Testes estatísticos semi paramétricos para discriminação de grafos

Gabriela Eleutério Soares

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

Programa: Ciência da Computação

Orientador: André Fujita

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, novembro de 2017

Testes estatísticos semi paramétricos para discriminação de grafos

Esta é a versão original da tese elaborada pela candidata Gabriela Eleutério Soares, tal como submetida à Comissão Julgadora.

Resumo

SOARES, G. E. **Testes estatísticos semi paramétricos para discriminação de grafos** 2017. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Grafos são utilizados para modelar redes em diversos campos científicos. Ao longo dos anos, o interesse em modelar redes do mundo real vem crescendo, com aplicações em diferentes áreas, como biologia molecular (redes genéticas regulatórias, redes de interação proteína-proteína), neurociência (rede funcional cerebral)e ciências sociais (redes sociais como facebook, instagram, foursquare). As redes do mundo real não podem ser modeladas por grafos determinístico adequadamente, pois tais grafos não apresentam componente aleatório desejável nestas situações. Além disso, algoritmos tradicionais e abordagem clássicas para grafos apresentam dificuldades e limitações que prejudicam tais análises. Com isso, grafos aleatórios são mais indicados para modelar as redes do mundo real e procedimentos estatísticos precisam ser formalizados com o intuito de facilitar a análise de dados destes grafos, como por exemplo: como comparar dois grafos reais? Uma abordagem tradicional seria verificar o isomorfismo entre gráficos, no entanto, tal abordagem apresenta alguns problemas, como, por exemplo, o custo computacional é alto (não existe algoritmo polinomial) e não considera a flutuação intrínseca presente em processos que modelam eventos do mundo real. Portanto, uma solução é considerar que os grafos reais são gerados por por processos probabilísticos (modelo de grafo aleatório) e testar se os grafos foram gerados pelo mesmo modelo e conjunto de parâmetros. Neste trabalho revisaremos os modelos de grafos aleatórios e suas aplicações na modelagem de redes do mundo real, bem como a teoria espectral para grafos e diversos resultados, medidas derivadas do seu estudo e trabalhos recentes que utilizam métodos baseados em medidas espectrais para seleção do modelo em conjuntos de grafos. Finalmente, introduzimos abordagens paramétricas para testar a igualdade entre dois ou mais grafos aleatórios desde que sejam grandes o suficiente. Primeiro é proposto uma generalização de teste de razão de verossimilhança (*likelihood ratio test - LRT*), a seguir o teste t (*t student*) é generalizado para grafos e finalmente, a análise de variância (ANOVA) é desenvolvida para testar a igualdade de dois ou mais grafos. Os testes desenvolvidos utilizam medidas baseadas no espectro do grafo, como a entropia espectral e a divergência de Kullback-Leibler no estimador dos parâmetros do modelo. Para estimar a variância, é usado um bootstrap paramétrico. O poder do teste é ilustrado com uso de curvas ROC e a

utilidade do método é demonstrada através de sua aplicação na comparação de redes reais de interação proteína-proteína em seis espécies.

Palavras-chave: grafos aleatórios, testes estatísticos, teste paramétrico, divergência de Kullback-Leibler

Abstract

SOARES, G. E. **A semi parametrical statistics test to discriminate graphs** 2017. Phd Thesis - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Graphs are used to model networks in various scientific fields. The interest to model real world data as graphs (networks) is increasing over the years. Several examples can be found with applications in different areas such as molecular biology (gene regulatory networks, protein-protein interaction networks), neuroscience (functional brain networks), and social science (social networks like facebook, instagram, foursquare). Different of deterministic graphs, real world graphs present a random component, which makes their analysis difficult by traditional computer science algorithms. In addition, traditional algorithms and classical approach to graphs present difficulties and limitations that undermine such analysis. Classical approaches do not handle intrinsic randomness. Thus, formal statistical procedures for graphs become necessary. Therefore, random graphs are better suited to model real world networks. In this context, one simple question is: how to compare two (or more) real graphs? A traditional solution would be to verify the isomorphism. However, this approach presents at least two problems: (i) there is no polynomial algorithm to verify isomorphism between graphs; and (ii) it does not take into account the intrinsic fluctuation present in real world graphs. Therefore, one solution is to imagine that these natural graphs are generated by random processes (models) and test if the graphs are generated by the same model and set of parameters. We present a solution for this problem by proposing generalizations of a likelihood ratio test, a t-test, and an analysis of variance for graphs. We demonstrate their performance in Monte Carlo simulations and illustrate their application in protein-protein interaction networks.

Keywords: random graphs, statistical tests, parametric test, Kullback-Leibler divergence

Sumário

Capítulo 1

Introdução

Grafos vêm sendo estudados ao longo dos anos, ganhando destaque em 1735 quando Leonardo Euler resolveu o problema da ponte de Königsberg (Chung e Lu, 2006) e desde então é um tema com aplicabilidade em várias áreas do conhecimento,como matemática discreta, ciência da computação, biologia, química, pesquisa operacional e ciência social. Em 1959, os estudos de (Erdös e Rényi, 1959) e (Gilbert, 1959) trazem novo fôlego e uma nova maneira de analisar grafos, com os grafos aleatórios, que, através de um algoritmo, geram grafos que seguem uma lei de probabilidade, sendo capazes de representar várias redes do mundo real. Interações do mundo real, tais como conectividade funcional em regiões cerebrais (Barabási e Albert, 1999; Bullmore e Sporns, 2009), a interação de pessoas em redes sociais (Borgatti *et al.*, 2009; Scott, 2012) e associações moleculares em redes genéticas (Barabási e Oltvai, 2004) podem e têm sido representadas através do uso de grafos.

No início da era da informação (*Information Age*), uma nova família de grafos foi estudada, ganhando muita importância atualmente. Esta família contém vários grafos, como WWW, co-autores, redes sociais, biológicos, todos com algumas características em comum, como por exemplo, número de arestas ser usualmente linear com o número de vértices (*sparsity*), estrutura de mundo pequeno (pequena distância entre vértices e a presença de grupos de vértices densamente conectados), e a distribuição de graus que segue a lei de potência (*power lee degree distribution*), quando o número de vértices com grau d é proporcional a $d^{-\beta}$, para algum expoente $\beta > 0$ (Chung e Lu, 2006). Tais grafos recebem o nome de redes complexas, pois não são totalmente aleatórios nem regulares. Para estudar esse tipo de rede, vários modelos foram propostos, como Watts-Strotgatz (Watts e Strogatz, 1998) e Barabási-Albert (Barabási e Albert, 1999).

Uma vez que os grafos de interação são obtidos passamos à análise das informações que estes grafos representam. Um problema comum é verificar se grafos de condições diferentes são iguais ou não. Por exemplo, pode ser interessante comparar redes regulatórias de genes ou redes funcionais cerebrais de pessoas diagnosticadas com uma doença em comparação ao grupo de controle. Ou ainda se a interação social de uma comunidade é igual a outra. Uma estratégia tradicional é utilizar algoritmos padrão em ciência da computação baseadas em algoritmos para determinar isomorfismo entre grafos. Entretanto, tal estratégia encontra como principal obstáculo o fato deste ser um problema é NP-completo. Além disso, essa abordagem não considera a flutuação intrínseca presente em grafos do mundo real. Por exemplo, redes biológicas podem mudar com o tempo e também entre indivíduos na mesma condição. Neste caso, algoritmos que identificam isomorfismo vão discriminar grafos pertencentes ao mesmo grupo ou condição.

