de grafos desconexos e, seus subconjuntos são ditos de componentes do grafo. Exemplos de grafos desconexos com número distinto de componentes podem ser vistos na Figura 2.3.



- (a) Grafo desconexo com dois componentes.
- (b) Grafo desconexo com três componentes.

Figura 2.3: Exemplos de grafos desconexos.

Em algumas aplicações de grafos há interesse de estudar a propriedade de transitividade, ou seja, se i e j são adjacentes e i é adjacente a k então j e k são adjacentes. A transitividade de um grafo pode ser obtida por meio da probabilidade de vértices adjacentes a um determinado vértice serem adjacentes entre si.

Nas aplicações da teoria de grafos em redes sociais a transitividade é uma propriedade importante, pois é mais provável que amigos do mesmo indivíduo sejam amigos entre si.

Considere um grafo simples em que o vértice i é adjacente a k_i vértices do grafo. Entre os k_i vértices adjacentes a i podem existir no máximo $\frac{k_i(k_i-1)}{2}$ arestas, ou seja, quando todos os k_i vértices são adjacentes entre si. Definimos o coeficiente de agrupamento local como a razão de arestas que realmente existem entre os k_i vértices como

$$C_i = \frac{2|\{x_{jk} : x_{ji} = 1, x_{ki} = 1 \text{ e } x_{jk} = 1\}|}{k_i(1 - k_i)}$$
(2.4)

Assim, o coeficiente de agrupamento de um grafo simples é dado pela Equação (2.5) como a média dos coeficientes de agrupamento locais do grafo.

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{2.5}$$

2.2 Modelos de grafos aleatórios

Um grafo é dito aleatório se sua estrutura for originada a partir de um processo aleatório. Assim, pode-se definir um conjunto de parâmetros do grafo que será fixo enquanto os outros parâmetros são definidos de modo aleatório. Dito isso, os grafos que são estudados nesta dissertação possuem o número de vértices fixados e todos os outros parâmetros serão aleatórios, como exemplo, a existência de arestas entre dois vértices. A distribuição de probabilidade de um grafo aleatório g será denotada por $\pi(g)$.

Denotaremos por X_{ij} a variável aleatória que indica se os vértices i e j são adjacentes, dada por

$$X_{ij} = \begin{cases} 1, & \text{se os v\'ertices } i \in j \text{ s\~ao adjacentes} \\ 0, & c.c. \end{cases}$$

A matriz de adjacência do grafo aleatório será denotada por \mathbf{X} . Dado um grafo observado denota-se por \mathbf{x} a matriz de adjacência desse grafo e x_{ij} a existência de aresta entre os vértices i e j.

Detota-se por \mathbf{x}_{-ij} a matriz de adjacência de um grafo sem considerar o par de vértices (i, j). Se em um grafo $x_{ij} = 1$ ou $x_{ij} = 0$ a sua matriz de adjacência será denotado por \mathbf{x}_{ij}^+ ou \mathbf{x}_{ij}^- , respectivamente.

Os principais modelos de grafos aleatórios encontrados na literatura são apresentados nas próximas seções.

2.2.1 Erdős-Rényi

A teoria de grafos aleatórios foi introduzida por Paul Erdős e Alfréd Rényi, onde os primeiros estudos de grafos com um número fixado de vértices e de arestas foram realizados. Nos trabalhos Erdős e Rényi (1959) e Erdős e Rényi (1960) foram introduzidos dois modelos de grafos aleatórios que, posteriormente, ficaram conhecidos como modelos de Erdős-Rényi.

No primeiro modelo considera-se o número de arestas m fixado e um grafo é escolhido aleatóriamente e com mesma probabilidade dentre os $\binom{\binom{N}{2}}{m}$ grafos que possuem N vértices e m arestas. Esse modelo também é conhecido na literatura como modelo G(n,m).

O segundo modelo, que é mais usual, consiste em todos os grafos em que as arestas são escolhidas independentemente com probabilidade p. Note que neste modelo o número de arestas do grafo é uma variável aleatória e sua esperança é dada por $\binom{N}{2}p$. Dessa forma, para que o grafo tenha em média m arestas devemos tomar $p=\frac{m}{\binom{N}{2}}$. Esse modelo também é conhecido como modelo G(n,p) e por modelo de Bernoulli.

Para formalizar o segundo modelo considere que o grafo g é uma coleção $\{X_{ij}: 1 \leq i < j \leq N\}$ de variáveis aleatórias independentes e identicamente distribuídas, onde $\mathbb{P}(X_{ij}=1)=p$ para $0 . Assim, se <math>g_0$ é um grafo com N vétices e possui m arestas, temos que

$$\mathbb{P}(G=g_0)=p^m(1-p)^{N-m}.$$

No estudo de grafos aleatórios, além da presença de arestas ser uma variável aleatória temos que todas as outras variáveis associadas com o grafo também são aleatórias, como por exemplo, o grau dos vértices.

Para estudar a distribuição do grau dos vértices de um grafo aleatório considere p_k como a probabilidade de um vértice escolhido ao acaso no grafo g ter grau k. Dado um grafo com N vértices seguindo o modelo de Erdős-Rényi, é fácil ver que

$$p_k = {N-1 \choose k} p^k (1-p)^{N-1-k}.$$

Quando aumenta-se o valor de N a distribuição do grau do grafo converge para a distribuição de Poisson com parâmetro Np. Porém, não são em todos os modelos de grafos aleatórios que esta propriedade é verificada. Esse tópico será discutido na Seção 2.2.4.

2.2.2 Condicionalmente Uniforme

Muitas vezes deseja-se estudar grafos aleatórios que possuem determinada característica, como por exemplo, grafos com o número de arestas fixado, grafos com caminhos pequenos ou grafos desconexos com um número pré-estabelecido de componentes. Assim, para extrair as informações desejadas dos grafos podem-se utilizar diferentes propriedades: grau dos vértices, número de arestas, de componentes e de triângulos do grafo.

Dizemos que a distribuição de um grafo é condicionamente uniforme se os grafos que satisfazem um conjunto de propriedades fixado têm a mesma probabilidade de serem escolhidos e os demais possuem probabilidade 0. Como a distribuição dos grafos é uniforme, condicionada a um conjunto predefinido, para este modelo temos que

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{S}(\mathbf{x}) = \mathbf{s}) = \frac{\mathbb{1}_{\{\mathbf{S}(\mathbf{x}) = \mathbf{s}\}}}{|\mathbf{x}': \mathbf{S}(\mathbf{x}') = \mathbf{s}|}$$

onde, X é a matriz de adjacência do grafo e S(x) é o vetor de propriedades do grafo.

Exemplo 2.2.1. Deseja-se calcular a probabilidade de que um grafo com 4 vértices tenha 2 triângulos, ou seja, $S_1(\mathbf{x}) = \sum_{i < j < k} x_{ij} x_{jk} x_{ki}$. Como o número de grafos com 4 vértices e 2 triângulos é igual a 16, temos

$$\mathbb{P}(X = x | S_1(x) = 2) = \frac{\mathbb{1}_{\{S_1(x) = 2\}}}{16}$$

2.2.3 Distribuição livre de escala

O trabalho de Albert et al. (1999) a respeito de redes de internet produziu resultados importantes para a aplicação da teoria de grafos aleatórios em redes reais. Nesse trabalho, redes de internet foram estudadas através de sua representação como um grafo orientado, de modo que os documentos presentes na Web são os vértices do grafo e os links presentes entre esses documentos são as arestas direcionadas. A partir do grafo, definiu-se a probabilidade de um vértice ter k arestas de saída p_k^- e ter k arestas de entrada p_k^+ .

Diferentemente do resultado anterior no qual para grafos do modelo de Erdős-Rényi a distribuição de probabilidade dos graus dos vértices converge para a distribuição de Poisson na medida em que o número de vértices do grafo aumenta, nesse trabalho de redes reais de internet obtiveram que a distribuição dos graus do grafo segue a lei de potência.

A distribuição do grau dos vértices segue a lei de potência se para algum α valer que

$$p_k = Ck^{-\alpha} \tag{2.6}$$

A distribuição dos graus de um grafo que segue a lei de potência apresenta assimetria à direita e cauda mais longa e pesada do que a distribuição de Poisson, indicando que a probabilidade de obter vértices com muitas arestas para a distribuição que segue a lei de potência é maior do que quando considera-se a distribuição de Poisson.

Os grafos em que a distribuição do grau segue a lei de potência são chamados de grafos livres de escala, uma vez que a lei de potência é invariante por escala, ou seja,

$$p_{ak} = C(ak)^{-\alpha} = Ca^{-\alpha}k^{-\alpha} \propto Ck^{-\alpha} = p_k$$
(2.7)

Outra propriedade importante da lei de potência é que em escala logarítmica resulta em uma função linear, Equação 2.8, tornando-se assim um meio de identificação da existência da lei da potência na distribuição dos graus do grafo.

$$ln(p_k) = -\alpha ln(k) + ln(C) \tag{2.8}$$

2.2.4 Distribuição do grau fixada

A partir dos estudos que comprovaram que em dados de redes reais a distribuição dos graus de um grafo segue a lei de potência, no trabalho de Newman *et al.* (2001) são propostos grafos com distribuição do grau dos vértices fixada.

Para descrever este modelo, considere um grafo com um número finito de vértices N em que a distribuição do grau dos vértices seja p_k , ou seja, a probabilidade de um vértice ter grau k é p_k . Seja g_1, \ldots, g_N variáveis independentes de modo que $P(g_i = k) = p_k$, assim essas variáveis representam a sequência de graus do grafo aleatório. Pode-se pensar que i é um vértice do grafo com g_i arestas emergindo deste vértice. Para obter todos os grafos que possuem essa sequência de graus basta adicionar as arestas ao grafo respeitando o grau de cada vértice. Note que se $\sum_{i=1}^{N} g_i$ for ímpar a probabilidade de selecionarmos um grafo com a sequência de graus dada é zero.

Exemplo 2.2.2. Considere um grafo com 5 vértices nomeados e fixados com a seguinte sequência de graus (3,1,2,2,2). Na Figura 2.4, obtemos 6 possíveis grafos em que o vértice 1 tem grau 3, o vértice

2 tem grau 1 e os outros vértices tem grau 2. Assim, a probabilidade de selecionarmos um grafo com a sequência de graus dada é por $\mathbb{P}\{(gr(1),gr(2),gr(3),gr(4),gr(5))=(3,1,2,2,2)\}=\frac{6}{2^{10}}$.

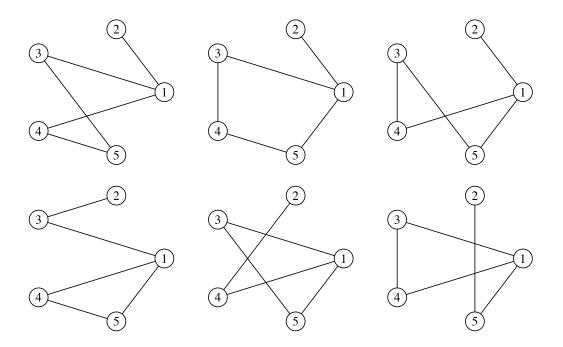


Figura 2.4: Grafos construídos a partir da sequência de graus (gr(1), gr(2), gr(3), gr(4), gr(5)) = (3, 1, 2, 2, 2)

2.2.5 Exponencial

O modelo de grafo aleatório exponencial é abordado na literatura principalmente pela sua vasta aplicação em dados de redes socias. Isso deve-se ao fato de que nesse modelo assume-se que o número de vértices do grafo é fixo e a existência de arestas entre pares de vértices depende de toda a estrutura da rede. Dito isso, imagine uma rede de amizade entre alunos de uma turma de estatística, nesse caso, cada vértice do grafo representa um indivíduos da turma e as arestas representam as relações de amizade existentes entre os alunos. A relação de amizade entre João e Maria pode influenciar na relação entre João e Alice, principalmente se o primeiro relacionamento for amoroso. Ou ainda, as relações de amizade entre Carlos e Pedro e Pedro e Manuel podem influenciar na exitência de amizade entre Carlos e Manuel. Para mais detalhes sobre a aplicação de grafos aleatórios exponenciais o leitor pode consultar Robins et al. (2007).

Cada parâmetro do modelo Exponencial corresponde a uma configuração do grafo e essas configurações representam as características estruturais do grafo. Dessa forma, o modelo de grafo aleatório exponencial é dado por

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \frac{exp(\boldsymbol{\theta}^T \mathbf{S}(\mathbf{x}))}{z(\boldsymbol{\theta})}$$
(2.9)

em que, θ é o vetor de parâmetros do modelo, $\mathbf{S}(\mathbf{x})$ é o vetor de estatísticas do grafo \mathbf{x} e $z(\theta)$ é a constante de normalização.

Note que o termo exponencial presente na Equação 2.9 é responsável pelo nome do modelo.

A constante de normalização do modelo será denotada por $z(\theta)$ para ressaltar a sua dependência

do vetor de parâmetros do modelo. Dada pela Equação 2.10, a constante de normalização é a soma sobre todos os possíveis grafos que possuem N vértices. Para grafos simples, tem-se que χ é o conjunto de todas as matrizes simétricas com N linhas e colunas, totalizando assim $2^{\binom{N}{2}}$ possíveis grafos.

$$z(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in \chi} exp(\boldsymbol{\theta}^T \mathbf{S}(\mathbf{x}))$$
 (2.10)

Na tabela 2.1 encontram-se alguns exemplos de estatísticas que podem ser usadas para definir o modelo de grafo exponencial.

Tabela 2.1: Exemplos de estatísticas utilizadas para descrever grafos simples.

Parâmetro	Representação	Estatística
arestas	①———①	$\sum_{i < j} x_{ij}$
2-estrelas	© ®	$\sum_{i < j < k} x_{ij} x_{ik}$
3-estrelas	0 8	$\sum_{i < j < k < l} x_{ij} x_{ik} x_{il}$
triângulo	0	$\sum_{i < j < k} x_{ij} x_{jk} x_{ki}$

No modelo de Erdős-Rényi, descrito na Seção 2.2.1, a existência de arestas entre os pares de vértices do grafo são variáveis aleatórias com distribuição de Bernoulli com parâmetro fixo p. Esse modelo pode ser representado pelo modelo Exponencial, descrito na Equação 2.9, tomando $p=\frac{exp(\theta)}{1+exp(\theta)}$ e é dado por

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \frac{exp\left(\theta \sum_{i < j} x_{ij}\right)}{z(\theta)}$$
(2.11)

De fato,

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \prod_{i < j} \mathbb{P}(X_{ij} = x_{ij}) = \prod_{i < j} \left[x_{ij} \frac{exp(\theta)}{1 + exp(\theta)} (1 - x_{ij}) \frac{1}{1 + exp(\theta)} \right]$$

$$= \frac{exp\left(\theta \sum_{i < j} x_{ij}\right)}{(1 + exp(\theta))^{\sum_{i < j} x_{ij}}} \cdot \frac{1}{(1 + exp(\theta))^{\sum_{i < j} (1 - x_{ij})}}$$

$$= \frac{exp\left(\theta \sum_{i < j} x_{ij}\right)}{(1 + exp(\theta))^{\binom{N}{2}}} \tag{2.12}$$

Os trabalhos encontrados na literatura associados ao modelo de grafo aleatório Exponencial frequentemente estão relacionados com a inferência estatística dos parâmetros do modelo. Os métodos tradicionais de inferência, como por exemplo, o método de máxima verossimilhança não é apropriado para redes com muitos vértices, pela dificuldade de calcular a constante de normalização dada por 2.10.

Os trabalhos recentes apresentam métodos de estimação a partir de métodos de Monte Carlo baseados em cadeias de Markov (MCMC) que simulam a distribuição do grafo aleatório a partir de um conjunto inicial de parâmetros e a cada iteração o conjunto de parâmetros é atualizado com base na comparação da distribuição do grafo aleatório com o grafo observado. Para detalhes dessa construção o leitor pode consultar Snijders (2002) e para uma abordagem bayesiana consultar Caimo e Friel (2011). O método de simulação de grafos Exponenciais será discutido no Seção 4.1.2.

2.2.6 Mundo Pequeno

Um dos fenômenos vastamente discutidos e estudados na literatura é o efeito de Mundo Pequeno em redes reais¹, em termos da teoria de grafos ocorre quando as distâncias entre os vértices da rede é pequena.

Na década de 1960, o psicólogo americano Stanley Milgram se interessou por estudos de redes sociais e com o objetivo de quantificar o tamanho do caminho entre dois indivíduos na rede propôs um experimento, conjuntamente com outros pesquisadores, que nos dias de hoje é conhecido como experimento de Mundo Pequeno.

No experimento de Mundo Pequeno um total de 96 pacotes foram enviados para pessoas selecionadas aleatóriamente da lista telefônica da cidade de Ohama, no estado de Nebraska. Cada pacote continha uma cópia do passaporte de um amigo de Stanley que morava em Boston, no estado de Massachusetts. Além disso, os pacotes contiam as instruções de que o passoporte deveria ser entregue para o dono, porém não diretamente. A pessoa que recebeu o pacote deveria reinviá-lo para algum amigo, que faria o mesmo procedimento até que o passaporte fosse entregue para seu verdadeiro dono.

Dos 96 pacotes que foram distribuidos, 18 chegaram no destino final. Com as informações desses 18 caminhos completos, o valor que Milgram obteve para o tamanho médio do caminho de um

¹conhecido em inglês por Small-World networks.

indivíduo do estado de Nebraska para seu amigo em Massachusetts foi 5.9. Esse resultado deu origem a ideia de "seis graus de separação", em que quaisquer duas pessoas do mundo estão ligadas por no máximo seis amigos em comum.

Porém, existem algumas críticas relacionadas à aproximação desse resultado. Utilizou-se nesse experimento apenas um "alvo" localizado em Boston, além disso as pessoas que receberam os pacotes, ou seja a origem dos caminhos, pertenciam à mesma cidade. O valor obtido é na verdade um limitante superior da distância obtida, pois não existem garantias de que o caminho tomado é o menor caminho entre todos os possíveis.

Por fim, alguns críticos argumentam que muitos pacotes não foram entregues, o que, por hipótese, pode ter ocorrido por terem se perdido devido a caminhos longos que foram tomados. Assim, os caminhos completos são uma amostra viesada do experimento e podem representar os caminhos que possuem tamanhos menores que a média de todos os caminhos. Para mais detalhes sobre este experimento e outros que foram feitos posteriormente o leitor pode consultar Newman (2010).

Ao leitor que se interessar pela simulação de grafos aleatórios com propriedades de mundo pequeno, ou seja, com coeficiente de agrupamente alto e distância geodésica pequena recomendá-se a leitura de Watts e Strogatz (1998).

Capítulo 3

Teste de hipóteses para grafos aleatórios

Neste capítulo são apresentados os testes de hipóteses não paramétricos para amostras de grafos proposto e utilizado ao longo desta dissertação. Alguns resultados assintóticos da estatística do teste também são apresentados e provados nas próximas seções.

3.1 Teste de hipóteses

3.1.1 Introdução

Uma revisão introdutória sobre teste de hipóteses é abordada nesta seção, como forma de familiarizar o leitor sobre o conteúdo discutido ao longo dos próximos capítulos.

Qualquer afirmação feita a respeito da função de probabilidade de uma ou mais variáveis aleatórias ou a respeiro da população da qual a amostra em estudo foi extraída é chamada de hipótese estatística. A hipótese estatística que será testada é chamada de hipótese nula e será denotada ao longo deste texto por H_0 . A hipótese alternativa, denotada por H_1 , é a conclusão do teste quando a hipótese nula é rejeitada.

Dada uma amostra $\mathbf{X} = (X_1, \dots, X_n)$, o teste de hipóteses para decidir se H_0 será rejeitada é construído com base em uma função $T = T(\mathbf{X})$ dessa amostra. Essa função $T(\mathbf{X})$ é chamada de estatística do teste.

A região crítica, ou região de rejeição, do teste de hipóteses, denotada por R, é o subconjunto de valores, definidos a partir da estatística do teste, que levam à rejeição de H_0 . O valor limite de R é chamado de valor crítico do teste.

Ao realizar o teste de hipóteses e decidir se rejeita-se ou não rejeita-se a hipótese nula é possível que essa decisão não esteja correta, existindo assim dois erros possíves de ocorrer em um teste de hipóteses: erro tipo I e o erro tipo II. O erro tipo I ocorre quando decide-se por rejeitar H_0 quando H_0 é realmente verdadeira e o erro de não reijeitar H_0 sendo que H_0 é falsa é chamado de erro tipo II. Um esquema desses erros pode ser visto na Tabela 3.1.

Tabela 3.1: Possíveis erros de decisão que podem ocorrer em um teste de hipóteses.

	H ₀ verdadeira	H ₀ falsa
Não rejeitar H ₀	Decisão correta	Erro tipo II
Rejeitar H ₀	Erro tipo I	Decisão correta

16

O nível de significância do teste é a probabilidade de rejeitar H_0 dado que H_0 é verdadeira. Esse nível pode ser fixado quando o teste de hipóteses for realizado como medida de controlar os erros que podem ser cometidos na tomada de decisão do teste. O nível de significância será denotado pela letra grega α .

A função poder do teste está relacionada com a probabilidade de rejeição da hipótese nula, em relação as hipóteses testadas. O valor calculado pelo poder do teste, dado que a hipótese alternativa é verdadeira, representa a probabilidade de que rejeitar H_0 quando H_1 é verdadeira, ou seja, quando mais próximo de 1 for este valor maior é a probabilidade de que a decisão tomada não esteja errada. A equação (3.1) representa a função poder de um teste de hipóteses em que θ é o parâmetro a ser testado.

$$\varphi(\theta) = \mathbb{P}(T(\mathbf{X}) \in \mathbf{R}|\mathbf{H}_1) \tag{3.1}$$

Um teste de hipóteses é dito consistente se sua função poder, determinada sob a hipótese alternativa, tender a 1 na medida em que o tamanho da amostra estudada crescer. Quando o poder de um teste de hipóteses for maior do que o poder de qualquer outro teste realizado, sob as mesmas hipóteses e com o mesmo nível de significância, o teste é dito teste mais poderoso. O poder do teste de hipóteses é afetado de algumas formas, como:

- Quando aumentamos o nível de significância do teste a região de rejeição do teste torna-se maior, com isso a probabilidade de rejeitar H₀ aumenta, aumentando assim o poder do teste.
- Quando o tamanho da amostra cresce o poder do teste também cresce.
- Quanto maior for a diferença entre o valor "real" do parâmetro e o valor testado maior será o poder do teste.

A estrutura e os principais conceitos de um teste de hipóteses foram descritos até esse ponto, porém a tomada de decisão de reijeitar ou não H_0 é feita para nível de significância α fixado. O pesquisador interessado na tomada de decisão pode se questionar a respeito do valor fixado para o nível de significância do teste. Dessa forma, o pesquisador pode questionar-se: dada a amostra em estudo, qual o valor mínimo de α que pode ser escolhido de modo que a hipótese nula seja rejeitada?

Para responder essa pergunta utiliza-se o nível descritivo do teste, também conhecido como p-valor. O p-valor, denotado por $\hat{\alpha}$, é definido como a probabilidade, dado que H_0 é verdadeira, de obter valores da estatística do teste extremos com relação ao valor obtido pela amostra. Assim, a hipótese nula deve ser rejeitada se $\alpha > \hat{\alpha}$, para algum nível de significância α .

3.1.2 Testes paramétricos e não paramétricos

Os testes de hipóteses podem ser classificados em paramétricos e não paramétricos. As principais diferenças entre estes dois tipos de testes são discutidas nesta seção.

Os testes paramétricos baseiam-se em parâmetros populacionais, como por exemplo, quando há interesse em testar valores para a média ou desvio padrão da população. Para este tipo de teste é normal que o pesquisador faça algumas pressuposições a respeito da população em estudo, assumindo assim que a distribuição dos dados em estudo segue algum modelo específico, em geral, o modelo adotado é o modelo Normal.

Por outro lado, para testes não paramétricos não é necessário assumir qualquer hipótese a respeito da distribuição dos dados. Os testes não paramétricos, em geral, tem como objetivo testar a aderência de modelos aos dados coletados, a indepêndencia entre duas amostras e aleatoriedade dos dados. Com isso, esse tipo de teste pode ser utilizado em dados qualitativos, ao contrário dos testes paramétricos. Porém, a desvantagem dos testes não paramétricos é que quando as hipóteses sobre a distribuição dos dados é verdadeira eles não são tão poderosos quanto a outra classe de testes.

Para mais detalhes sobre os principais testes paramétricos e não paramétricos o leitor interessado pode consultar (Gibbons e Chakraborti, 2010).

3.2 Teste de hipóteses não paramétrico para grafos

3.2.1 Definições

O teste de hipóteses para grafos proposto nessa dissertação é uma adaptação do teste de hipóteses proposto em (Busch *et al.*, 2009) para árvores aleatórias. Dessa forma, as definições utilizadas na construção do teste de hipóteses para árvores foram adaptadas para o caso de grafos e são apresentadas nesta Seção.

Primeiramente, é necessário definir uma medida de distância entre dois grafos. Como os grafos estudados são grafos simples é intuitivo pensar em uma medida que considere as diferenças de arestas presentes entre os grafos.

Definição 3.2.1. Dados dois grafos g_1 e g_2 , a distância entre g_1 e g_2 é definida por

$$d(g_1, g_2) = \sum_{i < j} (x_{ij}^1 - x_{ij}^2)^2$$
(3.2)

Como o teste de hipóteses proposto é para amostras de grafos é necessário definir um valor que mensure a discrepância entre um grafo para uma amostra de grafos.

Definição 3.2.2. Dado um grafo g e uma amostra de grafos $\mathbf{g} = (g_1, \dots, g_n)$, definimos a discrepância entre g e \mathbf{g} por

$$\delta_{\mathbf{g}}(g) = \frac{1}{n} \sum_{k=1}^{n} d(g, g_k)$$
 (3.3)

A partir de uma amostra de grafos é possível determinar o número médio de arestas presentes entre pares de vértices, como descrito pela Definição 3.2.3.

Definição 3.2.3. Dados um par de vértices (i, j) e uma amostra de grafos $\mathbf{g} = (g_1, \dots, g_n)$, definimos o número médio de conexões entre os vértices (i, j) por

$$\overline{g}(i,j) = \frac{1}{n} \sum_{k=1}^{n} x_{ij}^{k}$$
(3.4)

3.2.2 Teste de hipóteses para uma amostra única de grafos

Nesta Seção é construído um teste de hipóteses não paramétrico para uma única amostra de grafos, com o objetivo de testar se a distribuição de grafos da qual a amostra observada foi extraída é a mesma distribuição estabelecida pela hipótese estatística. Portanto, deseja-se testar

$$\begin{cases} H_0: \pi = \pi_0 \\ H_1: \pi \neq \pi_0 \end{cases}$$

A estatítica do teste de hipóteses baseada em Busch et al. (2009) é dada pela Definição 3.2.4.

Definição 3.2.4. Dada uma amostra aleatória de grafos $g = (g_1, \dots, g_n)$ definimos a estatística

$$W(\mathbf{g}) = \max_{g \in \mathbb{G}(V)} |\delta_{\mathbf{g}}(g) - \pi_0 D(g)|$$
(3.5)

onde

$$\pi_0 D(g) = \sum_{g' \in \mathbb{G}(V)} d(g, g') \pi_0(g')$$

A estatística W é o máximo de $g \in \mathbb{G}(V)$ da diferença absoluta entre a discrepância do grafo g e a amostra \mathbf{g} e da distância média de todos os grafos pertencentes a $\mathbb{G}(V)$ ao grafo g em relação a distribuição π_0 . Assim, espera-se que para pequenos valores dessa estatística existam evidências de que as amostras de grafos possuem a mesma distribuição. Com isso, a região crítica do teste baseado na estatística W é definida para um nível de significância α , onde $q_{(1-\alpha)}$ é o quantil de ordem $1-\alpha$ da distribuição de W, como

$$R = \{w : w > q_{(1-\alpha)}\}$$
(3.6)

Note que o cálculo da estatística W não é trivial computacionalmente, pois para amostras de grafos com V vértices tem-se que $|\mathbb{G}(V)| = 2^{\binom{V}{2}}$. Mostraremos nesta dissertação como computar o valor de W de forma eficiente. A partir da Proposição 3.2.1, tem-se que não é necessário explorar todo o espaço de grafos $\mathbb{G}(V)$ para o calculo da estatística W, basta calcular para cada par de vértices (i,j) a diferença absoluta entre o número médio de conexões da amostra e a esperança da variável X_{ij} sob H_0 .

Proposição 3.2.1. Dada uma amostra de grafos $g = (g_1, \ldots, g_n)$ vale que

$$W(\mathbf{g}) = \sum_{i < j} |\overline{\mathbf{g}}(i, j) - \mathbb{E}_{\pi_0}[X_{ij}]|$$

Demonstração. Para maximizar $|D_{\mathbf{g}}(g)| = |\delta_{\mathbf{g}}(g) - \pi_0 D(g)|$ em $\mathbb{G}(V)$ basta maximizar $D_{\mathbf{g}}(g)$ e $-D_{\mathbf{g}}(g)$.

$$\begin{split} D_{\mathbf{g}}(g) &= \delta_{\mathbf{g}}(g) - \pi_{0}D(g) = \frac{1}{n} \sum_{k=1}^{n} d(g, g_{k}) - \sum_{g' \in \mathbb{G}(V)} d(g, g')\pi_{0}(g') \\ &= \frac{1}{n} \sum_{k=1}^{n} \sum_{i < j} (x_{ij}^{k} - x_{ij})^{2} - \sum_{g' \in \mathbb{G}(V)} \pi_{0}(g') \sum_{i < j} (x_{ij}' - x_{ij})^{2} \\ &= \sum_{i < j} \left\{ \left[\sum_{k=1}^{n} \frac{x_{ij}}{n} - 2x_{ij} \sum_{k=1}^{n} \frac{x_{ij}^{k}}{n} + \sum_{k=1}^{n} \frac{x_{ij}^{k}}{n} \right] - \sum_{g' \in \mathbb{G}(V)} \pi_{0}(g') \left(x_{ij} - 2x_{ij} x_{ij}' + x_{ij}' \right) \right\} \\ &= \sum_{i < j} \left\{ x_{ij} - 2x_{ij} \overline{\mathbf{g}}(i, j) + \overline{\mathbf{g}}(i, j) - x_{ij} \sum_{g' \in \mathbb{G}(V)} \pi_{0}(g') - 2x_{ij} \sum_{g' \in \mathbb{G}(V)} x_{ij}' + \sum_{g' \in \mathbb{G}(V)} x_{ij}' \right\} \\ &= \sum_{i < j} 2x_{ij} \left[\sum_{g' \in \mathbb{G}(V)} x_{ij}' \pi_{0}(g') - \overline{\mathbf{g}}(i, j) \right] + x_{ij} \left(1 - \sum_{g' \in \mathbb{G}(V)} \pi_{0}(g') \right) + \overline{\mathbf{g}}(i, j) + \sum_{g' \in \mathbb{G}(V)} x_{ij}' \pi_{0}(g') \right] \end{split}$$

Como $\sum_{g' \in \mathbb{G}(V)} \pi_0(g') = 1$ e $\sum_{g' \in \mathbb{G}(V)} x'_{ij} \pi_0(g') = \mathbb{E}_{\pi_0}[X_{ij}]$ segue que

$$D_{\mathbf{g}}(g) = \sum_{i < j} 2x_{ij} (\mathbb{E}_{\pi_0}[X_{ij}] - \overline{\mathbf{g}}(i,j)) + \overline{\mathbf{g}}(i,j) + \mathbb{E}_{\pi_0}[X_{ij}]$$

$$= \sum_{i < j} (\overline{\mathbf{g}}(i,j) - \mathbb{E}_{\pi_0}[X_{ij}]) (1 - 2x_{ij})$$
(3.7)

A matriz de adjacência do grafo g^* que maximiza $D_{\mathbf{g}}(g)$ é dada por

$$x_{ij}^* = \begin{cases} 1, & \overline{\mathbf{g}}(i,j) \leqslant \mathbb{E}_{\pi_0}[X_{ij}] \\ 0, & c.c. \end{cases}, \quad 1 \le i < j \le |V|$$

Analogamente, a matriz de adjacência do grafo g^{**} que maximiza $-D_{\mathbf{g}}(g)$ é

$$x_{ij}^{**} = \begin{cases} 1, & \overline{\mathbf{g}}(i,j) \ge \mathbb{E}_{\pi_0}[X_{ij}] \\ 0, & c.c. \end{cases}$$
, $1 \le i < j \le |V|$

Agora, basta provar que $|D_{\mathbf{g}}(g^*)| = |D_{\mathbf{g}}(g^{**})|$. De fato,

$$D_{\mathbf{g}}(g^*) = \sum_{\substack{i < j \\ \overline{\mathbf{g}}(i,j) \leqslant \mathbb{E}_{\pi_0}[X_{ij}]}} (\mathbb{E}_{\pi_0}[X_{ij}] - \overline{\mathbf{g}}(i,j)) - \sum_{\substack{i < j \\ \overline{\mathbf{g}}(i,j) \leq \mathbb{E}_{\pi_0}[X_{ij}]}} (\mathbb{E}_{\pi_0}[X_{ij}] - \overline{\mathbf{g}}(i,j)) = -D_{\mathbf{g}}(g^{**})$$

Por fim, pela equação (3.7) segue que

$$W(\mathbf{g}) = D_{\mathbf{g}}(g^*) = \sum_{i < j} (\overline{\mathbf{g}}(i, j) - \mathbb{E}_{\pi_0}[X_{ij}])(1 - 2x_{ij}^*) = \sum_{i < j} (\overline{\mathbf{g}}(i, j) - \mathbb{E}_{\pi_0}[X_{ij}])I_{ij}$$

onde

Portanto,

20

$$W(\mathbf{g}) = \sum_{i < j} |\overline{\mathbf{g}}(i, j) - \mathbb{E}_{\pi_0}[X_{ij}]|$$

Para uma amostra aleatória de grafos mostraremos que a distribuição da estatística do teste W, sob H_0 , converge para a soma do módulo de variáveis aleatórias Normais quando o tamanho da amostra cresce, como descrito na Proposição 3.2.2.