Outra abordagem é baseada na análise de características de grafos, especialmente medidas de centralidade como cemtralidade de grau ou centralidade de informação (*degree*

centrality), centralidade de proximidade (*closeness centrality*), centralidade de intermediação (*betwenness centrality*), dentre outras (Borgatti *et al.*, 2009; Penrose, 2003). No caso do estudo de medidas de centralidade, geralmente, as medidas de centralidade são estimadas e comparadas entre grafos usando procedimentos estatísticos padrão. A dificuldade do ponto estatístico é devido ao fato de não existir um modelo que gere um grafo com uma centralidade específica, ou seja, grafos gerados por diferentes modelos podem apresentara mesma medida de centralidade, por exemplo, dois grafos aleatórios gerados pelo modelo Watts-Strogatz com o parâmetro p de probabilidade de religação (*rewiring probability*) diferentes produzem grafos com o mesmo grau de centralidade, uma vez que o número de arestas não muda ou grafos gerados pelo mesmo modelo podem ter medidas de centralidade diferentes (ex: redes livre de escala). Com isso, temos que uma solução para comparar grafos do mundo real é assumir que eles são gerados através de processos probabilísticos (modelos de grafos aleatórios) e então testar se as populações de grafos são geradas pelo mesmo modelo ou não.

O primeiro trabalho em testar se duas populações de grafos são gerados pelo mesmo modelo e conjunto de parâmetros foi descrito por (Takahashi *et al.*, 2012). Baseado no fato que grafos gerados por modelos diferentes apresentam distribuições espectrais diferentes (conjunto de autovalores da matriz de adjacência do grafo), eles propuseram utilizar a divergência de Jensen-Shannon entre as distribuições espectrais combinando com um procedimento de bootstrap (no qual os grafos do conjunto são reamostrados com substituição). Posteriormente, Fujita *et al.* (2017) generalizou essa ideia para teste simultâneo entre dois ou mais grupos de grafos semelhante à análise de variância (ANOVA). No entando, esta abordagem apresenta limitações. A principal limitação é o fato que elas só podem ser aplicadas a conjuntos de grafos, e portanto, é necessário que o grafo se replique em cada população. Na aplicação desta abordagem foram realizados testes em redes cerebrais funcionais (*functional brain networks*) para verificar se grafos de indivíduos controles e diagnosticados com uma doença podem ser gerados pelo mesmo modelo de grafo. Neste exemplo, centenas de grafos estão disponíveis e foram utilizados com um procedimento de reamostragem de indivíduos independentemente.

A maioria das aplicações reais não possuem réplicas, como é o caso de redes sociais, interações de rede proteína-proteína, WWW e rede de co-autores. Então, como testar se dois grafos com apenas uma ocorrência de cada um são gerados pelo mesmo modelo? Se o modelo que gerou os grafos for conhecido, este problema é reduzido a comparar os parâmetros do modelo e a dificuldade passa a ser obter estes parâmetros para testá-los.

Neste trabalho, apresentamos generalizações de testes estatísticos tradicionais (teste baseado em razão deverossimilhança, teste t Student e análise de variância - ANOVA) para comparar se dois ou mais grafos foram gerados por um mesmo modelo.

Para ilustrar a utilidade do teste proposto, a aplicação prática do teste é demonstrada através de testes realizados em redes de interação proteína-proteína de seis espécies diferentes. As redes de interação proteína-proteína, em geral, são não direcionais, hierárquicas e apresentam características de redes livre de escala e de pequeno mundo. O estudo de redes de interação proteína-proteína pode ser usado para identificar agentes patológicos de doenças (Mani *et al.*, 2008), para identificar proteínas que são potenciais alvos para vacina (Andrés F Flórez e Muskus, 2010; LaCount DJ, 2005) e até mesmo para identificar fungos que prejudicam lavouras (Lan V Zhang e Roth, 2004).

A maioria das aplicações reais não possuem réplicas, como é o caso de redes sociais, interações de rede proteína-proteína, WWW e rede de co-autores. Então, como testar se dois grafos com apenas uma ocorrência de cada um são gerados pelo mesmo modelo? Se o modelo que gerou os grafos for conhecido, este problema é reduzido a comparar os parâmetros do modelo e a dificuldade passa a ser obter estes parâmetros para testá-los.

1.1 Objetivos

Este trabalho tem como principais objetivos:

1. Revisar testes estatísticos e estratégias de comparação entre grafos aleatórios.
2. Desenvolver testes estatísticos para identificar diferenças entre dois ou mais grafos gerados por um dos modelos analisados.
3. Desenvolver um pacote em R que possa ser utilizado em aplicações onde comparar dois ou mais grafos seja necessário.
4. Mostrar o teste em funcionamento através da comparação de grafos doo mundo real como os grafos de rede de interação proteína-proteína, comparação de subgrafos de redes sociais, dentre outros.

1.2 Organização

Este trabalho está organizado em oito capítulos capítulos,o capítulo atual apresenta a motivação, os objetivos e uma descrição sucinta do teste proposto. Os capítulos restantes estão organizados da seguinte maneira:

- No capítulo 2, apresentamos as definições de teoria de grafos como definição de grafos, alguns tipos de grafos, matriz de adjacência, isomorfismo, teoria espectral e alguns resultados, algumas medidas adotadas na análise de grafos como as medidas de centralidade espectrais e não espectrais e os modelos utilizados em nosso estudo.
- No capítulo 3 são apresentadas medidas espectrais usadas no estudo de grafos aleatórios, como densidade espectral, entropia espectral e as divergências de Kullback-Leiber e Jensen-Shannon para grafos.
- No capítulo 4 é apresentada uma abordagem baseada em métodos estatísticos que usam a entropia espectral na obtenção de parâmetros do modelo proposta por (Takahashi *et al.* , 2012) para seleção do modelo. Um método de seleção do modelo baseado na divergência de Kullback-Leiber. Posteriormente é apresentado um teste de hipótese para grafos baseado na divergência de Jensen-Shannon. Finalmente, (Fujita *et al.*, 2017) generaliza a abordagem propondo o ANOGVa para coleções de grafos.
- Como as abordagens do capítulo 4 funcionam apenas para conjuntos de grafos, no capítulo 5 apresentamos os testes paramétricos para grafos que podem ser aplicados a dois ou mais grafos, desde que estes grafos sejam grandes o suficiente. Começamos generalizando o teste de verossimilhança para grafos, na seguida o mesmo é feito para o teste t e finalmente, um teste baseado em análise de variância (ANOVA para grafos)é apresentado.
- No capítulo 6 as simulações realizadas para avaliação do poder do teste: primeiro foram realizadas simulações para avaliar o poder do teste baseado em razão de verossimilhança, com relação ao teste baseado no teste t stutent e comparando com o ANOVA para dois grafos; posteriormente, o teste baseado no ANOVA é generalizado para dois ou mais grafos e também são avaliados a influência do tipo de estimador no poder do teste.

- No capítulo 7 apresentamos a aplicação do teste em dados de redes de interação proteína-proteína e explicamos a importância desta aplicação nos estudos de redes de interação proteína-proteína. O teste é aplicado em dados de seis espécies diferentes.
- No capítulo 8 analisamos os resultados obtidos, comparando-os com trabalhos anteriores relacionados e fazemos um panorama dos trabalhos que podem ser realizados no futuro para melhorar os testes propostos.

Capítulo 2

Definições

Neste capítulo serão apresentados os conceitos essenciais para entendimento do texto. Inicialmente serão apresentados os conceitos básicos e introdutórios de teoria de grafos. Na sequência serão apresentadas conceitos de diferentes redes complexas e modelos de grafos aleatórios. Finalmente, definiremos algumas medidas de centralidade, a definição de espectro do grafo como a definição de grafos e medidas de interesse, e alguns resultados importantes da teoria espectral de grafos.

2.1 Grafos

Um grafo é um par ordenado de conjuntos $G = (V, E)$, onde V é um conjunto finito e não vazio arbitrário com elementos que são chamados de vértices e E um subconjunto do conjunto de todos os pares dos elementos de V (chamado de arestas) que conectam os vértices de V. O número de vértices e o número de arestas são denotados por $|V|$ e $|E|$, respectivamente. Em um grafo não dirigido, cada elemento $e \in E$ é um par não ordenado $e = v_1, v_2$ onde $v_1, v_2 \in V$ e $v_1 \neq v_2$. Os vértices que são conectados por uma arestas são chamados de vértices vizinhos ou adjacentes. A cada aresta e podemos associar um valor real não negativo (peso) w_e. Dado $v \in V$, denotamos por $N(v)$ o conjunto de vértices de G que são adjacentes a v. Dizemos que $N(v)$ é a vizinhança de v e que $d(v) = |N(v)|$ é o grau de v (número de arestas que incidem em v). A seguir serão apresentados alguns tipos de grafos especiais e definições que serão úteis no entendimento deste texto: A seguir serão apresentados alguns tipos de grafos especiais e definições que serão úteis no entendimento deste texto:

- Grafo regular de grau k ou k-regular é um tipo especial de grafo em que todos os vértices têm o mesmo grau k. A figura 2.1 mostra um grafo 4 regular.