Proposição 3.2.2. Considere uma amostra independente de grafos $\mathbf{g} = (g_1, \dots, g_n)$. Tome $p_{ij} = \mathbb{P}(X_{ij} = 1)$ e $Y_{ij} = \overline{\mathbf{g}}(i,j) - \mathbb{E}_{\pi_0}[X_{ij}]$. Sob a hipótese nula H_0 vale que

$$W(\mathbf{g}) = \sum_{i < j} |Y_{ij}| \qquad onde \qquad \frac{Y_{ij}}{\sqrt{\frac{p_{ij}(1 - p_{ij})}{n}}} \xrightarrow{D} N(0, 1)$$
(3.8)

Demonstração. Para cada par de vértices i e j fixados temos que $X_{ij}^1, X_{ij}^2, \ldots, X_{ij}^n$ são independentes e identicamente distribuídas, onde $X_{ij}^k \sim Ber(p_{ij})$.

Pela Proposição 3.2.1, temos que

$$W(\mathbf{g}) = \sum_{i < j} |\overline{\mathbf{g}}(i, j) - \mathbb{E}_{\pi_0}[X_{ij}]| = \sum_{i < j} \left| \sum_{k=1}^n \frac{X_{ij}^k}{n} - \mathbb{E}_{\pi_0}[X_{ij}] \right|$$
(3.9)

Como $\mathbb{E}_{\pi_0}[X_{ij}] = p_{ij}$ segue pelo Teorema Central do Limite que

$$\frac{\sum_{k=1}^{n} \frac{X_{ij}^{k}}{n} - p_{ij}}{\sqrt{\frac{p_{ij}(1 - p_{ij})}{n}}} \xrightarrow[n \to \infty]{D} N(0, 1)$$
(3.10)

Fazendo $Y_{ij} = \overline{\mathbf{g}}(i,j) - \mathbb{E}_{\pi_0}[X_{ij}]$, segue o resultado.

Exemplo 3.2.1. Considere uma amostra independente de grafos $\mathbf{g} = (g_1, \dots, g_n)$ do modelo Exponencial dado pela equação 2.9. Deseja-se verificar se essa amostra de grafos segue o modelo Exponencial para um dado parâmetro $\boldsymbol{\theta}_0$. Desse modo, as hipóteses do teste são

$$\begin{cases} H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ H_1: \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \end{cases}$$

Para calcular $p_{ij} = \mathbb{P}(X_{ij} = 1)$, sob H_0 , fazemos

$$\mathbb{P}(X_{ij} = 1) = \sum_{x_{-ij}} \mathbb{P}(X_{ij} = 1, X_{-ij} = x_{-ij}) = \sum_{x_{-ij}} \mathbb{P}(X_{ij}^+ = x_{ij}^+)
= \frac{1}{z(\boldsymbol{\theta}_0)} \sum_{x_{-ij}} exp\left(\boldsymbol{\theta}_0^T \sum_{i < j} S(x_{ij}^+)\right)$$
(3.11)

Observe que os valores de p_{ij} dependem da constante de padronização do modelo $z(\boldsymbol{\theta}_0)$. Como essa constante, dada pela Equação 2.10, é uma soma sobre todos os grafos com N vértices, torna-se computacionalmente inviável o seu cálculo para grafos com muitas arestas.

Este teste será abordado novamente na Seção 4.2, onde seu desempenho será estudado a partir de dados simulados.

3.2.3 Teste de hipóteses para duas amostras de grafos

Dadas duas amostras de grafos, deseja-se testar se essas amostras provém de uma mesma distribuição. Supondo que as amostras têm distribuições π e π' , respectivamente, temos interesse em testar

$$\begin{cases} H_0: \pi = \pi' \\ H_1: \pi \neq \pi' \end{cases}$$

Para o teste de hipóteses para duas amostras de grafos a estatística W é definida como o máximo em $g \in \mathbb{G}(V)$ da diferença absoluta da distância média do grafo g a cada uma das amostras de grafos \mathbf{g}_1 e \mathbf{g}_2 , como descrito em 3.2.5.

Definição 3.2.5. Dadas duas amostras de grafos $\mathbf{g}_1 = (g_{11}, \dots, g_{1n})$ e $\mathbf{g}_2 = (g_{21}, \dots, g_{2m})$ de tamanhos n e m, respectivamente, definimos a estatística

$$W(\mathbf{g}_1, \mathbf{g}_2) = \max_{g \in \mathbb{G}(v)} |\delta_{\mathbf{g}_1}(g) - \delta_{\mathbf{g}_2}(g)|$$
(3.12)

A região crítica do teste de hipóteses para duas amostras é igual para o teste de uma amostra, como descrito por (3.6).

Também foi provado, Proposição 3.2.3, que não é necessário explorar todo o espaço de grafos $\mathbb{G}(V)$ para calcularmos a estatística W do teste para duas amostras de grafos, basta calcular para cada par de vértices a diferença absoluta entre o número médio de conexões para cada amostra.

Proposição 3.2.3. Dadas duas amostras de grafos $\mathbf{g}_1 = (g_{11}, \dots, g_{1n})$ e $\mathbf{g}_2 = (g_{21}, \dots, g_{2m})$ vale que

$$W(\boldsymbol{g}_1,\boldsymbol{g}_2) = \sum_{i < j} |\overline{\boldsymbol{g}}_1(i,j) - \overline{\boldsymbol{g}}_2(i,j)|$$

 $\mathbf{Demonstração.} \text{ Para maximizar } |\delta_{\mathbf{g}_1}(g) - \delta_{\mathbf{g}_2}(g)| = |D_{g_1,g_2}(g)| \text{ em } \mathbb{G}(V) \text{ basta maximizar } D_{g_1,g_2}(g)$

 $e - D_{g_1,g_2}(g)$.

$$D_{g_{1},g_{2}}(g) = \delta_{\mathbf{g}_{1}}(g) - \delta_{\mathbf{g}_{2}}(g) = \frac{1}{n} \sum_{k=1}^{n} d(g,g_{1k}) - \frac{1}{m} \sum_{l=1}^{m} d(g,g_{2l})$$

$$= \frac{1}{n} \sum_{k=1}^{n} \sum_{i < j} (x_{ij}^{1k} - x_{ij})^{2} - \frac{1}{m} \sum_{l=1}^{m} \sum_{i < j} (x_{ij}^{2l} - x_{ij})^{2}$$

$$= \sum_{i < j} \left[\frac{1}{n} \sum_{k=1}^{n} \left[x_{ij}^{1k} - 2x_{ij}^{1k} x_{ij} + x_{ij} \right] - \frac{1}{m} \sum_{l=1}^{m} \left[x_{ij}^{2l} - 2x_{ij}^{2l} x_{ij} + x_{ij} \right] \right]$$

$$= \sum_{i < j} \left[-2x_{ij} (\overline{\mathbf{g}}_{1}(i,j) - \overline{\mathbf{g}}_{2}(i,j)) + \overline{\mathbf{g}}_{1}(i,j) - \overline{\mathbf{g}}_{2}(i,j) \right]$$

$$= \sum_{i < j} (1 - 2x_{ij}) (\overline{\mathbf{g}}_{1}(i,j) - \overline{\mathbf{g}}_{2}(i,j)) = \sum_{i < j} (1 - 2x_{ij}) D_{\overline{\mathbf{g}}_{1},\overline{\mathbf{g}}_{2}}(i,j)$$

$$(3.13)$$

Como $g \in \mathbb{G}(V)$, segue que $(1 - 2x_{ij}) \in \{-1, 1\}$, $1 \le i < j \le |V|$. Segue da equação (3.13) que o grafo g' que maximiza $D_{g_1,g_2}(g)$ é dado por

$$x'_{ij} = \begin{cases} 1, & D_{\overline{\mathbf{g}}_1, \overline{\mathbf{g}}_2}(i, j) \leqslant 0 \\ 0, & c.c. \end{cases}, \quad 1 \le i < j \le |V|$$

Por outro lado, o grafo g'' que maximiza $-D_{g_1,g_2}(g)$ tem como matriz de adjacência

$$x_{ij}'' = \begin{cases} 1, & D_{\overline{\mathbf{g}}_1, \overline{\mathbf{g}}_2}(i, j) > 0 \\ 0, & c.c. \end{cases}, \quad 1 \le i < j \le |V|$$

Agora, basta mostrar que $|D_{g_1,g_2}(g')| = |D_{g_1,g_2}(g'')|$. De fato,

$$D_{g_1,g_2}(g') = -\sum_{\substack{i < j \\ D_{\overline{\mathbf{g}}_1,\overline{\mathbf{g}}_2}(i,j) \leqslant 0}} D_{\overline{\mathbf{g}}_1,\overline{\mathbf{g}}_2}(i,j) + \sum_{\substack{i < j \\ D_{\overline{\mathbf{g}}_1,\overline{\mathbf{g}}_2}(i,j) > 0}} D_{\overline{\mathbf{g}}_1,\overline{\mathbf{g}}_2}(i,j) = -D_{g_1,g_2}(g'')$$

Pela equação (3.13) temos que

$$W(\mathbf{g}_{1}, \mathbf{g}_{2}) = \sum_{i < j} (1 - 2X_{ij}^{*})(\overline{\mathbf{g}}_{1}(i, j) - \overline{\mathbf{g}}_{2}(i, j)) = \sum_{i < j} I_{ij}(\overline{\mathbf{g}}_{1}(i, j) - \overline{\mathbf{g}}_{2}(i, j))$$

em que

$$I_{ij} = \begin{cases} -1, & \overline{\mathbf{g}}_1(i,j) - \overline{\mathbf{g}}_2(i,j) \leq 0 \\ 1, & \overline{\mathbf{g}}_1(i,j) - \overline{\mathbf{g}}_2(i,j)) > 0 \end{cases}$$

Portanto,

$$W(\mathbf{g}_1, \mathbf{g}_2) = \sum_{i < j} |\overline{\mathbf{g}}_1(i, j) - \overline{\mathbf{g}}_2(i, j)|$$

Nessa dissertação provamos na Proposição 3.2.4 que a distribuição de W, sob H_0 , converge para a soma do módulo de variáveis aleatórias Normais quando o tamanho das amostras crescem.

Proposição 3.2.4. Considere duas amostras aleatórias independentes de grafos $\mathbf{g}_1 = (g_{11}, \dots, g_{1n})$ e $\mathbf{g}_2 = (g_{21}, \dots, g_{2m})$. Tome $p_{ij} = \mathbb{P}(X_{ij} = 1)$ e $Y_{ij} = \overline{\mathbf{g}}_1(i,j) - \overline{\mathbf{g}}_2(i,j)$. Sob a hipótese nula H_0 vale

que

$$W(\mathbf{g}_1, \mathbf{g}_2) = \sum_{i < j} |Y_{ij}| \qquad onde \qquad \frac{Y_{ij}}{\sqrt{p_{ij}(1 - p_{ij}) \left[\frac{1}{n} + \frac{1}{m}\right]}} \xrightarrow[m \to \infty]{D} N(0, 1)$$

Demonstração.

Dadas as amostras $\mathbf{g}_1 = (g_{11}, \dots, g_{1n})$ e $\mathbf{g}_2 = (g_{21}, \dots, g_{2m})$, para $1 \leq i < j \leq |V|$ tem-se que $X_{ij}^{11}, \dots, X_{ij}^{1n}$ são identicamente distribuídas com distribuição $Ber(p_{ij})$. Sob \mathbf{H}_0 , segue que $X_{ij}^{21}, \dots, X_{ij}^{2m}$ também são identicamente distribuídas com distribuição $Ber(p_{ij})$.

Fixado o par de vértices $(i,j), 1 \le i < j \le |V|$, segue que

$$\frac{\sum_{k=1}^{n} \frac{X_{ij}^{1k}}{n} - p_{ij}}{\sqrt{\frac{p_{ij}(1 - p_{ij})}{n}}} \xrightarrow[n \to \infty]{D} N(0, 1) \qquad e \qquad \frac{\sum_{k=1}^{m} \frac{X_{ij}^{2k}}{m} - p_{ij}}{\sqrt{\frac{p_{ij}(1 - p_{ij})}{m}}} \xrightarrow[m \to \infty]{D} N(0, 1)$$

Fazendo $Y_{ij} = \overline{\mathbf{g}}_1(i,j) - \overline{\mathbf{g}}_2(i,j)$ segue que

$$\frac{Y_{ij}}{\sqrt{p_{ij}(1-p_{ij})\left[\frac{1}{n}+\frac{1}{m}\right]}} \xrightarrow[m\to\infty]{} N(0,1)$$

Pela proposição (3.2.3) temos que $W(\mathbf{g}_1, \mathbf{g}_2) = \sum_{i < j} |Y_{ij}|$.

Capítulo 4

Simulações

Neste capítulo é apresentado o estudo do desempenho do teste de hipóteses proposto nesta dissertação via simulações. Além disso, comparações com outros testes de hipóteses encontrados na literatura são feitas com o objetivo de ampliar as discussões dos resultados obtidos para o teste de hipóteses proposto. Os programas utilizados nas simulações foram implementados no programa R e os principais programas podem ser encontrados no Apêndice A.

4.1 Simulação de grafos aleatórios

As amostras de grafos aleatórios utilizadas no estudo do desempenho do teste de hipóteses proposto são obtidas via simulação. Entre os modelos de grafos aleatórios apresentados no Capítulo 2 dois recebem maior enfoque, sendo esses o modelo de Erdős-Rényi e o modelo Exponencial. Nesta Seção é discutida a simulação de grafos aleatórios desses modelos.

4.1.1 Modelo de Erdős-Rényi

No modelo de Erdős-Rényi a probabilidade de existir uma aresta entre dois vértices do grafo é fixada e independente das outras arestas, ou seja, dado um grafo g com N vértices temos que $\{X_{ij}\}_{1 < i < j < N}$ são variáveis independentes e identicamente distribuídas com distribuição de Bernoulli(p). Portanto, para simular um grafo do modelo de Erdős-Rényi com parâmetro p basta que as variáveis $\{X_{ij}\}_{1 < i < j < N}$ sejam simuladas da distribuição de Bernoulli(p).

Exemplo 4.1.1. Deseja-se simular um grafo com 4 vértices do modelo de Erdős-Rényi com parâmetro p = 0.35. Para isso, as entradas da matriz de adjacência do grafo são simuladas a partir da distribuição de Bernoulli(0.35), totalizando assim $\binom{4}{2}$ variáveis simuladas. O resultado obtido foi

$$\boldsymbol{X} = \left(\begin{array}{cccc} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{array}\right)$$

4.1.2 Modelo Exponencial

Para obter uma amostra de grafos do modelo Exponencial pode-se utilizar métodos de simulação de Monte Carlo baseados em cadeia de Markov, conhecidos pela sigla MCMC. O objetivo desse método é construir um algoritmo de simulação de uma cadeia de Markov cuja distribuição estacionária é a distribuição da qual deseja-se simular.

É discutida nesta Seção a aplicação dos principais algoritmos do método MCMC para simulação de grafos do modelo Exponencial, conhecidos como amostrador de Gibbs e Metropolis—Hastings. Para mais detalhes sobre o método MCMC de simulação o leitor pode consultar Gamerman e Lopes (2006).

O amostrador de Gibbs é um algoritmo em que a cada iteração é selecionado, aleatóriamente, um par de vértices (i, j) do grafo e é decidido se $X_{ij} = 1$ ou $X_{ij} = 0$ de acordo com as probabilidades condicionais do grafo aleatório. Sendo $\Delta(\mathbf{S}(\mathbf{x}))_{ij} = \mathbf{S}(\mathbf{x}_{ij}^+) - \mathbf{S}(\mathbf{x}_{ij}^-)$, as probabilidades condicionais são dadas por

$$\mathbb{P}(X_{ij} = 1 | \mathbf{X}_{-ij} = \mathbf{x}_{-ij}) = \frac{\mathbb{P}(X_{ij} = 1, \mathbf{X}_{-ij} = \mathbf{x}_{-ij})}{\mathbb{P}(\mathbf{X}_{-ij} = \mathbf{x}_{-ij})} = \frac{\mathbb{P}(\mathbf{X}_{ij}^{+} = \mathbf{x}_{ij}^{+})}{\mathbb{P}(\mathbf{X}_{ij}^{+} = \mathbf{x}_{ij}^{+}) + \mathbb{P}(\mathbf{X}_{ij}^{-} = \mathbf{x}_{ij}^{-})}$$

$$= \frac{exp(\boldsymbol{\theta}^{T} \mathbf{S}(\mathbf{x}_{ij}^{+}))}{exp(\boldsymbol{\theta}^{T} \mathbf{S}(\mathbf{x}_{ij}^{+})) + exp(\boldsymbol{\theta}^{T} \mathbf{S}(\mathbf{x}_{ij}^{-}))} = \frac{exp(\boldsymbol{\theta}^{T} \Delta(\mathbf{S}(\mathbf{x}))_{ij})}{1 + exp(\boldsymbol{\theta}^{T} \Delta(\mathbf{S}(\mathbf{x}))_{ij})}$$

$$(4.1)$$

As iterações para simular um grafo Exponencial utilizando o amostrador de Gibbs são descritas abaixo.

- 1. Tome uma matriz de adjacência inicial $\mathbf{x}^{(0)}$ Na k-ésima iteração faça:
- 2. Selecione aleatóriamente um par de vértices (i, j) do grafo
- 3. Decida se $\mathbf{x}_{ij}^{(k)} = 1$ ou $\mathbf{x}_{ij}^{(k)} = 0$ de acordo com a distribuição condicional de \mathbf{x}_{ij}^k dado $\mathbf{x}_{-ij}^{(k-1)}$, dada pela Equação 4.1.

No algoritmo de Metropolis-Hastings, na k-ésima iteração uma nova matriz de adjacência $\mathbf{x}^{(k)}$ é determinada e $\mathbf{x}^{(k-1)}$ é substituida por $\mathbf{x}^{(k)}$ com probabilidade dada por

$$\min\left(1, \frac{\mathbb{P}(\mathbf{X}^{(k)} = \mathbf{x}^{(k)})}{\mathbb{P}(\mathbf{X}^{(k-1)} = \mathbf{x}^{(k-1)})}\right)$$
(4.2)

A cada iteração a nova matriz de adjacência é escolhida trocando o valor de \mathbf{x}_{ij} , para algum par de vértices (i, j) escolhido aleatóriamente. Dessa forma, como a matriz proposta $\mathbf{x}^{(k)}$ difere da matriz anterior $\mathbf{x}^{(k-1)}$ por apenas uma aresta no par de vértices (i, j), segue que

$$\frac{\mathbb{P}(\mathbf{X}^{(k)} = \mathbf{x}^{(k)})}{\mathbb{P}(\mathbf{X}^{(k-1)} = \mathbf{x}^{(k-1)})} = exp(\boldsymbol{\theta}^T [\mathbf{S}(\mathbf{x}^{(k)}) - \mathbf{S}(\mathbf{x}^{(k-1)})]) = \exp(\pm \boldsymbol{\theta}^T \Delta(\mathbf{S}(\mathbf{x}^k))_{ij})$$
(4.3)

Os passos para a construção do algoritmo são descritos abaixo

- 1. Tome uma matriz de adjacência inicial $\mathbf{x}^{(0)}$ Na k-ésima iteração faça:
- 2. Selecione aleatóriamente um par de vértices (i, j) do grafo
- 3. Aceita-se a mudança desse vértice com probabilidade dada pelas Equações 4.2 e 4.3.

Observe que para os dois algoritmos não é necessário o cálculo da constante de padronização do modelo. Porém, alguns estudos experimentais desenvolvidos por Snijders (2002) mostraram que dependendo do parâmetro do modelo e do grafo inicial $\mathbf{x}^{(0)}$ os algoritmos podem tomar um longo tempo até convergirem para a distribuição do modelo.

Para simular as amostras de grafos Exponenciais estudadas nesta dissertação foi utilizado o pacote ergm, Handcock et al. (2013) e Hunter et al. (2008), do programa R. Neste pacote, os grafos são simulados utilizando o algoritmo de Metropolis-Hastings e para as simulações dos grafos utilizados nesta dissertação foi tomado o valor de burn-in igual a 100000.

4.2 Teste de hipóteses para uma única amostra de grafos

Para avaliarmos o desempenho do teste de hipóteses para uma amostra de grafos foi utilizado o modelo Exponencial dado por 4.4.

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \frac{exp(\theta_1 S_1(\mathbf{x}) + \theta_2 S_2(\mathbf{x}))}{z(\boldsymbol{\theta})}$$
(4.4)

onde

$$S_1(\mathbf{x}) = \sum_{i < j} x_{ij}$$
 e $S_2(\mathbf{x}) = \sum_{i < j < k} x_{ij} x_{jk} x_{ki}$

Note que nesse modelo as estatísticas $S_1(\mathbf{x})$ e $S_2(\mathbf{x})$ representam, respectivamente, o número de arestas e de triângulos do grafo cuja matriz de adjacência é dada por \mathbf{x} . Por isso, este modelo será chamado de modelo aresta-triângulo para facilitar o entendimento e leitura desta dissertação.

Dessa forma, para uma amostra de grafos do modelo Exponencial tem-se interesse em testar

$$\begin{cases}
H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\
H_1: \boldsymbol{\theta} \neq \boldsymbol{\theta}_0
\end{cases}$$
(4.5)

4.2.1 Distribuição empírica de W

Os estudos de simulação abordados neste capítulo são baseados na distribuição empírica da estatística W, proposta no Capítulo 3, que é obtida a partir de amostras aleatórias simuladas de modelos de grafos aleatórios.

Para realizar qualquer estudo que avalie o desempenho do teste proposto é necessário determinarmos sua região crítica para um nível de significância fixado. Para isso, a distribuição empírica de W é obtida sob a hipótese nula. No caso do teste, descrito na Seção 3.2.2, para uma única amostra de grafos a distribuição empírica de W é obtida assumindo que a amostra de grafos tem distribuição π' estabelecida em H_0 .

Para determinarmos a distribuição de W pode-se utilizar dois métodos de simulação. O primeiro método é baseado no método de simulação via Monte Carlo. O método de Monte Carlo é muito utilizado para determinar aproximações para integrais e para probabilidades a partir de um número elevado de simulações de variáveis aleatórias. Para mais detalhes desse método o leitor pode consultar Casella e Robert (2005).

Os passos para a construção da distribuição de W utilizando o método de Monte Carlo, sob H_0 , são apresentados abaixo.

- 1. Simule uma amostra de grafos do modelo Exponencial $\mathbf{g}=(g_1,\ldots,g_n)$ com parâmetro $\boldsymbol{\theta}_0$
- 2. Calcule a estatística $W(\mathbf{g})$
- 3. Repita os passos 1 e 2 um número fixado S de vezes

O segundo método consiste na utilização do resultado obtido na Proposição 3.2.2, em que a distribuição de W pode ser obtida através dos seguintes passos

- 1. Seja $p_{ij} = \mathbb{P}(X_{ij} = 1|\mathbf{H}_0)$
- 2. Simule $Y_{ij} \sim N\left(0, \frac{p_{ij}(1-p_{ij})}{n}\right)$, para $1 \leq i \leq j \leq N$
- 3. Faça $w = \sum_{i < j} |Y_{ij}|$
- 4. Repita os passos 2 e 3 um número fixado S de vezes

Para construir a região crítica do teste definida em (3.6) fixa-se um nível de significância α e calcula-se o valor de $q_{(1-\alpha)}$ de modo que $\mathbb{P}(W > q_{(1-\alpha)}|H_0) = \alpha$. Como a distribuição de W não é conhecida utiliza-se a sua distribuição empírica para obter o quantil de ordem $(1-\alpha)$ da distribuição, determinando assim $q_{(1-\alpha)}$.

A distribuição de W obtida utilizando os dois métodos descritos acima estão representadas na Figura 4.1.

4.2.2 Poder do teste de hipóteses

Um meio de avaliar o desempenho do teste de hipóteses proposto é através de sua função poder. Como apresentado na Seção 3.1.1, há interesse em verificar se o teste de hipóteses proposto nesta dissertação é consistente, ou seja, quando a função poder, obtida sob H_1 , tender a 1 na medida que o tamanho da amostra cresce.

O poder do teste de hipóteses foi determinado utilizando o método de Monte Carlos e foi estudado para grafos gerados pelo modelo Exponencial aresta-triângulo. Nesse caso, a função poder do teste para um nível de significância α é dada por

$$\varphi(\boldsymbol{\theta}) = \mathbb{P}(\text{rejeitar } \mathbf{H}_0|\mathbf{H}_1) = \mathbb{P}(W > q_{(1-\alpha)}|\boldsymbol{\theta} \neq \boldsymbol{\theta}_0)$$
 (4.6)

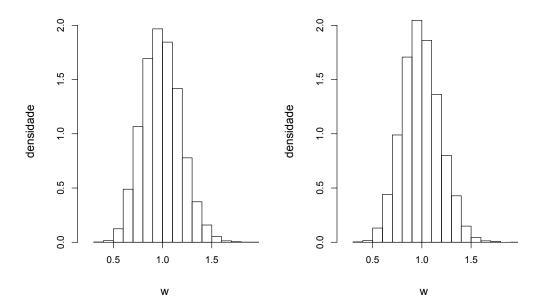


Figura 4.1: Distribuição empírica da estatística W, sob H_0 , utilizando o modelo Exponencial descrito por 4.4 para $\theta_0 = (0.3, 0.08)$. A figura da direita foi obtida utilizando o método de Monte Carlo e na figura da esquerda utilizou-se o método baseado na Proposição 3.2.2.

Os passos para obter o poder do teste via simulação são descritos abaixo.

- 1. Determinar a distribuição de W, como descrito em 4.2.1, e obter $q_{(1-\alpha)}$
- 2. Simule uma amostra de grafos exponenciais ${\bf g}$ com parâmetro ${\boldsymbol \theta}_1$
- 3. Calcule $w(\mathbf{g})$

Repita os passos (2) e (3) S vezes

4. Faça
$$\varphi(\pmb{\theta}_1) = \frac{|\{w: w(\mathbf{g}) > q_{1-\alpha}\}|}{S}$$

Para o modelo Exponecial aresta-triângulo o poder do teste de hipóteses foi computado utilizando a hipótese nula em que $\theta_0 = (0.3, 0.08)$ e grafos com 6 vértices. Diferentes amostras foram obtidas variando o valor do parâmetro θ referente à estatística S_1 , que representa o número de arestas do grafo. Pode-se observar, Figura 4.2, que o poder do teste é maior para amostras com mais grafos e torna-se pequeno para valores de θ próximos ao valor desse parâmetro sob H_0 .

4.2.3 Taxas de rejeição

Outro meio de avaliar o desempenho do teste proposto é através das taxas de rejeição. As taxas de rejeição referem-se a proporção de vezes que o teste de hipóteses rejeitou H_0 para diferentes amostras de grafos geradas de uma mesma distribuição.

As taxas de rejeição do teste, para um nível de significância α , são obtidas utilizando o método de Monte Carlo utilizando os passos descritos abaixo.

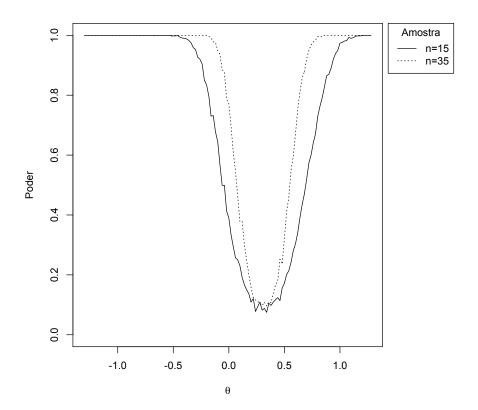


Figura 4.2: Poder do teste de hipóteses para uma amostra de grafos do modelo Exponencial aresta-triângulo para $\theta_0 = (0.3, 0.08)$ e $\theta_1 = (\theta, 0.08)$.

- 1. Determinar a distribuição de W, como descrito em 4.2.1, e obter $q_{(1-\alpha)}$
- 2. Gere uma amostra de grafos \mathbf{g} com distribuição π'
- 3. Calcule $w_{\text{obs}} = w(\mathbf{g})$
- 4. Rejeite a hipótese nula com nível α se $w_{\rm obs}>q_{(1-\alpha)}$ Repita os passos acima um número S de vezes
- 5. Calcule a taxa de rejeição para cada nível fazendo $\frac{|\{w_{\text{obs}}:w_{\text{obs}}>q_{1-\alpha}\}|}{S}$

Para o modelo aresta-triângulo foram determinadas as taxas de rejeição do teste proposto, para as hipóteses descritas em (4.5), com $\theta_0 = (0.3, 0.08)$ e para amostras de grafos com parâmetro $\theta = (0.8, 0.07)$ e $\theta = (-0.9, 0.07)$. Pela Tabela 4.1, observa-se que as taxas de rejeição para a amostra originada da distribuição aresta-triângulo com parâmetro $\theta = (0.8, 0.07)$ são maiores para amostras maiores e para grafos com mais vértices. Além disso, no caso em que $\theta = (-0.9, 0.07)$ a hipótese nula sempre foi rejeitada. Esse fato pode ser justificado ao analisar a Figura 4.3, pois é possível verificar as diferenças entre amostras de quatro grafos do modelo aresta-triângulo com os parâmentro em estudo.

1

1

1

0.887

0.997

0.991

0.968

		n	n=35		
Vértices	α	$\theta = (0.8, 0.07)$	$\theta = (-0.9, 0.07)$	$\theta = (0.8, 0.07)$	$\theta = (-0.9, 0.07)$
	0.1	0.673	1	0.984	1
6	0.05	0.54	1	0.970	1

1

1

1

1

Tabela 4.1: Taxas de rejeição do teste proposto para uma amostra de grafos do modelo aresta-triângulo, em que na hipótese nula $\theta_0 = (0.3, 0.08)$.

4.3 Teste de hipóteses para duas amostras de grafos

4.3.1 Distribuição empírica de W

0.01

0.1

0.05

0.01

7

Para determinar a distribuição da variável W, sob a hipótese H_0 , para o teste de hipóteses para duas amostras de grafos é utilizado o método Bootstrap. Esse método é um método de reamostragem que consiste em obter a estimativa de alguma função da distribuição de probabilidade de uma variável aleatória a partir de uma amostra extraída dessa distribuição. Para mais detalhes sobre esse método o leitor pode consultar Efron e Tibshirani (1994).

Os passos para a simulação da distribuição empírica de W, sob H_0 , são elencados abaixo para duas amostras $\mathbf{g}_1 = (g_{11}, \dots, g_{1n})$ e $\mathbf{g}_2 = (g_{21}, \dots, g_{1m})$.

1. Reamostre $(g_{11}^*, \dots, g_{1n}^*)$ de $(g_{11}, \dots, g_{1n}, g_{21}, \dots, g_{1m})$

0.350

0.773

0.660

0.426

- 2. Reamostre $(g_{21}^*, \ldots, g_{2m}^*)$ de $(g_{11}, \ldots, g_{1n}, g_{21}, \ldots, g_{1m})$
- 3. Calcule a estatística $W(\mathbf{g}_1^*, \mathbf{g}_2^*)$
- 4. Repita os passos 1, 2 e 3 um número fixado S de vezes

O valor crítico, para um nível de significância α , é obtido utilizando o quantil de ordem $(1-\alpha)$ da distribuição empírica de W.

4.3.2 Taxas de rejeição do teste

Estudos do desempenho do teste proposto em dados simulados de grafos aleatórios são apresentados nesta Seção e, para isso, assume-se que a distribuição empírica de W seja obtida como descrito na Seção 4.3.1.

A taxa em que o teste de hipóteses proposto rejeita a hipótese nula para um nível de significância α é calculada utilizando o método de Monte Carlo como descrito nos passos abaixo.

- 1. Gere duas amostras de grafos aleatórios \mathbf{g}_1 e \mathbf{g}_2 , com distribuições π e π'
- 2. Calcule $w_{\text{obs}} = w(\mathbf{g}_1, \mathbf{g}_2)$
- 3. Rejeite a hipótese nula com nível α se $w_{\text{obs}} > q_{(1-\alpha)}$
- 4. Repita os passos acima um número S de vezes

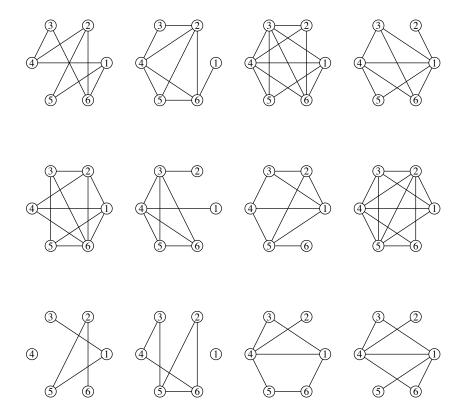


Figura 4.3: Amostra de quatro grafos, por linha, do modelo Exponencial aresta-triângulo com parâmetro $\theta = (0.3, 0.08), \theta = (0.8, 0.07)$ $\theta = (-0.9, 0.07)$.

5. Calcule a taxa de rejeição para cada nível fazendo
$$\frac{|\{w_{\rm obs}:w_{\rm obs}>q_{1-\alpha}\}|}{S}$$

Com o objetivo de verificar se o teste rejeita a hipótese nula na medida em que a diferença entre os parâmetros do modelo de Erdős-Rényi das duas amostras aumenta computou-se as taxas de rejeição do teste fixando o valor do parâmetro do modelo da primeira amostra e variando o parâmetro da segunda amostra. Além disso, afim de avaliar a influência em que o tamanho da amostra e o número de vértices dos grafos podem ter na decisão do teste, variou-se os tamanhos das amostras e o número de vértices dos grafos simulados. Com base na Tabela 4.2, pode-se observar que na medida em que o número de vértices e o tamanho das amostras aumentam a taxa de rejeição do teste também aumenta. Além disso, quando a diferença entre os parâmetros do modelo não é muito grande, no caso é igual a 0.1, é necessário que o tamanho das amostras e o número de vértices do grafo sejam grandes para que a taxa de rejeição aumente.