- Grafo completo é o grafo no qual para todo par de vértice distintos v_i e v_j existe uma aresta os conectando. Ou seja, um grafo em que todos os vértices são adjacentes. O grafo completo com n vértices é $(n-1)$-regular.

- Cadeia (*walk*)é uma sequência finita de vértices adjacentes. Formalmente, uma sequência finita de vértices $v_1, v_2, ..., v_k$ é dita uma cadeia de v_1 a v_k quando $\{v_i, v_{i+1}\} \in E$ para $\leq i \leq k-1$. Para o caso em que $v_1 = v_k$ temos uma cadeia fechada, enquanto quando $v_1 \neq v_k$ é uma cadeia aberta.

- Caminho (*path*) é uma cadeia que possui todos os vértices distintos.

- Chamamos de ciclos os caminhos que são fechados ($v_1 = v_k$).

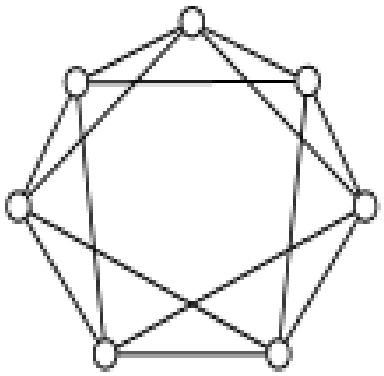


Figura 2.1: *Grafo 4-regular*

- O comprimento de um caminho ou ciclo é o número de arestas que ocorrem em cada um.
- Grafo conexo é um grafo no qual para cada par de vértices existe um caminho entre eles. Se existe pelo menos um par de vértices para o qual não existe um caminho, o grafo é desconexo.
- Grafo k-partido: seja $G = (V, E)$ um grafo com uma partição do conjunto de vértices V em k subconjuntos disjuntos não vazios dois a dois, ou seja, não existem vértices adjacentes em um mesmo conjunto da partição.

2.2 Grafos aleatórios

O termo grafos aleatórios é um abuso de linguagem, uma vez que os modelos de grafos aleatórios tem significado preciso e são bem definidos. Grafos aleatórios podem ser definidos como sendo uma variável aleatória definida em um espaço de probabilidade $(\Omega, \mathcal{F}, P)$com uma distribuição P, onde o espaço amostral Ω é um conjunto não vazio de grafos, o conjunto de eventos $\mathcal{F}$ é uma coleção de subconjuntos do espaço amostral e P é uma função que atribui uma probabilidade para cada etento. Um algoritmo capaz de gerar grafos que seguem uma determinada lei de probabilidade recebe é um modelo de grafo aleatório (Chung, 1994). Uma das maneiras de descrever um grafo aleatório é através de uma sequência de passos para construí-lo.

Os grafos aleatórios ganharam atenção a partir de 1959, com os trabalhos de (Erdös e Rényi , 1959) e (Gilbert, 1959). Desde então, vem sendo usados em aplicações para análise de redes de todos os tipos. Algumas questões acerca de grafos se tornam relevantes na análise destes grafos, como: dado um grafo gerado por um modelo de grafo aleatório é possível dizer que tal grafo possui certas propriedades (diâmetro pequeno, determinar qual o modelo o gerou ou ainda, se a estrutura de um modelo é predizível). A importância dos grafos aleatórios está em garantir a existência de grafos com certas propriedades.

Entender grafos aleatórios ajudam a entender estruturas de vários tipos de redes, como por exemplo redes sociais, redes pequenas de amigos. E com esse entendimento, é possível identificar comportamentos inadequados na rede, como contas duplicadas em redes sociais. Particularmente, queremos determinar quão similares são dois ou mais grafos aleatórios. Para tanto, podemos construir métodos estatísticos em grafos, uma abordagem não trivial

devido à complexidade para lidar com grafos, especialmente à medida em que os grafos ficam maiores e mais complexos.

Existem vários modelos de grafos aleatórios, uma vez que para obter um espaço de probabilidade basta atribuir a mesma probabilidade para todos os grafos em um conjunto finito, gerando grafos k-regulares, árvores aleatórias, etc. No decorrer deste capítulo apresentaremos alguns modelos de grafos aleatórios, com destaque especial aos quatro modelos que foram utilizados nas simulações: Erdos-Rényi Erdös e Rényi (1959), Geométrico, Barabási-Albert Barabási e Albert (1999) e Watts-Strotgatz Watts e Strogatz (1998).

2.3 Redes complexas

Com o surgimento da era da informação, redes como WWW, rede de coautores, redes sociais e biológicos despertaram o interesse em estudos apresentando similaridades inesperadas e formando, dessa forma, uma nova família de grafos. Com estrutura não trivial, nem totalmente aleatório ouregular, essa nova família recebe o nome de redes complexas. Dentre as características comuns presentes nessas redes, podemos citar:

- ser esparsas, i.e. presença de vértices densamente conectados e com pequenas distâncias entre eles;
- e distribuíção de graus que segue a lei de potência (*power law*), ou seja, o número de vértices com grau d é proporcional a $d^{-\beta}$ para algum expoente $\beta \geq 0$, o que significa que quanto mais conectado um vértice, maior a probabilidade de receber uma nova aresta.

As redes complexas são utilizadas para modelar redes do mundo real envolvendo em seu estudo diversas áreas como a ciência da computação, matemática, física, biologia e sociologia e vários modelos têm sido propostos para estudar este tipo de grafo. São exemplos de redes complexas: redes aleatórias, redes de mundo pequeno e redes livres de escala. Neste trabalho, veremos particularmente dois modelos com tais características: Barabási-Albert (Barabási e Albert, 1999) e Watss-Strogatz (Watts e Strogatz, 1998).

2.4 Mundo Pequeno

O efeito mundo pequeno é caracterizado por padrões altamente conectado, formando pequenas quantidades de conexões em cada vértice. Foi identificado por (Watts e Strogatz, 1998), que propuseram um modelo semelhante ao de (Erdös e Rényi, 1959), no qual grande parte das conexões são estabelecidas entre os vizinhos mais próximos, de maneira que a distância média entre dois vértices quaisquer não ultrapasse um número pequeno de vértices.

Nas redes de mundo pequeno, a maioria dos vértices se conecta a outros através de um caminho mínimo, também chamado de caminho geodésico ou distância geodésica, que é formado pelo menor número de arestas que conectam um vértice de origem a um vértice destino. Em 1960, (Milgram, 1963) realizou um experimento no qual ao enviar uma carta, se esta carta fosse entregue a um indivíduo que não fosse o destinatário, e ele repassasse a um outro indivíduo e assim sucessivamente, em aproximadamente seis envios a carta chegaria ao destinatário. Esse experimento ilustra o efeito mundo pequeno.

2.5 Redes livres de escala

As redes livres de escala são redes nas quais as distribuições de grau seguem leis de potência ($p(k) = k^{-\gamma}$), i.e. a probabilidade de um nó ser escolhido é diretamente proporcional ao seu grau. Dessa forma, os vértices dessas redes tem poucas ligações, com alguns vértices com elevado número de ligações. Os nós que concentram a maioria das ligações são chamados *hubs*.

O modelo de Barabási-Albert (Barabási e Albert, 1999) foi proposto para modelar redes livres de escala e, portanto, apresenta duas propriedades de redes livres de escala: crescimento e anexação preferencial. O crescimento significa que novos nós são adicionados à rede com o passar do tempo. A anexação preferencial é consequência da distribuição de grau, e significa que quanto mais conectado um vértice for, maior será a sua probabilidade de receber novas ligações. Este tipo de rede é adequado para modelar redes sociais, redes como o Google, redes de coautores, etc **?**.

2.6 Modelos

Nesta seção serão apresentados cinco modelos de grafos: o modelo de Erdös-Rényi (Erdös e Rényi , 1959), o modelo Geométrico (Penrose, 2003), o modelo Regular, o modelo de Wattz-Strogatz (Watts e Strogatz, 1998) e o modelo de Barabási-Albert (Barabási e Albert, 1999). Na figura 2.2 é possível ver a estrutura de grafos de 500 vértices gerados pelos modelos descritos nesta seção.

2.6.1 Erdös-Rényi

O modelo Erdös-Rényi (ER) foi proposto em 1959 (Erdös e Rényi, 1959) e é um dos modelos de grafo mais estudados. É gerado acrescentando-se uma aresta a um par de vértices (v_i, v_j) com probabilidade p em um grafo com n vértices. A construção do grafo ER pode ser formalizada como se segue:

Definição: seja n um inteiro positivo e p ($0 \leq p \leq 1$) a probabilidade. O grafo $G(n, p)$ é o grafo não direcionado com n vértices com arestas determinadas como se segue: para todos os pares de vértices v_i, v_j existe uma aresta (v_i, v_j) com probabilidade p.

Esse processo pode gerar qualquer grafo, mas é pouco provável que não exista nenhuma aresta com p grande o suficiente. O grafo $G(n, p)$ pode ser visto como uma distribuíção de probabilidade sobre o conjunto de todos os grafos de n vértices.

2.6.2 Geométrico

Um grafo aleatório geométrico $G(n, r(n))$ é um grafo que é construído considerando um componente espacial, ou seja, para construir um grafo aleatório geométrico (GRG) n pontos são dispostos aleatoriamente (seguindo uma distribuíção uniforme) em um disco (ou quadrado) unitário e então, arestas são adicionadas conectando dois pontos se, e somente se, a distância euclidiana entre eles for no máximo $r(n)$. Este modelo de grafo tem sido utilizado em testes de hipótese, física estatística e redes de sensores (Penrose, 2003).

2.6.3 Regular

Um modelo especial de grafo aleatório é o grafo regular, no qual cada vértice tem o mesmo grau. Se o grau for igual a r o grafo é r-regular. Tal modelo pode ser definido como

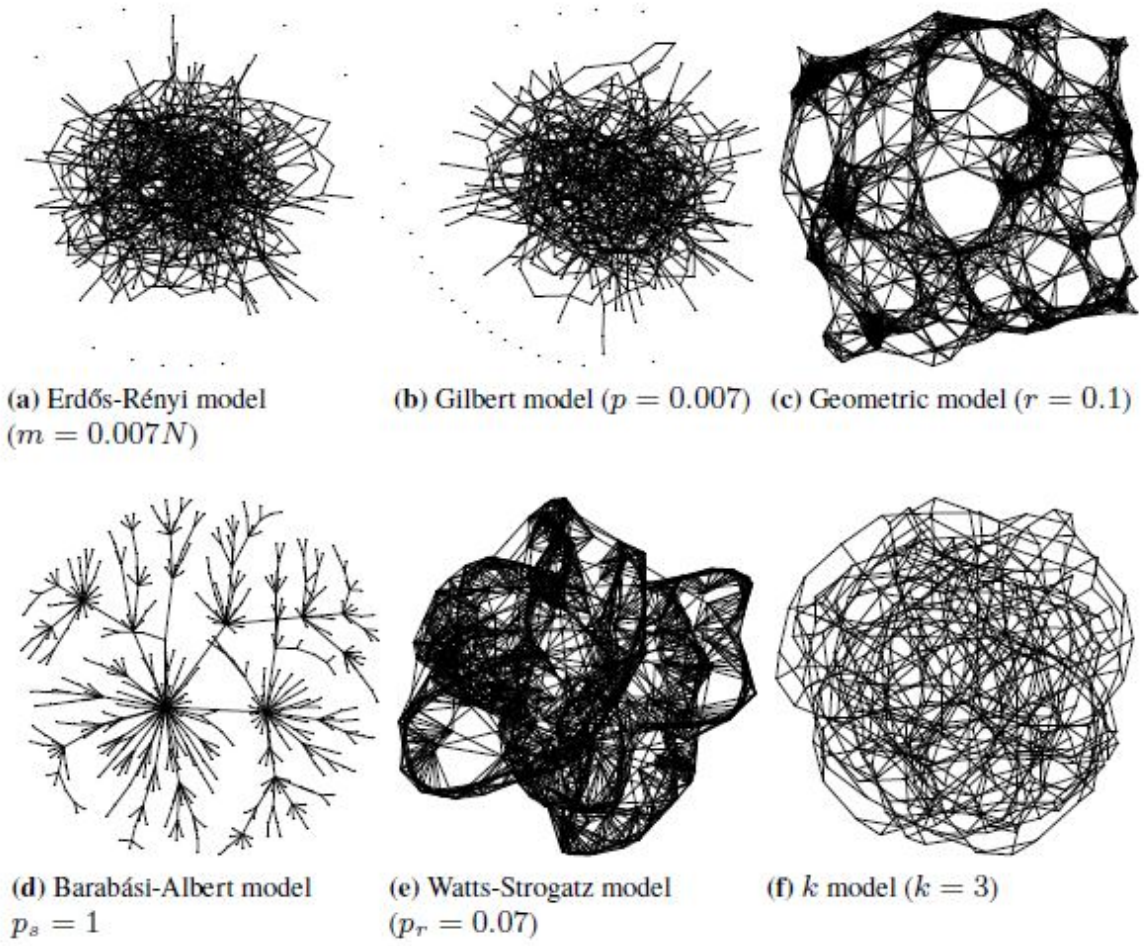


Figura 2.2: *Modelos de grafos aleatórios com 500 vértices gerados pelos modelos Erdõs-Rényi (a), Gilbert (b), Barabási-Albert (d), Watts-Strogatz (e) e k-regular(f). Em (a) o número de arestas é igual a* $0.007N = o.0007(5002)$. *Em (b), a probabilidade p de conectar dois vértices é igual a* 0.0007. *Em (c) o raio* $r = 0.1$. *Em (d) o expoente de escala* $p_S = 1$. *Em (e) a probabilidade de reconectar um vértice* p_r *é 0.07. Em (f) o grau de cada vértice é* $k = 3$. *Fonte: (de Siqueira Santos et al., 2016)*

se segue: Suponha n e $r = r(n)$ naturais tais que $3 \leq r < n$ e $rn = 2m$ é par, com essa definição, pelo menos um entre r e n precisa ser par. Seja o conjunto dos grafos $r-regulares$ denotados por G_{r-reg}. Como na suposição inicial $r \geq 3$, o conjunto G_{r-reg} não é vazio. Desta forma, G_{r-reg} é transformado em um espaço de probabilidade, no qual todos elementos tem a mesma probabilidade.

2.6.4 Wattz-Strogatz

É um modelo de grafo aleatório proposto por (Watts e Strogatz, 1998) que está entre um grafo regular e um ER. É conhecido por apresentar propriedades de mundo pequeno, e, portanto, apresenta as seguintes propriedades:

- os caminhos apresentam baixa média de tamanho, uma vez que a maioria dos vértices não são vizinhos uns dos outros, mas todo vértice pode ser alcançado a partir de qualquer vértice por um pequeno número de passos;
- coeficiente de clusterização (número de triângulos no grafo) mais alto que no modelo ER.

O algoritmo para construir um grafo Wattz-Strogatz (WS) é:

Entrada: n - número de vértices, nei - número de vizinhos (grau médio) e p^w (rewiring probability)

1. Construa uma estrutura de anel com n vértices e conecte cada vértice aos seus primeiros nei vizinhos ($\frac{nei}{2}$ de cada lado).
2. Escolha um vértice v_i e a aresta e que o conecta ao seu vizinho mais próximo no sentido horário.
3. Com probabilidade p^w substitua a aresta e a um vértice escolhido aleatoriamente entre todos os vértices do anel.
4. Os passos 2 e 3 são repetidos no sentido horário em volta do anel, para cada vértice.
5. A seguir, repita os passos 2 a 4, alterando o passo 2 de maneira que a sejam escolhidas as arestas que conectam vértices aos seus segundos vizinhos mais próximos, depois aos terceiros, quartos e assim sucessivamente ao redor do anel em direção ao exterior para os vizinhos mais distantes até que toda aresta tenha sido considerada pelo menos uma vez.