Para comparar os resultados obtidos para o teste de hipóteses proposto nesta dissertação, estudou-se o desempenho de outro teste, o teste de hipóteses para duas proporções. Nas simulações, como sabemos que os grafos gerados são do modelo de Erdős-Rényi, ou seja, as variáveis aleatórias $\{X_{ij}\}_{1 < i < j < N}$ que representam a existência de arestas entre os dois vértices são independentes e têm distribuição de Bernoulli, pode-se utilizar o teste de hipóteses para comparação de duas proporções por aresta. Dessa forma, para um grafo com N vértices temos um total de $\binom{V}{2}$ testes. Assim, como são utilizadas comparações múltiplas é necessário utilizar a correção de Bonferroni em que rejeita-se a hipótese nula se o p-valor do teste for menor que $\frac{\alpha}{k}$, onde k é o número

Tabela 4.2: Taxas de rejeição do teste proposto para duas amostras de grafos do modelo de Erdős-Rényi, com a primeira amostra de tamanho n e parâmetro 0.3 e segunda amostra com tamanho m e parâmetro variando entre 0.2, 0.5 e 0.75.

		n=	25 e m=1	17	n=3	35 e m = 2	27	n=4	45 e m = 3	37
Vértices	α	0.2	0.5	0.75	0.2	0.5	0.75	0.2	0.5	0.75
	0.1	0.3037	0.7716	1	0.441	0.9238	1	0.5628	0.9754	1
4	0.05	0.1949	0.6656	1	0.3144	0.8702	1	0.4405	0.9534	1
	0.01	0.072	0.4594	1	0.1382	0.7153	1	0.2284	0.8600	1
	0.1	0.451	0.9682	1	0.6542	0.9983	1	0.8090	1	1
6	0.05	0.3177	0.9344	1	0.5230	0.9958	1	0.7024	0.9996	1
	0.01	0.1234	0.8344	1	0.2757	0.9757	1	0.4803	0.9983	1
	0.1	0.5964	0.9979	1	0.8331	1	1	0.9474	1	1
8	0.05	0.4735	0.9942	1	0.7516	1	1	0.904	1	1
	0.01	0.2241	0.9972	1	0.5403	0.9996	1	0.7433	1	1

total de testes. Para este estudo adotou-se $\frac{\alpha}{\binom{V}{2}}$.

Na Tabela 4.3, são apresentadas as taxas de rejeição utilizando comparações múltiplas para amostras de grafos do modelo de Erdős-Rényi. Observando as taxas de rejeição obtidas utilizando testes de comparações múltiplas e o teste proposto observa-se que o teste proposto teve melhor desempenho, ou seja, as taxas de rejeição calculadas para o teste proposto foram maiores do que para os testes múltiplos. Com isso, o teste proposto torna-se mais vantajoso uma vez que é necessário realizar um único teste de hipóteses para obter um resultado tão bom quanto ao realizar múltiplos testes de hipóteses.

Tabela 4.3: Taxas de rejeição dos múltiplos testes para duas amostras de grafos do modelo de Erdős-Rényi, com a primeira amostra de tamanho n e parâmetro 0.3 e segunda amostra com tamanho m e parâmetro variando entre 0.2, 0.5 e 0.75.

		n=	$n{=}25 e m{=}17$		n=	=35 e m=	=27	n=45 e m=37		
Vértices	α	0.2	0.5	0.75	0.2	0.5	0.75	0.2	0.5	0.75
	0.1	0.2207	0.5885	0.9996	0.2938	0.7495	1	0.4005	0.8727	1
4	0.05	0.1192	0.4066	0.9963	0.1933	0.6521	1	0.2526	0.7652	1
	0.01	0.0175	0.1903	0.9486	0.04	0.3516	0.9985	0.075	0.4727	1
	0.1	0.1872	0.7207	1	0.3554	0.858	1	0.4813	0.9711	1
6	0.05	0.0955	0.5103	1	0.1686	0.7509	1	0.2847	0.9105	1
	0.01	0.0093	0.2176	0.9925	0.0417	0.4514	1	0.085	0.6343	1
	0.1	0.1705	0.8085	1	0.3691	0.9449	1	0.5346	0.9887	1
8	0.05	0.0842	0.641	1	0.188	0.8688	1	0.3024	0.9518	1
	0.01	0.0126	0.3053	0.9996	0.0371	0.4586	1	0.0884	0.6859	1

Entretanto, nem sempre o teste proposto apresenta melhores resultados do que o teste de comparações múltiplas. Considere que a distribuição π segue o modelo de Erdős-Rényi com parâmetro p e a distribuição π' é dada por

$$\mathbb{P}(X_{ij} = 1) = \begin{cases} p', & i = 1, j = 2\\ p, & c.c \end{cases}$$
 (4.7)

Note que no modelo descrito por (4.7) a distribuição de π' difere de um modelo de Erdős-Rényi

por apenas uma aresta, ou seja, a probabilidade de existir uma aresta para um par de vértices fixado do grafo é diferente nestes dois modelos. A partir desse novo modelo será possível verificar se o teste proposto nesta dissertação rejeita H_0 quando a distribuição das duas amostras difere apenas em um par de vértices.

Pelas Tabelas 4.4 e 4.5 nota-se que para amostras geradas do modelo de Erdős-Rényi e do modelo descrito por (4.7) os testes de comparações múltiplas apresentaram taxas de rejeição maiores do que as taxas obtidas utilizando o teste proposto. Esse fato é importante na avaliação do teste de hipóteses proposto, uma vez que, com esse resultado é possível observar que o teste proposto não é adequado para amostras de grafos com pequenas diferenças em sua distribuição. Essa discussão será abordada novamente nos estudos do poder do teste, na Seção 4.3.3.

Tabela 4.4: Taxas de rejeição do teste proposto para duas amostras de grafos. A primeira amostra de tamanho n é do modelo de Erdős-Rényi com parâmetro 0.3 e segunda amostra com tamanho m é do modelo descrito por (4.7) com parâmetro p' assumindo valores 0.3 e 0.8.

		n=25 e m=17		n=35 e	e m=27	n=45 e m=37		
Vértices	α	0.3	0.8	0.3	0.8	0.3	0.8	
	0.1	0.1267	0.4519	0.1241	0.5454	0.1393	0.6863	
4	0.05	0.0634	0.3045	0.0651	0.3871	0.069	0.5193	
	0.01	0.0113	0.1151	0.0138	0.148	0.0101	0.2097	
	0.1	0.1004	0.2641	0.0895	0.3897	0.1174	0.4381	
6	0.05	0.0387	0.1465	0.0402	0.2355	0.0565	0.2887	
	0.01	0.0047	0.0237	0.0063	0.0621	0.0094	0.0906	
	0.1	0.0907	0.1389	0.0747	0.2201	0.0966	0.2847	
8	0.05	0.0333	0.0631	0.0264	0.1051	0.0434	0.1513	
	0.01	0.0036	0.0057	0.0026	0.0215	0.0043	0.0314	

Tabela 4.5: Taxas de rejeição dos múltiplos testes para duas amostras de grafos. A primeira amostra de tamanho n é do modelo de Erdős-Rényi com parâmetro 0.3 e segunda amostra com tamanho m é do modelo descrito por (4.7) com parâmetro p' assumindo valores 0.3 e 0.8.

			e m=17	n=35 ϵ	m=27	n=45 e m=37		
Vértices	α	0.3	0.8	0.3	0.8	0.3	0.8	
	0.1	0.123	0.6278	0.1257	0.8235	0.1460	0.9298	
4	0.05	0.0655	0.5114	0.0773	0.7489	0.0831	0.8779	
	0.01	0.0095	0.2583	0.0188	0.5219	0.0216	0.7198	
	0.1	0.1013	0.5167	0.1122	0.7365	0.1261	0.8697	
6	0.05	0.0453	0.3867	0.0592	0.6395	0.0614	0.7965	
	0.01	0.007	0.1833	0.0141	0.4191	0.0124	0.6201	
	0.1	0.0982	0.4226	0.1062	0.6578	0.1157	0.8272	
8	0.05	0.0502	0.3256	0.0565	0.5572	0.0565	0.7379	
	0.01	0.0099	0.1487	0.0095	0.3449	0.0127	0.5749	

A taxa em que o teste rejeita a hipótese nula foi estudada para grafos em que os pares de vértices são independentes. Para avaliar o caso em que a existência de aresta entre pares de vértices depende das demais arestas utiliza-se o modelo Exponencial. O modelo Exponencial com parâmetro $\theta = (\theta_1, \theta_2)$ utilizado é dado pela Equação 4.8, onde as estatísticas S_1 e S_2 representam, respectivamente, o número de arestas e o número de 2-estrelas no grafo. Para simplificar, esse modelo será chamado

de modelo aresta-2estrelas.

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \frac{exp(\theta_1 S_1(\mathbf{x}) + \theta_2 S_2(\mathbf{x}))}{z(\boldsymbol{\theta})}$$
(4.8)

onde

$$S_1(\mathbf{x}) = \sum_{i < j} x_{ij}$$
 e $S_2(\mathbf{x}) = \sum_{i < j < k} x_{ij} x_{ik}$

Nas simulações, o parâmetro da primeira amostra é fixado e igual a $\theta^1 = (0.3, 0.07)$. Na tabela 4.6 foram registradas as taxas de rejeição do teste proposto para duas amostras de grafos, sendo o parâmetro da primeira amostra fixado e igual a $\theta^1 = (0.3, 0.07)$ e o parâmetro da segunda amostra θ^2 é tomado variando apenas o parâmetro θ_1^2 referente à estatística S_1 . Pela tabela, observa-se que para $\theta^2 = (0.8, 0.07)$ as taxas de rejeição foram menores, principalmente para grafos com número pequeno de vértices, do que quando tomado $\theta^2 = (-0.9, 0.07)$. Uma possível explicação para esse fato pode ser observada na Figura 4.4, uma vez que para amostras de grafos exponenciais com os parâmetros em estudo é possível notar, visualmente, as diferenças e semelhanças em suas estruturas.

Tabela 4.6: Taxas de rejeição do teste proposto para duas amostras de grafos do modelo Exponencial aresta-2 estrelas, com a primeira amostra de tamanho n e parâmetro $\theta^1 = (0.3, 0.07)$ e segunda amostra com tamanho m e parâmetro θ^2 .

-		n=25	e m=17	n=45 e m=37			
Vértices	α	$\theta^2 = (0.8, 0.07)$	$\boldsymbol{\theta}^2 = (-0.9, 0.07)$	$\theta^2 = (0.8, 0.07)$	$\boldsymbol{\theta}^2 = (-0.9, 0.07)$		
	0.1	0.338	0.991	0.648	1		
4	0.05	0.219	0.978	0.529	1		
	0.01	0.063	0.876	0.294	1		
	0.1	0.487	1	0.838	1		
6	0.05	0.328	1	0.746	1		
	0.01	0.106	1	0.498	1		
	0.1	0.64 1	0.955	1	1		
8	0.05	0.465	1	0.898	1		
	0.01	0.171	1	0.75	1		

4.3.3 Poder do teste de hipóteses

Primeiramente, o comportamento das distribuições empíricas de W, sob H_0 e H_1 , foi estudado. A Figura 4.5 representa a sobreposição das distribuições empíricas da estatística W, sob H_0 e H_1 , para duas amostras de grafos do modelo de Erdős-Rényi. Visualmente é possível notar que o poder do teste de hipóteses aumenta quando aumenta-se o tamanho das amostras de grafos, pois observa-se que para tamanho de amostras maiores, Figura 4.5(b), os histogramas estão mais distantes do que para amostras com tamanhos menores, Figura 4.5(a).

Para analisar o poder do teste de hipóteses para duas amostras de grafos \mathbf{g}_1 e \mathbf{g}_2 seguindo o modelo de Erdős-Rényi com parâmetros p e q, respectivamente, apenas o parâmetros p será fixado. A partir da região crítica definida por (3.6), para um nível de significância α fixado, a função poder

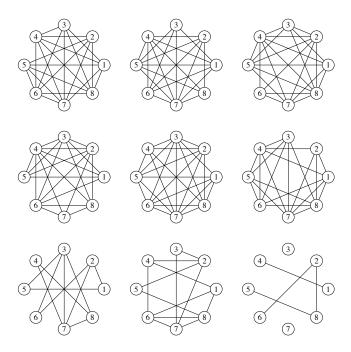


Figura 4.4: Amostra de três grafos, por linha, do modelo Exponencial aresta-2estrelas com parâmetro $\boldsymbol{\theta} = (0.3, 0.07), \ \boldsymbol{\theta} = (0.8, 0.07) \ e \ \boldsymbol{\theta} = (-0.9, 0.07).$

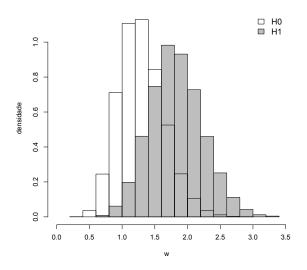
é dada por

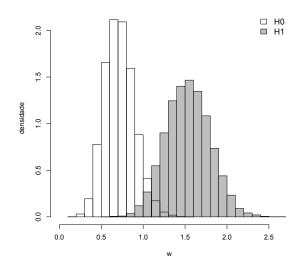
$$\varphi(q) = \mathbb{P}(W(\mathbf{g}_1, \mathbf{g}_2) > q_{(1-\alpha)}|\mathbf{H}_1) = \mathbb{P}(W(\mathbf{g}_1, \mathbf{g}_2) > q_{(1-\alpha)}|q \neq p)$$
(4.9)

A função poder é estimada pelo método de Monte Carlo seguindo os passos abaixo.

- 1. Obtenha a distribuição empírica de W, sob a hipótese nula, como descrito em (4.3.1)
- 2. Determine a região crítica do teste para um nível de significância α obtendo o quantil $q_{1-\alpha}$ de ordem $(1-\alpha)$ da distribuição empírica de W
- 3. Simule uma amostra \mathbf{g}_1 da distribuição π e outra amostra \mathbf{g}_2 da distribuição π'
- 4. Calcule $w(\mathbf{g}_1, \mathbf{g}_2)$
- 5. Repita os passos 3 e 4 um número fixado S de vezes

6. Faça
$$\varphi(q) = \frac{|\{w: w(\mathbf{g}_1, \mathbf{g}_2) > q_{1-\alpha}\}|}{S}$$





- (a) Amostras geradas com tamanho n=20 e m=15.
- (b) Amostras geradas com tamanho n=60 e m=55.

Figura 4.5: Distribuição empírica da estatística W, sob a hipótese nula e alternativa, para duas amostras de grafos Erdős-Rényi com 5 vértices e parâmetros p=0.7 e q=0.55.

Inicialmente, a função poder do teste foi estudada para diferentes tamanhos amostrais. Na Figura 4.6, nota-se que o poder do teste é próximo de 1 na medida em que a diferença entre os parâmetros das duas amostras aumenta e assume valores próximos de zero quando essa diferença é pequena. O teste torna-se mais poderoso tanto quando o tamanho das amostras cresce, Figura 4.6, quanto quando aumenta-se o número de vértices dos grafos, Figura 4.7.

Para o modelo descrito em (4.7) o poder do teste de hipóteses proposto foi calculado para diferentes valores de p'. Pela Figura 4.8 é possível observar que para amostras do modelo de Erdős-Rényi e do modelo de Erdős-Rényi com uma aresta de diferença é necessário que o tamanho das amostra seja grande para que o poder do teste seja próximo de 1. Dessa forma, dependendo do tamanho das amostras o teste de hipóteses proposto pode não ser adequado para avaliar amostras de grafos com pequenas diferenças, pois a probabilidade de cometer o erro tipo II, aceitar H_0 quando H_1 é verdadeira, é grande para tamanhos de amostras pequenos.

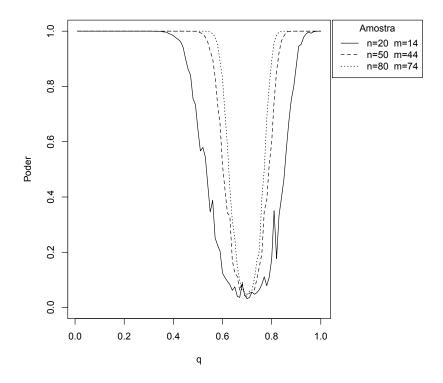


Figura 4.6: Poder do teste obtido para amostras de grafos com 5 vértices do modelo de Erdős-Rényi. A primeira amostra possui parâmetro fixado igual a 0.7 e a segunda amostra possui o parâmetro q variável.

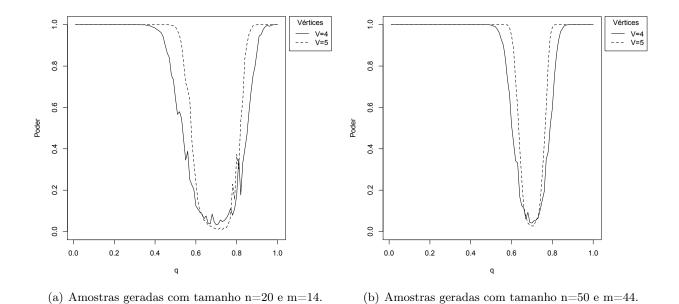


Figura 4.7: Poder do teste obtido para amostras de grafos do modelo de Erdős-Rényi. A primeira amostra foi gerada pelo modelo com parâmetro fixado e igual a 0.7 e o parâmetro q da segunda amostra variável.