Saída: WS grafo aleatório

2.6.5 Barabási-Albert

Barabási-Albert (BA) é um modelo proposto por (Barabási e Albert, 1999) para gerar redes sem escala (i. e. redes complexas nas quais o grau de distribuíção segue a lei de potência) de maneira aleatória. Para gerar tais redes é usado o processo de anexação preferencial no qual as variações aleatórias são reforçadas, ou seja, se um nó possui mais ligações que outro ele terá maior probabilidade de ser anexado, gerando uma rede com poucos nós com grau elevado e muitos nós com grau baixo.

O algoritmo para gerar o grafo consiste em: dado um número de vértices inicial pequeno (n_{V0}), a cada passo (tempo $t \in N$) é adicionado um novo vértice com $m_1 (\leq n_{V0})$ arestas que conectam o novo vértice a m_1 vértices diferentes já presentes na rede. A probabilidade de um novo vértice ser conectado ao vértice i é proporcional ao grau de i (número de arestas de i) com ordem de proporcionalidade dada pelo expoente de escala p^s ($P(v_i)$ $grau(v_i)^{p^s}$, onde $grau(v_i$ é o número de arestas adjacentes do vértice v_i naquele instante). O processo de escolha preferencial das arestas é como se segue:

- Com probabilidade $p \in [0, 1]$, esse novo vértice se conecta aos vértices existentes com probabilidade igual para todos os vértices.
- Com probabilidade $1 - p$, esse novo vértice se conecta com os m nós existentes com probabilidade proporcional ao grau do vértice com o qual ele será conectado.

Nos grafos gerados por este modelo, a distribuição dos graus é, assintoticamente, regida pela Lei de Potência (Bollobás *et al.*, 2001), propriedade amplamente observada em sistemas naturais e artificiais.

2.7 Matriz de adjacência

A matriz de adjacência é formada a partir das relações de adjacências entre os vértices do grafo e pode ser utilizada para obter propriedades estruturais de grafos. Pode ser definida como se segue:

Seja n_v a cardinalidade do conjunto de vértices V, a matriz de adjacência de G é uma matriz A com n_v linhas e n_v colunas, onde

$$A_{ij} = w_{ij}, \text{ se existe os vértices } v_i \text{ e } v_j \text{ são vizinhos.}$$
$$A_{ij} = 0, \text{ se não existe os vértices } v_i \text{ e } v_j \text{ não são vizinhos.}$$

Caso o grafo não tenha peso (w) nas arestas, como é o caso dos grafos analisados neste trabalho, todas as arestas têm peso igual a 1. Além disso, em um grafo não dirigido, a matriz de adjacência A é simétrica, isto é, $A_{ij} = A_{ji}$. Portanto, a matriz A de G será uma matriz real formada por zeros e uns com todos autovalores reais.

Seja A a matriz de adjacência de um grafo G, o polinômio característico de G ($p_G(\lambda)$) é o polinõmio característico da matriz A, ou seja, $\det(\lambda I - A)$. Se λ é uma raíz de $p_G(\lambda)$ é um autovalor do grafo G. Se λ é um autovalor de G, então um vetor não nulo $x \in \mathcal{R}^{\setminus}$ que satisfaça $Ax = \lambda x$ é chamado de autovetor de G.

2.8 Isomorfismo de grafos

Dois grafos g_1 e g_2 são isomorfos quando for possível obter um a partir do outro usando uma permutação das rotulações de seus vértices. Formalmente, isso quer dizer que existe uma correspondência biunívica entre seus conjuntos de vértices, de modo que as adjacências sejam preservadas. Se g_1 e g_2 são isomorfos as matrizes de adjacência $A(g_1)$ e $A(g_2)$ são semelhantes, ou seja, existe uma matriz de permutação P tal que $P^T A(g_1)P = A(g_2)$.

2.9 Espectros de grafos

Seja $G = (V, E)$ um grafo não direcionado com n vértices (i.e., $n = |V|$). O espectro de G é o conjunto dos autovalores da matriz de adjacência A. Como G é não direcionado, se dois vértices i e j são conectados por uma aresta, então $A_{G_{ij}} = A_{G_{ji}}$, ou seja, A é simétrica e, portanto, todos seus autovalores são reais.

Sejam $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ o espectro de G de tal forma que $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. O maior autovalor de G é denominado índice de G.

Dois grafos isomorfos têm o mesmo espectro, no entanto, dois grafos coespectrais (grafos que possuem o mesmo espectro) não são, obrigatoriamente, isomorfos.

A teoria espectral de grafos estuda propriedades dos espectros do grafo e suas associações com a estrutura do grafo. Os resultados podem facilitar a obtenção do espectro de um grafo usando operações de conjuntos, como complementar e união. Já outros resultados podem fornecer informações sobre alguns autovalores e medidas derivadas destes. Alguns exemplos são listados abaixo:

1. Alguns resultados descritos por Cvetković *et al.* (1980):

- Seja $d(i)$ o grau do vértice i. O autovalor λ_i é pelo menos

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

e no máximo $\max_{i \in V} d(i)$ (Cvetković *et al.*, 1980)

- O grafo G é bipartido (ou bigrafo) se e somente se $\lambda_n = -\lambda_1$
- O maior autovalor λ_1 de G e o grau máximo de um vértice Δ estão relacionados da seguinte maneira:

$$\sqrt{\Delta} \leq \lambda_1 \leq \Delta$$

2. Se G é conexo (existe um caminho que leva a todo vértice), então o autovalor λ_1 é maior que λ_2 e existe um autovetor positivo associado a λ_1 (Bapat e Raghavan, 1997).

3. Cada vértice em V é conectado exatamente a λ_1 vértices se, e somente se, o vetor de 1s é um autovetor de λ_1 (Spielman, 2012).

4. Seja $C \subseteq V$ tal que cada par de vértice em C está conectado em G (C é um clique de G). Então, o tamanho de C é menor ou igual a $\lambda_1 + 1$ (Lovász, 2007).

5. Seja k o diâmetro de G. Se G é conexo, então a matriz de adjacência A de G tem pelo menos $k + 1$ autovalores distintos (Brouwer e Haemers, 2012).

2.10 Medidas de centralidade

Ao usar grafos para representar redes é possível analisar a importância de um vértice. Por exemplo, as redes sociais podem ser modeladas de maneira que cada vértice representa uma pessoa e as arestas representam a relação entre duas pessoas. Como saber quais são os vértices mais importantes? Intuitivamente, os vértices mais importantes são aqueles que estão mais frequentemente relacionados a outros, os mais centrais, ou seja, aqueles a partir dos quais podemos atingir qualquer outro com facilidade e rapidez pois essas relações os tornam mais visíveis, sendo os que possuem a maior parte do acesso ou controle (Wasserman e Faust , 1994).

As medidas de centralidade surgiram com a intenção de quantificar esta importância, descrevendo as propriedades da localização de um vértice na rede. O conceito de centralidade em redes sociais foi introduzido por (**?**) e muitas outras medidas foram apresentadas ao longo dos anos. Tais medidas são baseadas na interação do vértice com o restante da rede, levando em consideração as diferentes maneiras de interação entre os elementos da rede.

Nesta seção serão apresentadas medidas de centralidade não espectrais que são obtidas através da posição estrutural de um vértice em um grafo e medidas de centralidade espectrais, que usam as propriedades dos autovalores e autovetores das matrizes associadas ao grafo para obter as propriedades estruturais..