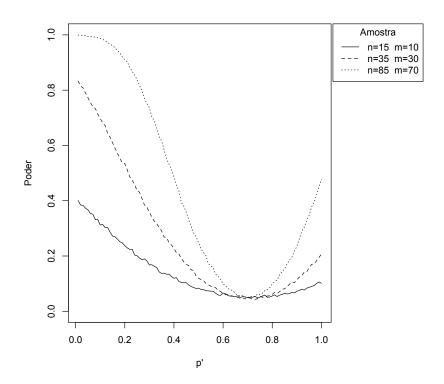


Figura 4.8: Poder do teste obtido para amostras de grafos com 5 vértices do modelo de Erdős-Rényi com uma aresta de diferença. A primeira amostra possui parâmetro fixado igual a 0.7 e a segunda amostra possui o parâmetro de apenas um par de vértices p' variável.

Capítulo 5

Aplicação

Até este ponto foram discutidos alguns resultados teóricos obtidos para o teste de hipóteses proposto nesta dissertação e o seu desempenho em dados simulados de grafos aleatórios. Neste capítulo são apresentados os resultados obtidos utilizando o teste de hipóteses para dados de eletroencefalograma.

5.1 Dados

Eletroencefalograma (EEG) é uma técnica utilizada para medir a atividade neural através do monitoramento de sinais elétricos do cérebro. Esses sinais são captados utilizando eletrodos fixados no escalpo dos indivíduos. O EEG possui boa resolução temporal, porém, em comparação com outras técnicas, sua resolução espacial é inferior. Isso deve-se ao fato dos eletrodos capturarem sinais elétricos de populações de neurônios próximas.

Os dados de EEG analisados nesta dissertação foram primeiramente analisados em Saunier et al. (2013). O estímulo utilizado no experimento é composto por 10 pontos luminosos brancos, em um fundo preto, que representam 10 marcadores do corpo humano (cabeça, ombro, cotovelo, mão, quadril, joelho e tornozelo). A animação desses pontos luminosos permite a visão de um indivíduo caminhando.

O estímulo tem duração de 5200ms e é composto por 3 fases diferentes: a fase vísivel (0 - 1600ms) representa o indivíduo caminhando, a fase de oclusão (1600ms - 3900ms) em que os pontos luminosos desaparecem permitindo a visão de que o indivíduo desaparece por trás de uma parede e a fase de reaparecimento (3900ms - 5200ms) em que o indivíduo torna-se visível novamente e continua caminhando.

O outro estímulo que foi utilizado no experimento consiste na modificação da posição dos pontos luminosos brancos, fazendo com que a animação perca a impressão de um indivíduo caminhando. Esse estímulo será chamado de movimento não biológico. Os estímulos utilizados no experimento podem ser observados na Figura 5.1.

A atividade do EEG foi registrada utilizando BrainNet BNT 36 (EMSA) que consiste em vinte Ag/AgCl eletrodos distribuídos no escalpo do indivíduo como representado na Figura 5.2.

42 APLICAÇÃO 5.1

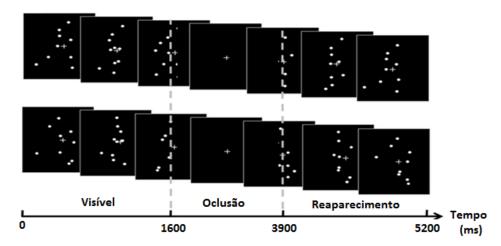


Figura 5.1: Estímulo utilizado no experimento e suas três fases: visível (0 - 1600ms), oclusão (1600ms - 3900ms) e reaparecimento (3900ms - 5200ms). A figura superior representa o estímulo para o movimento biológico e a inferior o movimento não biológico.

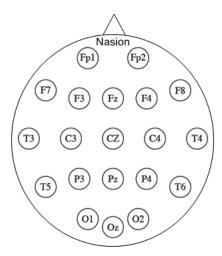


Figura 5.2: Representação da posição dos eletrodos no escalpo do indivíduo.

Para investigar a resposta cerebral ao estímulo, Figura 5.1, cada indivíduo assistia a animação 25 vezes enquanto o sinal obtido através do EEG era registrado. A partir dessas repetições, foram obtidas as curvas dos potenciais relacionados a eventos¹ (ERP) que correspondem ao sinal médio das repetições obtido para cada eletrodo. Com o objetivo de comparar as propriedades do ERP para o estímulo biológico e não biológico em Saunier et al. (2013) foram realizados testes de hipóteses pareados t, para comparação de médias, entre as condições para cada eletrodo. Observe que nessa abordagem os eletrodos são estudados separadamente, com isso é possível identificar em quais eletrodos os sinais obtidos pelo EEG apresentam diferenças.

Em estudos posteriores do mesmo conjunto de dados, Fraiman et al. (2014) propôs a utilização de grafos de interações para estudar as diferenças entre as condições biológico e não biológico. Para isso, determinou-se a matriz de correlação de Spearman Σ_t entre os eletrodos para cada janela temporal [t, t+333ms]. Dessa forma, para cada série temporal do par de eletrodos e_t^i e e_t^j , com $t \in \{1, 2, ..., T\}$ a correlação de Spearman é calculada, pela Equação (5.1), através da correlação

¹do inglês, event related potencials

de Pearson entre os postos x_t e y_t , com $x_t, y_t \in \{1, 2, \dots, T\}$, das sérires e_t^i e e_t^j , respectivamente.

$$\rho_{ij} = \frac{\sum_{t=1}^{T} (x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^{T} (x_t - \overline{x})^2 \sum_{t=1}^{T} (y_t - \overline{y})^2}}$$
(5.1)

Para a construção dos grafos de interações foi fixado o número de arestas presentes nos grafos. Assim, para um número fixado de arestas, por exemplo 30 arestas, foi definido que para os 30 pares de eletrodos com maior valor absoluto de correlação existiria uma aresta no grafo de interações. Para a análise dos grafos para as diferentes condições foram estudadas as propriedades globais e não globais desses grafos, como por exemplo, tamanho do menor caminho, coeficiente de agrupamento. Essas propriedades foram descritas na Seção 2.1.1. Observe que nessa metodologia também é possível distinguir para quais pares de eletrodos as duas condições estudadas apresentam diferenças. Os resultados obtidos em Saunier et al. (2013) e Fraiman et al. (2014) serão discutidos na Seção 5.3, juntamente com os resultados obtidos nesta dissertação.

Nas análises realizadas nesta dissertação foram utilizadas as correlações obtidas em Fraiman et al. (2014) para fases visível e oclusão das condições biológico e não biológico. O número de indivíduos e repetições do experimento são dados na Tabela 5.1.

Indivíduo Total Biológico Visível Não biológico Biológico Oclusão Não biológico

Tabela 5.1: Número de repetições do experimento.

Como a matriz de correlação de Spearman Σ_t entre os eletrodos foi determinada para cada janela temporal [t, t+333ms], para valores de t variando a cada 16.66ms, obteve-se um total de 80 janelas para a fase visível e 100 janelas para a fase oclusão. Como essas janelas são sobrepostas, optou-se por tomar janelas disjuntas para cada fase para as análises. As janelas selecionadas e o intervalo de tempo dessas janelas em ms estão descritas na Tabela 5.2.

Tabela 5.2: Janelas temporais em ms.

	J_1	J_{21}	J_{41}	J_{61}	J_{81}
Visível	[1.7, 335]	[335, 668.3]	[668.3, 1001.7]	[1001.7, 1335]	-
Oclusão	[1601.7,1935]	[1935,2268.3]	[2268.3,2601.7]	[2601.7,2935]	[2935,3268.3]

5.2 Construção dos grafos de interações

A partir da matriz de correlação de Spearman, utiliza-se um critério para a construção dos grafos de interações entre eletrodos. Esse critério pode ser tomado fixando o número de arestas do grafo ou adotando um valor limiar, onde para pares de eletrodos com correlação maior que esse valor assume-se a existência de aresta no grafo. Os grafos obtidos dependem do critério que será

44 APLICAÇÃO 5.2

tomado e nesta dissertação optou-se por adotar o critério de um valor limiar para a construção dos grafos de interações.

Primeiramente, para cada par de eletrodos será tomado um valor limiar distinto que considere a característica da curva de correlações ao longo das janelas temporais desse par de eletrodos.

Além disso, serão definidos valores limites distintos para as fases visível e oclusão. Para compreender essa escolha observe pela Figura 5.3 que as fases vísivel e oclusão apresentam valores grandes de correlação entre o par de eletrodos Pz e Oz. Se fosse tomado um valor limite igual para as duas curvas de correlação, por exemplo 0.8, os grafos construídos por volta da trigésima janela temporal não teriam arestas entre os pares de eletrodos Pz e Oz para a fase de oclusão e teriam para a fase visível. Porém, ao considerar apenas a curva da fase de oclusão o valor das correlações por volta da trigésima janela são próximos ao valor máximo das correlações de todo o intervalo temporal. Assim, ao considerar apenas a curva de correlação da fase de oclusão para determinar o valor limiar, os grafos construídos para as janelas temporais por volta da trigésima janela teriam arestas para esse par de eletrodos. Portanto, ao adotar um valor limite igual para as duas fases é possível que as diferenças entre os grafos para a fase visível e oclusão sejam induzidas por essa escolha.

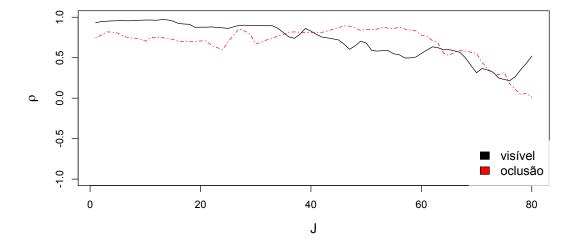


Figura 5.3: Correlações ao longo das janelas temporais entre os eletrodos Pz e Oz para a primeira repetição do experimento durante as fases visível e oclusão do movimento biológico realizado no indivíduo 1.

Por fim, ao analisar a curva de correlação para a fase vísivel em diferentes repetições do experimento observa-se, Figura 5.4, grandes diferenças entre essas curvas. Desse modo, optou-se por determinar um valor de corte diferente para cada repetição do experimento, considerando assim as características da curva de correlação.

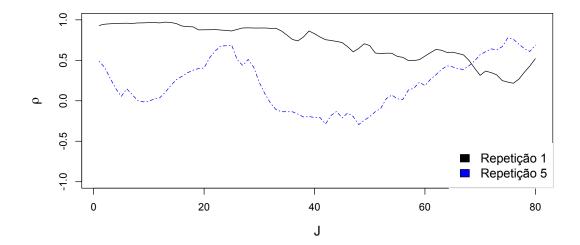


Figura 5.4: Correlações ao longo das janelas temporais entre os eletrodos Pz e Oz para a fase visível para a primeira e quinta repetição do experimento realizado no indivíduo 1.

Até este ponto, definiu-se que os valores limites tomados para a construção do grafo de interações entre os eletrodos serão tomados para cada indivíduo, par de eletrodos, repetição do experimento e para cada uma das fases visível e olcusão. Agora, é necessário definir como esse valor limite será adotado.

Com o objetivo de utililizar a característica da curva de correlação ao longo das janelas temporias para definir um valor limiar para a construção dos grafos e estabelecendo que valores absolutos altos de correlação correspondem a existência de arestas no grafo foi escolhido utilizar o primeiro e terceiro quartil das correlações para selecionar os valores limites.

Porém, o valor do primeiro e terceiro quartil não são suficientes para determinar os valores limiares. Observe nas duas figuras superiores da Figura 5.5 que ao determinar o valor do primeiro quartil, por volta de zero, o grafo teria aresta entre o par de eletrodos F7 e T6 mesmo que para esse par as correlações sejam pequenas em valor absoluto. O contrário ocorre nas figuras inferiores, uma vez que o terceiro quartil assume valor negativo e pequeno. Para evitar esses problemas ao utilizar apenas o primeiro e terceiro quartil defini-se uma constante c de modo que o grafo terá uma aresta para valores de correlação que sejam pelo menos maiores em valor absoluto que c. Na Figura 5.5 foi tomado c=0.3.

Para cada indivíduo e cada uma das fases vísivel e oclusão será fixado um par de eletrodos (i, j) $(1 \le i < j \le 20)$, uma repetição r do experimento e uma janela temporal J (para a fase visível $J \in \{1, ..., 80\}$ e para oclusão $J \in \{1, ..., 100\}$).

Assim, sendo $q_1^{r,ij}$ e $q_3^{r,ij}$ o primeiro e terceiro quartil das correlações $(\rho_{ij}^{r,J_1}, \rho_{ij}^{r,J_2}, \dots, \rho_{ij}^{r,J_M})$, correspondentes ao par de eletrodos (i,j) e repetição r, definimos

$$X_{ij}^{r,J} = \begin{cases} 1, & \text{se } \rho_{ij}^{r,J} \ge \max(c, q_3^{r,ij}) \text{ ou } \rho_{ij}^{r,J} \le \min(-c, q_1^{r,ij}) \\ 0, & c.c. \end{cases}$$
 (5.2)

Portanto, para cada fase do experimento a matriz de adjacência do grafo de interações para a janela J e repetição r será dada pela matriz $\mathbf{X}^{r,J} = (X_{ij}^{r,J})_{1 \leq i < j \leq 20}$.

46 Aplicação 5.2

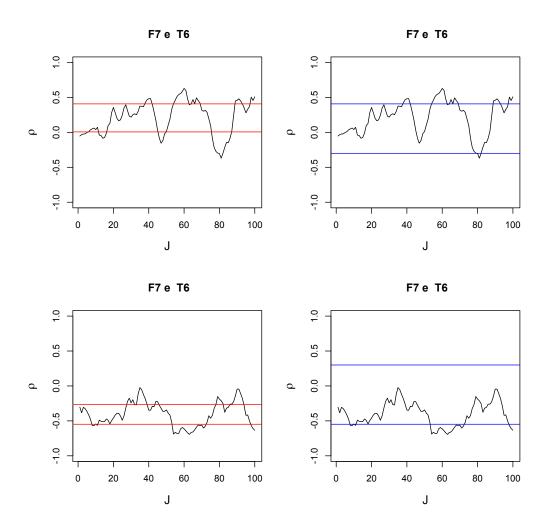


Figura 5.5: Correlações entre os eletrodos F7 e T6 das 100 janelas temporais da terceira repetição do experimento para a fase da oclusão do movimento biológico. As figuras da esquerda correspondem aos indivíduos 1 e 2 e as linhas horizontais o primeiro e terceiro quartil da série de correlações. As figuras da direita correspondem aos mesmos indivíduos e as linhas horizontais correspondem ao valor limite tomado como o máximo entre 0.3 e o terceiro quartil e o mínimo entre -0.3 e o primeiro quartil.

As amostras de grafos utilizadas serão obtidas para cada fase e janela temporal J, sendo compostas pelas matrizes de adjacência dos grafos de todos os indivíduos e repetições para a janela J.

As médias dos valores absolutos das correlações para todas as repetições e janelas temporais são dadas na Tabela 5.3. Observe que todos os valores médios são menores que 0.5 e por isso foram escolhidos os seguintes valores para a constante c para determinar os grafos: 0.3, 0.4, 0.5, 0.6, 0.7 e 0.8.

Para cada valor da constante c e para cada janela temporal selecionada os grafos para as diferentes fases foram construídos. Para o movimento biológico foram obtidas amostras de grafos de tamanho 132 para as fases visível e oclusão. Foi obtida uma amostra de tamanho 142 para a fase vísivel e tamanho 137 para a fase de oclusão para o movimento não biológico.

O número médio de arestas dos grafos presentes nas amostras para as diferentes fases do experimento e constantes c pode ser encontrado na Tabela 5.4. Note que pelo método utilizado para definir a existência de aresta entre dois vértices o conjunto de arestas dos grafos obtidos utilizando a constante c_1 contém o conjunto de arestas dos grafos obtidos utilizando c_2 , para $c_1 < c_2$. Por isso,

o número médio de arestas dos grafos é maior para valores de constantes menores.

Tabela 5.3: Média do valor absoluto das correlações.

	Indivíduo	1	2	3	4	6	7	8	Média
Visível	Biológico	0.381	0.440	0.511	0.473	0.480	0.466	0.412	0.395
	Não biológico	0.393	0.429	0.518	0.449	0.453	0.457	0.407	0.388
Oclusão	Biológico	0.363	0.420	0.500	0.461	0.452	0.459	0.424	0.385
	Não biológico	0.362	0.420	0.505	0.442	0.468	0.450	0.416	0.383

Tabela 5.4: Média do número de arestas das amostras de grafos construídos utilizando a constante c para as janelas temporais selecionadas.

				Janela		
	c	1	21	41	61	81
	0.3	40.45	33.30	33.65	30.19	-
	0.4	36.21	29.60	29.77	26.59	-
Visível - Biológico	0.5	31.25	25.41	25.54	22.85	-
Visiver - Diologico	0.6	25.62	20.51	20.36	18.22	-
	0.7	18.77	14.73	14.71	12.65	-
	0.8	10.64	7.9	8.26	6.57	-
	0.3	33.78	34.07	33.06	35.43	30.73
	0.4	29.61	29.73	29.15	31.18	26.46
Oclusão - Biológico	0.5	25.07	25.47	24.57	26.55	22.29
Octusão - Biológico	0.6	19.94	20.36	19.56	21.24	17.43
	0.7	13.89	14.33	13.85	15.43	12.47
	0.8	7.47	7.71	7.18	8.67	6.89
	0.3	41.77	37.47	32.56	33.60	-
	0.4	36.98	33.13	28.26	29.30	-
Visível - Não Biológico	0.5	31.37	28.56	23.38	24.20	-
Visivei - Não Diologico	0.6	24.73	23.12	18.40	18.70	-
	0.7	17.21	16.23	12.74	12.65	-
	0.8	9.14	8.49	6.55	6.36	-
	0.3	37.85	36.08	33.71	32.66	33.46
	0.4	33.31	31.77	29.59	28.17	29.38
Oclusão - Não Biológico	0.5	28.09	27.04	24.91	23.50	25.01
Octusão - Mão Diológico	0.6	21.93	21.20	19.52	18.28	19.56
	0.7	14.93	14.63	13.97	12.69	13.40
	0.8	7.41	7.33	7.37	6.90	7.03

5.3 Resultados do teste de hipóteses para grafos e discussão

A partir dos grafos construídos tem-se interesse de verificar se as interações entre os eletrodos descritas por esses grafos são semelhantes quando os indivíduos são submetidos a diferentes estímulos visuais. No caso desse experimento há interesse em estudar as condições biólogico e não biológico. Também optou-se por estudar as fases vísivel e oclusão dentro de cada condição.

48 APLICAÇÃO 5.3

Portanto, deseja-se testar se as amostras de grafos obtidas para diferentes estímulos provém de uma mesma distribuição sobre grafos. Nesse caso, nenhuma hipótese a respeito da distribuição das amostras é assumida, assim o teste de hipóteses não paramétrico para duas amostras de grafos proposto nesta dissertação foi utilizado.

5.3.1 Teste de hipóteses para as condições biológico e não biológico

Primeiramente, o teste de hipóteses foi utilizado para uma amostra de grafos da fase visível e outra amostra para a fase oclusão, para as quatro janelas temporais selecionadas, como descrito na Tabela 5.2. Assim, foram realizados quatro testes de hipóteses para cada condição do experimento (biológico e não biológico).

O p-valor obtido em cada teste de hipóteses realizado para amostras de grafos das condições biológio e não biológico estão descritos na Tabela 5.5. Para a fase vísivel, o p-valor obtido para o teste realizado para as amostras de grafos das janelas temporais 1 (entre 1.7 ms e 335 ms) e 41 (entre 668.3 ms e 1001.7 ms) foram pequenos, em relação aos demais valores, para os diferentes valores da constante c, indicando assim diferença entre as condições para essas janelas. Para o intervalo final da fase de oclusão o p-valor obtido foi pequeno, principalmente para a janela temporal 81 (entre 2935 ms e 3268.3 ms).

Tabela 5.5: P-valor do teste de hipóteses para amostras de grafos das fases visível do movimento biológico e não biológico e para amostras de grafos para a fase oclusão para o movimento biológico e não biológico. Os valores em negrito correspondem a p-valor<0.10 e em vermelho a p-valor<0.05

Biológico e não biológico				Janela		
	c	1	21	41	61	81
	0.3	0.1842	0.4639	0.0723	0.4606	-
	0.4	0.0899	0.6914	0.0373	0.7074	-
Visível	0.5	0.1014	0.6621	0.0295	0.5910	-
Visivei	0.6	0.0781	0.5470	0.0437	0.6122	-
	0.7	0.0915	0.5578	0.0836	0.5448	-
	0.8	0.0373	0.6539	0.1819	0.4467	-
	0.3	0.7434	0.8548	0.2557	0.1682	0.0251
	0.4	0.6241	0.9157	0.4173	0.1200	0.0117
Oclusão	0.5	0.8227	0.8816	0.3764	0.1292	0.0185
Octusão	0.6	0.9421	0.9028	0.3456	0.0778	0.0433
	0.7	0.9706	0.8289	0.4190	0.0441	0.0542
	0.8	0.7486	0.2685	0.4036	0.1252	0.0851

No trabalho Saunier et al. (2013) foram encontradas diferenças significativas entre o movimento biológico e não biológico para a fase visível no eletrodo T6 no intervalo temporal de 167 ms à 245 ms e no intervalo entre 430 ms e 950 ms na região centro-parietal (eletrodos C3, Cz e Pz). Observe que esses intervalos temporais estão contidos nos intervalos temporais em que o p-valor do teste de hipóteses realizado para a fase visível foi pequeno. Como discutido nesse trabalho, as diferenças obtidas no intervalo temporal da fase vísivel que precede a fase de oclusão sugere um processo preditivo da chegada da oclusão.

Para a fase de oclusão foram encontradas diferenças significativas nos eletrodos C3, Cz, Pz e O2 no intervalo temporal de 2495 ms à 3585 ms, no eletrodo T6 no intervalo entre 2560 ms e 3600

ms e no eletrodo T5 no intervalo entre 2340 ms e 2955 ms. Observe novamente que a maior parte desdes intervalos estão contidos nas janelas 61 e 81 (entre 2601.7 ms e 3268.3 ms) em que o p-valor do teste de hipóteses foi pequeno, em comparação as outras janelas temporais.

Com o objetivo de resumir os resultados obtidos nesta dissertação e no trabalho de Saunier et al. (2013), na Figura 5.6 estão destacadas as janelas temporais em que os testes de hipóteses rejeitaram a hipótese nula, no caso do teste proposto, de os grafos provenientes das condições biológico e não biológico provém da mesma distribuição.

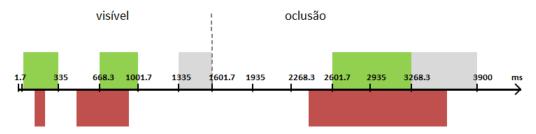


Figura 5.6: As janelas temporais destacadas em verde representam as janelas em que o teste de hipóteses proposto rejeitou a hipótese nula. As janelas temporais em vermelho representam as janelas em que foram encontradas diferenças significativas entre os eletrodos em Saunier et al. (2013). As janelas em cinza representam as regiões que não foram testadas.

Vale ressaltar que o teste de hipóteses proposto e utilizado não discrimina em qual eletrodo a diferença entre as condições para as janelas temporais selecionadas é significativa. Assim, para compreender melhor os resultados obtidos com o teste de hipóteses e para comparar com os resultados obtidos nas análises realizadas anteriormente, foram obtidos grafos selecionando as 30 arestas mais frequêntes nas amostras de grafos. Esse critério foi adotado para que fosse tomado o mesmo número de arestas para as duas amostras de grafos.

Dessa forma, nesse novo grafo a existência de uma aresta entre dois eletrodos indica que esta aresta está entre as 30 arestas mais frequêntes da amostra de grafos. Uma vez que o número médio de arestas das amostras de grafos obtidos a partir das constantes c=0.7 e c=0.8 é menor que 20 arestas (ou seja, 10% do número possível de arestas) foram consideradas apenas as amostras de grafos obtidos utilizando como constante c os valores 0.3, 0.4, 0.5 e 0.6.

Na Figura 5.7 estão representados os grafos com as 30 arestas mais frequentes nas amostras de grafos da janela 1 da fase visível. Não é possível notar diferença na presença de arestas no eletrodo T6, como foi observado em Saunier et al. (2013). Porém, é possível observar para o movimento não biológico as arestas mais frequêntes pertencem a região frontal enquanto para o movimento biológico as arestas pertencem a região occipital. Ao analisar a janela 41, Figura 5.8, nota-se que entre as 30 arestas mais frequentes para o movimento não biológico muitas arestas estão conectadas aos eletrodos temporais, principalmente aos eletrodos T5 e T6. Para o movimento biológico percebe-se que muitas arestas estão conectadas aos eletrodos O1, O2 e Oz.

50 APLICAÇÃO 5.3

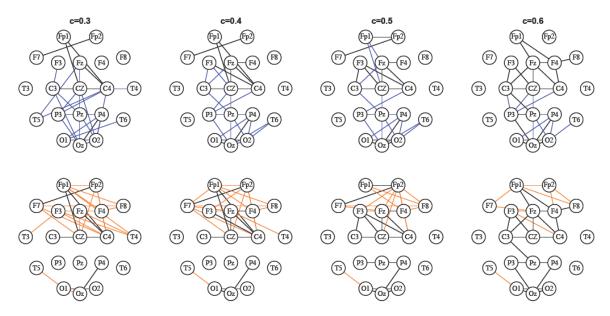


Figura 5.7: Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da janela 1 da fase visível para o movimento biológico (arestas em azul) e não biológico (arestas em laranja). As arestas em preto são as arestas em comum nas duas condições.

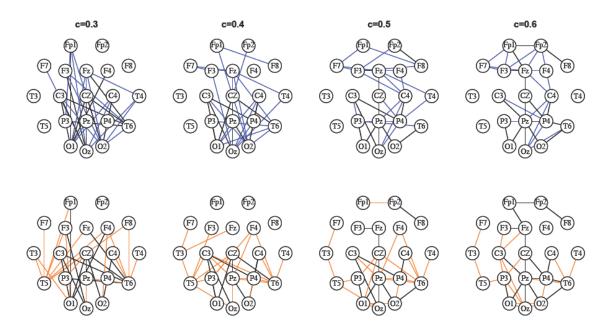


Figura 5.8: Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da janela 41 da fase visível para o movimento biológico (arestas em azul) e não biológico (arestas em laranja). As arestas em preto são as arestas em comum nas duas condições.

Em Fraiman $et\ al.\ (2014)$ foram encontradas diferenças significativas (p-valor <0.05/20) quando comparadas as propriedades locais dos grafos (grau, centralidade e coeficiente de agrupamento) para o movimento biológico e não biológico. No eletrodo F7 foram obtidas diferenças com relação ao grau e centralidade dos grafos no período de 100 ms à 750 ms. Com relação ao coeficiente de agrupamento foram obtidas diferenças no eletrodo P3 no intervalo de 150 ms à 550 ms e no eletrodo O2 de 500 ms à 850 ms. Por fim, no intervalo de 350 ms à 700 ms foram encontradas diferenças no grau do vértice do grafo referente ao eletrodo Pz. Ao comparar esses resultados com os grafos obtidos selecionando as 30 arestas mais frequentes nas amostras de grafos temos na Figura 5.7 que o eletrodo F7 possui muitas arestas para o grafo do movimento não biológico no intervalo de 1.7 ms à 385 ms e o eletrodo P3 possui poucas arestas para o movimento biológico e nenhuma aresta para o não biológico. Nota-se pela Figura 5.8 que o eletrodo O2 possui muitas arestas para o movimento biológico no intervalo de 668 ms à 1001.7 ms.

Para a fase de oclusão ao analisar as duas janelas temporais próximas ao fim dessa fase (janela 61 e 81 referente ao intervalo de 2601.7 ms à 3268.3 ms) nota-se que para a janela 61, Figura 5.9, muitas arestas estão conectadas ao eletrodo F8 e F7 para o movimento não biológico. Pela Figura 5.10, é possível observar que as 30 arestas mais frequentes nas amostras de grafos para o movimento biológico estão presentes na região frontal e, por outro lado, para o movimento não biológico as arestas estão presentes na região occipital.

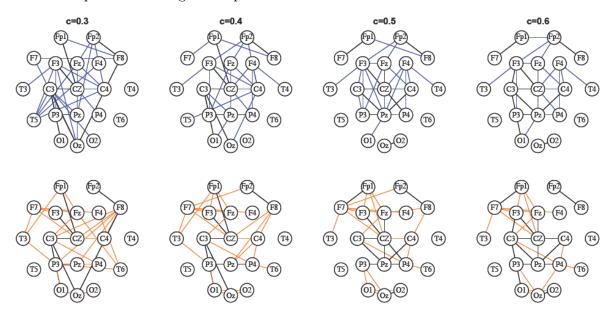


Figura 5.9: Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da janela 61 da fase oclusão para o movimento biológico (arestas em azul) e não biológico (arestas em laranja). As arestas em preto são as arestas em comum nas duas condições.

52 APLICAÇÃO 5.3

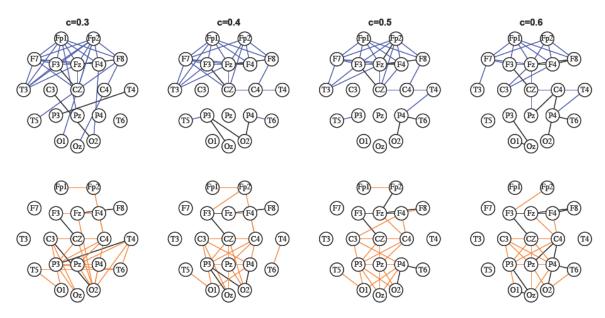


Figura 5.10: Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da janela 81 da fase oclusão para o movimento biológico (arestas em azul) e não biológico (arestas em laranja). As arestas em preto são as arestas em comum nas duas condições.

5.3.2 Teste de hipóteses para as fases visível e oclusão

Por fim, o teste de hipóteses foi realizado para comparar as fases visível e oclusão para as condições biológico e não biológico. Pela Tabela 5.6 é possível observar que tanto para o movimento biológico quanto para o não biológico na janela temporal 1 (referente ao intervalo de 1.7 ms à 335 ms na fase visível e 1601.7 ms à 1935 ms na fase oclusão) os p-valores obtidos foram pequenos para todas as constantes c tomadas. Essa diferença no período inicial das duas fases pode estar associada a presença de potenciais evocados na fase vísivel e sua ausência na fase de oclusão. Potencial evocado é um potencial elétrico que surge como resposta a um estímulo apresentado ao indivíduo, no caso desse experimento ao estímulo visiual. Em geral, esse potencial pode ser registrado no intervalo de 333 ms após o estímulo visual.

Novamente na Tabela 5.6 os p-valores para a janela temporal 61 (referente ao intervalo de 1001.7 ms à 1335 ms na fase visível e de 2601.7 ms à 2935 ms na fase de oclusão) foram pequenos para a fase vísivel comparados aos valores obtidos para a fase de oclusão. Esse resultado pode indicar a presença de processos preditivos da chegada de oclusão quando o indivíduo é submetido ao estímulo que simula o movimento biológico do caminhar de um sujeito.

Tabela 5.6: P-valor do teste de hipóteses para amostras de grafos das fases visível e oclusão para o movimento biológico e não biológico. Os valores em negrito correspondem a p-valor<0.10 e em vermelho a p-valor<0.05

			Jar	nela	
	С	1	21	41	61
	0.3	0.0059	0.4444	0.2322	0.0106
	0.4	0.0030	0.5836	0.2675	0.0204
Biológico	0.5	0.0019	0.4294	0.1984	0.0278
Diologico	0.6	0.0035	0.3827	0.3037	0.0563
	0.7	0.0019	0.4352	0.4513	0.0548
	0.8	0.0013	0.5340	0.5022	0.1338
	0.3	0.0024	0.7959	0.0407	0.6918
	0.4	0.0016	0.8777	0.0802	0.5380
Não Biológico	0.5	0.0016	0.8278	0.1249	0.6673
Nao Diologico	0.6	0.0014	0.6362	0.2222	0.5629
	0.7	0.0023	0.6046	0.1686	0.3582
	0.8	0.0066	0.5784	0.2959	0.3360

Para facilitar a compreensão de quais regiões o teste de hipóteses rejeitou a hipótese de que a distribuição de probabilidade dos grafos obtidos para a fase visível e oclusão são diferentes, na figura 5.11 as regiões em laranja representam o intervalo temporal em que o teste entre as fases visível e oclusão rejeitou a hipótese nula para a condição biológico. Já a região em azul representa o intervalo temporal em que a hipótese nula foi rejeitada para a condição não biológico para o teste entre as fases visível e oclusão.

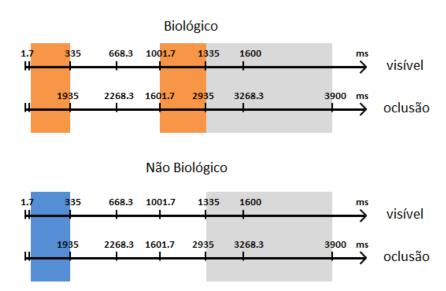


Figura 5.11: As janelas temporais destacadas em verde representam as janelas em que o teste de hipóteses proposto rejeitou a hipótese nula. As janelas temporais em vermelho representam as janelas em que foram encontradas diferenças significativas entre os eletrodos em Saunier et al. (2013). As janelas em cinza representam as regiões que não foram testadas.