São exemplos de medidas de centralidade não espectrais como a centralidade de grau (que conta o número de arestas incidentes a um vértice do grafo), de proximidade (relacionada com a distância total de um vértice aos demais vértices do grafo), a centralidade de eficiência (que minimiza as distâncias de um vértice) e de intermediação (mede a quantidade de geodésicas entre todos os pares de vértice do grafo passam através de determinado vértice). Já dentre as medidas de centralidade espectrais, podemos destacar a centralidade de autovetor (mede a relevância de um vértice em função da sua relação com os vizinhos, de

maneira que se um vértice for vizinho de um vértice importante será importânte também) e a centralidade via conectividade algébrica (mede o grau de relevância de um vértice em relação a vulnerabilidade que ele oferece à rede caso seja retirado).

2.10.1 Centralidade de grau

O grau de um vértice, denido como o número de arestas conectadas ao vértice, é uma das medidas de centralidade mais utilizadas. Se as arestas tiverem peso, nós obtemos o grau de um vértice pela soma dos pesos das arestas incidentes ao vértice. Assim, podemos escrever o grau do vértice vi de um grafo G com nV vértices como

A concepção mais simples e intuitiva no que diz respeito à centralidade de um vértice é o número de contatos diretos que ele possui. Uma pessoa que se encontra em uma posição que permite o contato direto com muitos outros é vista pelos demais como um canal maior de informações, razão pela qual dizemos ser mais central. SHAW 58[ ], em 1964, usou o grau do vértice como medida de centralidade e, em seguida, ACKENZIE 55[ ] e BEAUCHAMP 57[ ] aplicaram variantes do grau em problemas de redes sociais. Segundo FREEMAN 21[ ], Nieminem, em 1974, foi quem chamou a medida usada por Shaw e os demais pesquisadores de centralidade de grau, e ainda afirma ser esta a medida mais adequada (pelo menos até aquela data, 1979). Assim chamada, a centralidade de grau nada mais é que a contagem do número de adjacências de um vértice k v , ou seja, este parâmetro coincide com o próprio grau de k v . Formalmente, temos:

2.10.2 Centralidade de intermediação *betweenness centrality*

A centralidade de intermediação *betweenness centrality* de um vértice v (importância relativa de v na rede) é definida como o número de caminhos mais curtos entre todos os pares de vértices que passam por v (Freeman, 1978). A centralidade de *betweenness* média é soma das centralidades de *betweenness* dividida pelo número de vértices.

Se o grafo tiver peso nas arestas, a medida pode ser calculada usando o algoritmo de Dijkstra para achar os caminhos mais curtos (**?**). Neste estudo, os pesos denotarão a força de interação entre dois genes e, portanto, o algoritmo de Dijkstra deve utilizar valores nas arestas inversamente proporcionais ao peso (assim, as arestas com associação mais forte estarão mais "próximas"). **?** propôs utilizar o algoritmo de Dijkstra associando a cada aresta e o valor $\frac{1}{w_e}$, onde w_e é o peso da aresta e.

2.10.3 Coeficiente de *clustering* médio

O coeficiente de *clustering* local de um vértice v é o número de arestas entre seus vizinhos dividido pelo número de arestas que poderiam existir na vizinhança de v (Watts e Strogatz, 1998).

O coeficiente de *clustering* médio é a soma dos coeficientes de *clustering* locais dividida pelo número de vértices.

Se o grafo tiver peso nas arestas, consideraremos o método proposto por **?**, que define o coeficiente de *clustering* do vértice v_i como:

$$c_i = \frac{1}{s_i(k_i - 1)} \sum_{\{v_j, v_h\} \in E, v_j, v_h \in N(v_i)} w_{ij} + w_{ih}$$

onde s_i é a soma dos pesos das arestas com ponta em v_i, k_i é o número de arestas com ponta

em v_i (grau de v_i), $N(v_i)$ é a vizinhança de v_i (conjunto dos vértices adjacentes a v_i), E é o conjunto das arestas do grafo e w_{ab} denota o peso da aresta $\{a, b\}$.

2.10.4 Comprimento médio do caminho mais curto

O comprimento do caminho mais curto entre dois vértices v_i e v_j é o número de arestas do caminho mais curto entre eles. O comprimento médio do caminho mais curto é a média dos comprimentos de caminhos mais curtos entre todos os pares de vértice v_i e v_j, com $i \neq j$.

No caso em que as arestas têm pesos, os caminhos mais curtos podem ser obtidos pelo algoritmo de Dijkstra, associando a cada aresta e o valor $\frac{1}{w_e}$, onde w_e é o peso de e.

2.10.5 Centralidade de autovetor

A centralidade de autovetor de um vértice considera, além do grau do próprio vértice, o grau dos vértices vizinhos, sendo portanto, uma medida mais conveniente quando se deseja observar a propagação de um fenômeno em uma rede, como em estudos de propagação de doenças (**??**) ou de efeitos de trânsito (**?**).

Podemos definir a centralidade de autovetor do vértice v_i do grafo como sendo a i-ésima coordenada x_i do autovetor unitário positivo $x = [x_1 ... x_n]^T$ associado ao índice λ_1 do grafo, ou seja, é o número:

$$x_i = \frac{1}{\lambda_1} \sum_{j=1}^{n} a_{ij} x_j,$$

onde os a_{ij} são os elementos de sua matriz de adjacência.

Capítulo 3

Entropia de grafos

No capítulo anterior foram apresentados modelos de grafos aleatórios e medidas que podem ser usadas para analisar grafos. Uma vez que temos tais definições, outras questões a cerca de grafos aleatórios surgem, como dado um grafo aleatório g e um modelo de grafo aleatório $\mathcal{M}$ é possível medir o tanto que tal modelo descreve o grafo g? Dentre um conjunto de modelos de grafos aleatórios, qual descreve melhor g? Essas e outras questões motivam o estudo de grafos aleatórios. Para responder essas perguntas é importante construir métodos estatísticos para grafos, embora não seja uma abordagem trivial dada à complexidade de arestas e vértices.

Visando responder perguntas como se é possível descobrir qual foi o modelo que gerou um grafo observado, (Takahashi *et al.*, 2012) apresentaram uma abordagem de seleção de modelo, usando o espectro do grafo para definir a entropia espectral de um grafo. Neste capítulo, apresentaremos esta estratégia.

3.1 Densidade espectral

Seja G um grafo gerado por um modelo no espaço amostral g. A densidade espectral empírica de G é Rogers (2010):

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

onde δ é o delta de Dirac é uma medida de probabilidade que satisfaz:

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

A densidade espectral empírica esperada é obtida quando calculamos o limite $n \to \infty$ da esperança (denotado por "$\langle . \rangle$") com respeito a lei de probabilidade de g da densidade espectral empírica:

$$\rho(\lambda) = \lim_{n\to\infty}\langle\frac{1}{n}\sum_{j=1}^{n}\delta(\lambda - \lambda_j)/\sqrt{n}\rangle$$

A figura 3.1 mostra a diferença da distribuição espectral em diferentes modelos de grafos aleatórios.

3.2 Entropia espectral do grafo

Na seção anterior e no capítulo 2 é possível observar a relação entre o espectro de um grafo e sua estrutura. Sabendo desta relação, e tendo definido a densidade espectral, é possível apresentar outras medidas que medem e comparam estruturas de grafos aleatórios. A medida de entropia, proposta por (Shannon, 1948), quantifica a quantidade de incerteza associada ao seu valor.

Sejam duas variáveis discretas X e Y, que podem assumir valores 0 e 1. Suponha que as probabilidades para que X e Y assumam os valores possíveis são: $P(X = 0) = P(X = 1) = 0,50$, $P(Y = 0) = 0,95$ e $P(Y = 1) = 0,05$. Intuitivamente, a variável X está associada a uma quantidade maior de incerteza que a variável Y, uma vez que prever o valor de X é difícil, mas o valor de Y será "quase certamente" 0. Suponha ainda uma terceira variável Z com probabilidades $P(Z = 0) = 0,75$ e $P(Z = 1) = 0,25$, a incerteza esperada de Z está entre a incerteza de X e Y. A entropia é uma medida que representa bem essa noção intuitiva de cada variável aleatória discreta X que pode assumir n valores com probabilidade $p_1, p_2, \therefore, p_n$ e é definida como:

$$H(X) = \sum_{i=1}^{n} p_i log p_i$$

onde $p_i log p_i = 0$ quando $p_i = 0$ e o logaritmo tem uma base arbitrária, mas fixa Shannon (1948).

Observe que $H(X)$ vale zero se, e somente se, um dos números de $p_1, p_2, \therefore, p_n$ é um e todos os demais valem zero. Nesse caso, os valores de X podem ser preditos com absoluta certeza. Para todos os demais casos, a entropia será positiva. Intuitivamente, o valor de X terá maior incerteza se $p_1 = p_2 = \therefore = p_n = 1/n$. De fato, podemos verificar que a entropia será máxima nesse caso, isto é, quando $H(X)$ vale

$$-\sum_{i=1}^{n} \frac{1}{n} log \frac{1}{n}$$

Se f é uma função convexa contínua, então ela satisfaz a propriedade

$$f(\frac{1}{n} \sum x_i) \leq \frac{1}{n} \sum_{i=1}^{n} f(x_i$$

onde $x_1, x_2, \therefore, x_n$ são números reais positivos Khinchin (1957). Como $f = xlogx$ é convexa contínua nos reais positivos, segue que

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

$$= n(\frac{1}{n} \sum_{i=1}^{n} p_i) log(\frac{1}{n} \sum_{i=1}^{n} p_i)$$

$$\leq n(\frac{1}{n} \sum_{i=1}^{n} p_i log p_i)$$

$$= \sum_{i=1}^{n} p_i log p_i = -H(X)$$

Temos que

$$H(X) \leq -\sum_{i=1}^{n} \frac{1}{n} log \frac{1}{n}$$

isto é, a entropia é máxima quando $p_1, p_2, \therefore, p_n = 1/n$.

A entropia de variáveis contínuas, também conhecida como entropia diferencial, é definida similarmente. O somatório é substituído por uma integral no suporte da função de densidade de probabilidade. Contudo, diferentemente da entropia definida para distribuições discretas, a entropia diferencial pode assumir valores negativos.

Para um grafo, a entropia quantifica a aleatoriedade da sua estrutura. Formalmente, a entropia espectral de um grafo é definida como segue. Sejam g um grafo aleatório e ρ sua densidade espectral. A entropia espectral de g é

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

onde $0log0 = 0$ Takahashi *et al.* (2012).

Com isso, é possível aproximar a entropia para grafos gerados pelo modelo Erdös-Rényi e Gilbert por

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

onde p é a probabilidade de um par de vértices ser conectado por uma aresta Takahashi *et al.* (2012). Neste caso, o valor máximo da entropia é atingido quando $p = 0,5$, o que é consistente com a ideia intuitiva que é mais difícil prever se dois vértices do grafo serão conectados quando $p = 0,50$, e, portanto, a quantidade de incerteza é alta. De maneira análoga, quando $p = 0$ e $p = 1$, a quantidade de incerteza associada ao modelo será menor, pois os grafos serão sempre vazios (quando $p = 0$) e completos (quando $p = 1$).

Para aproximar a entropia de grafos aleatórios, primeiro o modelo é utilizado para construir os grafos e então, estima-se a densidade espectral do conjunto. Por exemplo, dado um modelo de grafo aleatório, sejam $\{G_1, G_2, ..., G_p\}$ os grafos obtidos através do modelo com n vértices. Para cada G_i, $i \leq 1 \leq p$ o estimador da densidade espectral é aplicado. O estimador aplicado nos exemplos descritos seguem um estimador baseado no Kernel Gaussiana (de Siqueira Santos *et al.*, 2016; Takahashi *et al.*, 2012), que pode ser interpretada como sendo uma versão suavizada do histograma. Dado um grafo G_j e sem espectro $\{\lambda_1^(j), \lambda_2^(j), ..., \lambda_n^(j)\}$, cada autovalor λ_i contribui na função estimada num ponto λ considerando a sua diferença com λ_i. Tal contribuíção recebe um peso calculado pela função do Kernel, que depende de um intervalo h, que controla o tamanho da vizinhança ao redor de λ. O estimador da função de densidade pode ser descrito por:

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

onde

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

Em seu trabalho, (Takahashi *et al.*, 2012) utilizou o critério de Silverman para selecionar o intervalo h. Para obter um estimador para o grafo aleatório, é calculada a estimativa da densidade para cada grafo $G_1, G_2, ..., G_p$ e então obter a média de todos os estimadores. Na figura 3.2 estão graficos que representam a entropia espectral empírica para os modelos Erdös-Rényi, Gilbert, geométrico, Barabási-Albert, Watts-Strogatz e k-regular. Para cada

modelo foram gerados 50 grafos com 500 vértices.

Analisando a entropia espectral, podemos perceber que a relação da incerteza dos modelos com os seus parâmetros, para alguns modelos, são apontados os comportamentos mais importantes que relacionam seus parâmetros com a entropia:

- Para o modelo Erdös-Rényi, com número de arestas m, a entropia atingirá seu valor máximo quando $\binom{m=n}{2}$, onde n é o número de vértices. Quando o grafo aproximar de um grafo completo ($\binom{m \to n}{2}$) ou vazio ($m \to 0$) a entropia atingirá os menores valores.

- Os grafos do modelo k-regular também se aproximam do menor valor de entropia quando o grafo se aproxima de um grafo vazio ($k \to 0$). Quando k aumenta a entropia aumenta até k atingir um valor intermediário.

- No modelo geométrico, quando o raio r é próximo a zero, o grafo provavelmente terá poucas arestas, enquanto quanto mais próximo r for de $\sqrt{2}$ mais provavelmente o grafo se aproximará de um grafo completo. Nestes dois casos, a incerteza é baixa, no entanto, com r em um valor intermediário entre 0 e $\sqrt{2}$, a entropia terá os valores mais altos.

- A entropia do grafo Barabási-Albert é inversamente proporcional ao expoente de escala (p_s). Quando p_s é baixo, a construção do grafo se torna mais aleatória, pois a influência do grau dos vértices sobre a probabilidade de conectar um vértice a outro é pequena. Já, quando p_s é alto, o grau dos vértices tem um peso maior na escolha dos pares de vértices a serem conectados e, assim, a quantidade de incerteza será pequena.

- Já no modelo Watts-Strogatz, a entropia espectral aumenta com o crescimento do parâmetro p_r, que é a probabilidade de substituir a aresta recém-inserida no grafo, que conecta um vértice v_i a outro vértice que está próximo a ele na estrutura de anel, por uma aresta que conecta v_i a um vértice escolhido aleatoriamente. Assim, quando $p_r = 1$, temos um grafo construído de forma aleatória, como o grafo de Erdös-Rényi, e quando $p_r = 0$, temos um grafo determinado pela estrutura de anel, onde cada vértice está conectado aos K vértices mais próximos.

Uma vez definida a entropia espectral, a entropia cruzada entre duas densidades espectrais f_1 e f_2 (quantidade de incerteza quando f_2 é usada para estimar f_1) é definida como

$$H(f_1, f_2) = -\int_{-\infty}^{+\infty} f_1(\lambda) log f_2(\lambda) d\lambda$$

onde $0log0 = 0$.

3.3 Divergência de Kullback-Leiber

Enquanto a entropia quantifica a incerteza associada a uma variável aleatória, a divergência de Kullback-Leiber (KL) mede a quantidade de informação perdida quando uma distribuição de probabilidade é utilizada para aproximar outra. Para grafos, a divergência de KL pode ser utilizada para discriminar distribuições de probabilidade e para selecionar o modelo de grafo que melhor descreve o grafo observado. Pela sua definição, quando duas densidades espectrais forem diferentes, os grafos aleatórios correspondentes são diferentes. Entretanto, o contrário não é necessariamente verdade, uma vez que grafos aleatórios diferentes podem ter a mesma densidade espectral.