Na Figura 5.12, referente ao movimento biológico, é possível observar que algumas das 30 arestas mais frequentes nas amostras de grafos na fase de oclusão estão conectadas aos eletrodos T5, F4, O2 e O1. Já para a fase vísivel as arestas mais frequentes são praticamente as mesmas para as diferentes constantes c tomadas. Para o movimento não biológico as arestas mais frequentes estão

54 APLICAÇÃO 5.3

concentradas na região frontal para a fase vísivel e para a fase de oclusão algumas arestas estão concentradas na região temporal e occiptal. É possível observar pela Figura 5.14 que muitas das arestas mais frequentes da amostra de grafos para a fase visível do movimento biológico estão conectadas ao eletrodo T6.

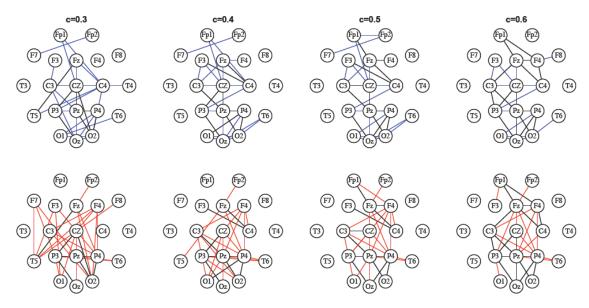


Figura 5.12: Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da janela 1 das fases visível (arestas em azul) e oclusão (arestas em vermelho) para o movimento biológico. As arestas em preto são as arestas em comum nas duas fases

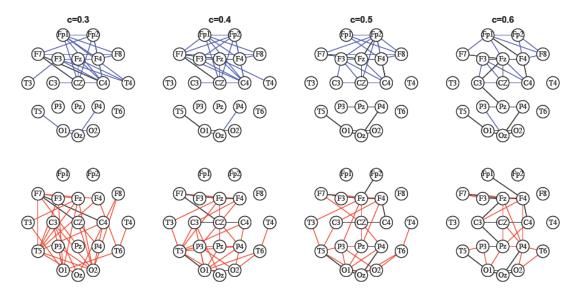


Figura 5.13: Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da janela 1 das fases visível (arestas em azul) e oclusão (arestas em vermelho) para o movimento não biológico. As arestas em preto são as arestas em comum nas duas fases.

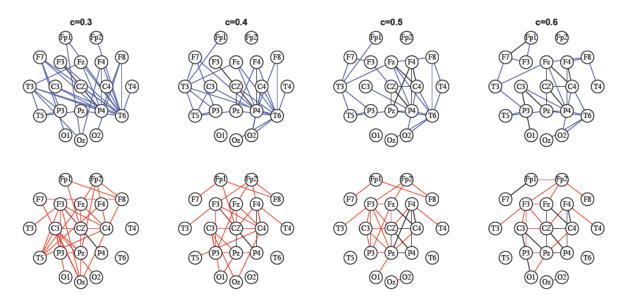


Figura 5.14: Grafos obtidos selecionando as 30 arestas mais frequentes das amostras de grafos da janela 61 das fases visível (arestas em azul) e oclusão (arestas em vermelho) para o movimento biológico. As arestas em preto são as arestas em comum nas duas fases.

APLICAÇÃO 5.3

Capítulo 6

Conclusões

Nesta dissertação foi feita uma revisão bibliográfica sobre grafos aleatórios apresentando os principais modelos de grafos aleatórios com o conjunto de vértices fixado. Com o objetivo principal de realizar testes de hipóteses a respeito das distribuições de amostras de grafos aleatórios sem considerar propriedades locais e globais desses grafos foi proposto um novo teste de hipóteses não paramétrico. Nesse teste deseja-se testar a hipótese de que duas amostras de grafos provém da mesma distribuição de probabilidade sobre grafos e no teste para uma única amostra de grafos deseja-se testar se a distribuição da qual a amostra foi originada é a mesma que a distribuição especificada na hipótese nula.

Provamos como computar de maneira eficiente a estatística do teste de hipóteses e provamos que a distribuição dessa estatística, tanto no teste para uma única amostra quanto no teste para duas amostras, converge para a soma do módulo de variáveis aleatórias normais na medida em que o tamanho da amostra cresce.

O desempenho do teste de hipóteses foi estudado a partir de amostras de grafos simuladas dos modelos de Erdos-Rényi e Exponencial. É importante ressaltar que os grafos Exponenciais foram simulados a partir de um pacote disponível no programa R que utiliza o método de Metropolis-Hastings. A convergência para a distribuição desejada desse método é discutida na literatura e depende tanto do parâmetro do modelo quanto do grafo inicial tomado. Como nesta dissertação os grafos simulados eram formados por no máximo 8 vértices, totalizando no máximo 28 possíveis arestas, foi tomado o valor para o burn-in da simulação igual a 100000. Apesar disso, é importante destacar que não há indícios de que a convergência para a distribuição desejada foi de fato verificada.

Nos estudos do desempenho do teste de hipóteses prosposto obteve-se que as taxas em que o teste rejeita a hipótese nula, quando ela é falsa, aumenta na medida em que o tamanho da amostra e o número de vértices dos grafos aumenta. O mesmo ocorre quando estuda-se o poder do teste de hipóteses que aproxima-se do valor 1 na medida em que o tamanho da amostra e o número de vértices dos grafos aumenta. Porém, ao tomar amostras de grafos do modelo de Erdos-Rényi e outra do mesmo modelo mudando apenas a probabilidade de ocorrência de uma aresta obtemos taxas de rejeição pequenas. Ao estudar o poder do teste verificamos que é necessário amostras de grafos grandes (maiores que 80 grafos) para que o poder se aproxime de 1.

Dessa forma, comprovamos que para amostras "pequenas" de grafos com diferenças presentes na ocorrência de algumas arestas o teste de hipóteses proposto não é adequado. Além disso, o teste de hipóteses proposto para duas amostras de grafos não é adequado quando o número médio de conexões para pares de vértices dessas amostras são iguais.

Os resultados teóricos do teste de hipóteses apresentado nessa dissertação dependem da distância entre grafos que foi definida na Seção 3.2.1. Como um trabalho futuro é possível estudar as propriedades do teste de hipóteses construído a partir de diferentes distâncias entre grafos. Com isso, é possível estudar se os novos testes construídos seriam adequados nos casos mencionados acima.

O teste de hipóteses foi utilizado para determinar diferenças nas interações neurais, descritas por meio de grafos originados a partir de dados de Eletroencefalograma, de indivíduos que foram submetidos a dois estímulos visuais diferentes. Como resultado foi obtido que a distribuição das amostras de grafos para o movimento biológico e não biológico são diferentes na fase vísivel no intervalo de 1.7 ms à 335 ms (p-valor < 0.1) e no intervalo de 668.3 ms à 1001.7 ms (p-valor < 0.05). Para a fase oclusão foram encontradas diferenças no intervalo de 2935 ms à 3268.3 ms (p-valor < 0.05).

Em Saunier et al. (2013) foram analisadas as curvas de potenciais relacionados a eventos de cada eletrodo das condições biológico e não biológico e em Fraiman et al. (2014) as propriedades locais dos grafos de interações para as duas condições foram testadas. Os intervalos de tempo, ou os subintervalos de tempo, em que alguns eletrodos apresentaram diferenças entre o movimento biológico e não biológico desses trabalhos estão contidos, ou contém, os intervalos descritos acima em que o teste de hipóteses proposto nessa dissertação rejeita a hipótese nula.

Com base nos intervalos de tempo em que o teste de hipóteses rejeita a hipótese de que a distribuição dos grafos para o movimento biológico e não biológico é a mesma é possível dizer que a diferença obtida no intervalo que precede a fase de oclusão sugere um processo preditivo da chegada da oclusão. Além disso, as diferenças obtidas no intervalo final da fase de oclusão sugerem também um processo preditivo, porém nesse caso para o ressurgimento da animação.

Ao analisar o grafo obtido selecionando as 30 arestas mais frequentes das amostras de grafos nota-se que para o intervalo temporal que precede o ressurgimento dos pontos luminosos (2935 ms à 3268.3 ms) as arestas estão concentradas na região frontal, responsável pelo pensamento abstrado, para o movimento biológico e estão concentradas na região occipital para o movimento não biológico. Ao contrário, na fase inicial do experimento (1.7 ms à 335 ms) as arestas estão concentradas na região frontal para o movimento não biológico que indica a atividade de abstração realizada pelos indivíduos o estímulo visual.

É importante ressaltar que, ao utilizar o teste de hipóteses proposto nesse trabalho, não é possível discriminar em quais eletrodos as diferenças entre os estímulos é significativa. Porém, é possível determinar em quais intervalos de tempo os grafos de interações dos estímulos não são originados da mesma distribuição realizando apenas um único teste de hipóteses ao invés de realizar um teste de hipóteses para cada par de eletrodos (190 testes de hipóteses no total!).

Apêndice A

Apêndice

A.1 Programas em R

Programa para calcular a estatística W para uma amostra de grafos.

```
#amostra: matriz contendo a amostra; cada linha linha da matriz contém a
diagonal superior da matriz de adjacência de um grafo da amostra
# n:tamanho da amostra
# p: prob. de existir uma aresta entre pares de vértices do grafo
calculow <- function(amostra,p){
    n <- dim(amostra)[1]
    g <- vector()
    g <- apply(amostra,2,sum)
    g <- g/n
    D <- g - p
    w <- sum(abs(D))
    return(w)
}</pre>
```

Programa para calcular a estatística W para duas amostras de grafos.

```
#amostral: matriz contendo a amostra 1; cada linha linha da matriz contém a
diagonal superior da matriz de adjacência de um grafo da amostra
# n:tamanho da amostra1
# m:tamanho da amostra2
calculow <- function(amostra1,amostra2) {
    n <- dim(amostra1)[1]
    m <- dim(amostra2)[1]
    g1 <- apply(amostra1,2,sum)
    g2 <- apply(amostra2,2,sum)
    g1 <- g1/n
    g2 <- g2/m
    D <- g1 -g2
    w <- sum(abs(D))</pre>
```

```
return (w)
}
Programa par
```

Programa para gerar uma amostra de grafos Exponenciais aresta-triângulo.

```
#Este código gera uma amostra de grafos do tipo exponencial
#m:tamanho da amostra
#v: número de vértices do grafo
#teta: vetor de paramêtros do modelo
#amostra: matriz com m linhas e v*(v-1)/2 colunas, onde cada linha
representa a diagonal superior da matriz de adjacêndia do grafo
require(ergm)
require(gdata)
amostra.exp <- function(m, v, teta) {</pre>
    num <- v*(v-1)/2
    amostra <- matrix(nrow=m, ncol=num)</pre>
    for(i in 1:m) {
        #g.use: o grafo que será usadado para gerar a amostra
        q.use <- network(v,density=0.1,directed=FALSE)</pre>
        #g.sim: o grafo simulado
        q.sim <- simulate(~edges+triangles,nsim=1,coef=teta,basis=q.use,</pre>
control=control.simulate(MCMC.burnin=100000,MCMC.interval=1000))
        matriz <- as.matrix.network(g.sim)</pre>
        #salvar apenas a diagonal superior do grafo simulado
        amostra[i,] <- upperTriangle(matriz, diag=FALSE)</pre>
     return (amostra)
}
```

Programa para gerar uma amostra de grafos Exponenciais aresta-2estrelas.

```
#Este código gera uma amostra de grafos do tipo exponencial
#m:tamanho da amostra
#v: número de vértices do grafo
#teta: vetor de paramêtros do modelo
#amostra: matriz com m linhas e v*(v-1)/2 colunas, onde cada linha
representa a diagonal superior da matriz de adjacêndia do grafo
require(ergm)
require(gdata)
amostra.exp <- function(m,v,teta){
    #n. de pares de vértices
    num <- v*(v-1)/2
    amostra <- matrix(nrow=m,ncol=num)</pre>
```

```
for(i in 1:m) {
             #q.use: o grafo que será usadado para gerar a amostra
             g.use <- network(v,density=0.1,directed=FALSE)</pre>
             #g.sim: o grafo simulado
             g.sim <- simulate(~edges+kstar(2),nsim=1,coef=teta,basis=g.use,</pre>
control=control.simulate(MCMC.burnin=100000,MCMC.interval=1000))
        matriz <- as.matrix.network(g.sim)</pre>
        #salvar apenas a diagonal superior do grafo simulado
        amostra[i,] <- upperTriangle(matriz, diag=FALSE)</pre>
    return (amostra)
}
  Programa para gerar uma amostra do modelo Erdos-Rényi.
#Este código gera uma amostra de grafos do modelo de Erdos-Rényi
#m: tamanho da amostra
#v: número de vértices do grafo
#amostra: matriz com m linhas e v*(v-1)/2 colunas, onde cada linha
representa a diagonal superior da matriz de adjacêndia do grafo
amostra <- function(m,p,v) {</pre>
    num < -(v * v - v)/2
    amostra <- matrix(nrow=m, ncol=num)</pre>
        for(1 in 1:m) {
             amostra[1,] <- rbinom(num,1,p)</pre>
        }
```

Referências Bibliográficas

- **Albert** et al. (1999) Réka Albert, Hawoong Jeong e Albert-László Barabási. Diameter of the world-wide web. Nature, 401:130–131. Citado na pág. 10
- Bullmore e Sporns (2009) E. Bullmore e O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*, 10:186–198. Citado na pág. 1
- Busch et al. (2009) J. R. Busch, P. A. Ferrari, A. G. Flesia, R. Fraiman, S. P. Grynberg e F. G. Leonardi. Testing statistical hypothesis on random trees and applications to the protein classification problem. The Annals of Applied Statistics, 3(2):542–563. Citado na pág. 17, 18
- Caimo e Friel (2011) Alberto Caimo e Nial Friel. Bayesian inference for exponentical random graph models. *Social Networks*, 33:41–55. Citado na pág. 1, 13
- Casella e Robert (2005) George Casella e Christian Robert. Monte Carlo Statistical Methods. Springer Texts in Statistical. Springer. Citado na pág. 28
- Csardi e Nepusz (2006) Gabor Csardi e Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695. URL http://igraph.sf.net. Citado na pág. 5
- **Dethlefsen e Højsgaard (2005)** Claus Dethlefsen e Søren Højsgaard. A common platform for graphical models in R: The gRbase package. *Journal of Statistical Software*, 14(17):1–12. URL http://www.jstatsoft.org/v14/i17/. Citado na pág. 5
- Efron e Tibshirani (1994) Bradley Efron e Robert J. Tibshirani. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability. Chapman and Hall. Citado na pág. 31
- Erdős e Rényi (1959) Paul Erdős e Alfréd Rényi. On random graphs i. *Publicationes Mathematicae*, 6:290–297. Citado na pág. 8
- Erdős e Rényi (1960) Paul Erdős e Alfréd Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61. Citado na pág. 8
- Fraiman et al. (2014) Daniel Fraiman, Ghislain Saunier, Eduardo F. Martins, e Claudia D. Vargas. Biological motion coding in the brain: analysis of visually-driven eeg funcional networks. *Plos One. No prelo 2014.* Citado na pág. 2, 42, 43, 51, 58
- Gamerman e Lopes (2006) Dani Gamerman e Hedibert F. Lopes. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Texts in Statistical Science. Chapman and Hall. Citado na pág. 26
- Gibbons e Chakraborti (2010) Jean D. Gibbons e Subhabrata Chakraborti. *Nonparametric Statistical Inference*. Textbooks and Monographs. Chapman and Hall. Citado na pág. 17

- Handcock et al. (2013) Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky e Martina Morris. ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks. The Statnet Project (http://www.statnet.org), 2013. URL CRAN.R-project.org/package=ergm. R package version 3.1-0. Citado na pág. 27
- Hunter et al. (2008) David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, Morris e Martina. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3). Citado na pág. 27
- Newman et al. (2001) Mark Newman, Steven Strogatz e Duncan Watts. Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E, 64:026118. Citado na pág. 10
- Newman (2010) Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press. Citado na pág. 1, 14
- R Development Core Team (2008) R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0. Citado na pág. 5
- Robins et al. (2007) Garry Robins, Pip Pattison, Yuval Kalish e Dean Lusher. An introduction to exponential random graph models (p*) for social networks. Social Networks, 29:173–191. Citado na pág. 11
- Saunier et al. (2013) Ghislain Saunier, Eduardo F. Martins, Elisa C. Dias, José M. de Oliveira, Thierry Pozzo e Claudia D. Vargas. Electrophysiological correlates of biological motion permanence in humans. *Behavioural Brain Research*, 236:166–174. Citado na pág. 41, 42, 43, 48, 49, 53, 58
- Snijders (2002) Tom A. B. Snijders. Markov chain monte carlo estimation of exponential random graph models. SJournal of Social Struture, 3. Citado na pág. 13, 27
- Wang et al. (2010) L. Wang, C. Yu, H. Chen, W. Qin, Y. He e et al. Dynamic functional reorganization of the motor execution network after stroke. *Brain*, 133:1224–1238. Citado na pág. 2
- Watts e Strogatz (1998) Duncan J. Watts e Steven H. Strogatz. Collective dynamics of 'smallworld' networks. *Nature*, 394:440–442. Citado na pág. 14
- Wu et al. (2012) Junjie Wu, Junsong Zhang, Chu Liu, Dongwei Liu, Xiaojun Ding e Changle Zhou. Graph theoretical analysis of eeg functional connectivity during music perception. Brain Research, 1483:71–81. Citado na pág. 2