Sejam ρ_1 e ρ_2 densidades espectrais dos grafos g_1 e g_2, respectivamente. A divergência de Kullback-Leibler entre os grafos g_1 e g_2 (Takahashi *et al.*, 2012) quando o suporte de ρ_2 contém o suporte de ρ_1 é:

$$KL(\rho_1|\rho_2) = \int_{-\infty}^{+\infty} \rho_1(\lambda) \log \frac{\rho_1(\lambda)}{\rho_2(\lambda)} d\lambda \tag{3.1}$$

onde $0 \log 0 = 0$ e ρ_2 é chamada de medida de referência. Se o suporte de ρ_2 não contém o suporte de ρ_1, $KL(\rho_1|\rho_2) = +\infty$.

Esta medida é não negativa e assume o valor zero se, e somente se, ρ_1 e ρ_2 são iguais. Em muitos casos, quando $\rho_1 \neq \rho_2$, $KL(\rho_1, \rho_2) \neq KL(\rho_2, \rho_1)$, isto é, a divergência de Kullback-Leibler é assimétrica. Esta propriedade pode ser útil quando é necessário encontrar uma medida que melhor descreva o espectro observado, entretanto, se o objetivo for diferenciar grafos baseados em suas propriedades estruturais o mais adequado é usar uma divergência simétrica, como a divergência de Jensen-Shannon, que será descrita na próxima seção.

3.4 Divergência de Jensen-Shannon entre grafos

Em algumas aplicações pode ser necessário utilizar uma divergência que seja simétrica. A divergência de Jensen-Shannon (JS) é uma alternativa simétrica á divergência de KL.

Dados dois grafos aleatórios g_1 e g_2, definição de distância entre dois grafos baseado na entropia. Objetivo: identificar grafos que são gerados pelo menos processo aleatório, e não isomorfismo entre grafos (um isomorfismo entre os grafos g_1 e g_2 é uma bijeção f de um conjunto de vértices de g_1 para um conjunto de vértices de g_2 tal que quaisquer dois vértices u e v de g_1 são adjacentes se, e somente se $f(u)$ e $f(v)$ são adjacentes em g_2).

A divergência de Jensen-Shannon entre os grafos G_1 e G_2 com densidades espectrais ρ_1 e ρ_2 é definida como (Takahashi *et al.*, 2012):

$$JS(\rho_1, \rho_2) = \frac{1}{2} KL(\rho_1|\rho_M) + \frac{1}{2} KL(\rho_2|\rho_M) \tag{3.2}$$

onde $\rho_M = \frac{1}{2}(\rho_1 + \rho_2)$

Esta medida pode ser interpretada como uma medida de diferenças estruturais entre dois grafos. Além disso, JS é uma medida de distância (métrica) e satisfaz as seguintes propriedades:

(i) é igual a zero se, e somente se, ρ_1 e ρ_2 são iguais;

(ii) é simétrica;

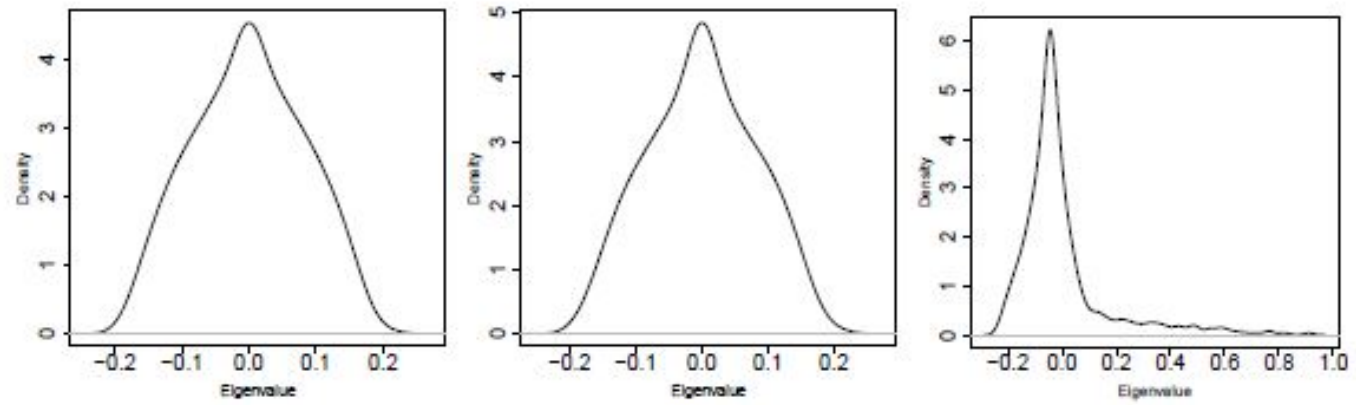
(iii) é não negativa,

(iv) satisfaz a desigualdade triângular.

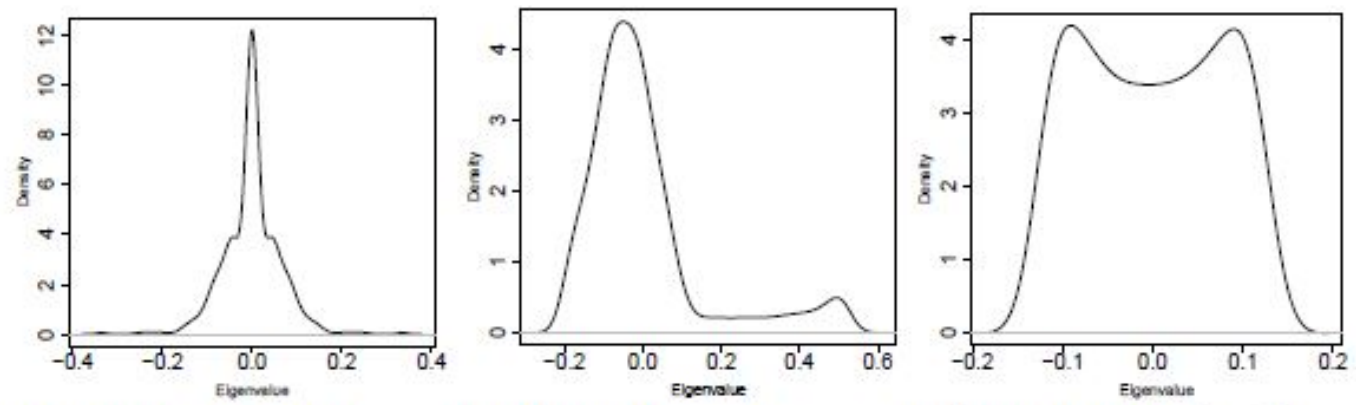
A divergência de KL serve ao propósito da estimação de parâmetro e seleção de modelo. Entretanto, não é simétrico, i. e., em geral, $KL(\rho_1|\rho_2) \neq KL(\rho_2|\rho_1)$. Por essa razão, a divergência de KL não é adequada quando não se sabe ao certo qual distribuição é a distribuição de referência. Este é o caso para o teste estatístico comparando espectros ρ_1 e ρ_2. Enquanto a entropia quantifica a incerteza associada a uma variável aleatória, a divergência de Kullback-Leiber (KL) mede a quantidade de informação perdida quando uma distribuição de probabilidade é utilizada para aproximar outra. Para grafos, a divergência de KL pode ser utilizada para discriminar distribuições de probabilidade e para selecionar o modelo de

grafo que melhor descreve o grafo observado. Pela sua definição, quando duas densidades espectrais forem diferentes, os grafos aleatórios correspondentes são diferentes. Entretanto, o contrário não é necessariamente verdade, uma vez que grafos aleatórios diferentes podem ter a mesma densidade espectral.

No próximo capítulo a divergência de KL será usada em uma aplicação como critério de seleção do modelo para selecionar bons modelos entre um conjunto de candidatos. Mais especificamente, dado um grafo, é importante decidir se o grafo mais provavelmente foi gerado pelo ER,*scale-free* ou *small-world*. A divergência de KL entre espectros de grafos e o espectro de diferentes classes de grafos pode ser interpretada como a qualidade de *'fitting'* o grafo a um modelo.



(a) Erdős-Rényi model ($m = 0.007N$) (b) Gilbert model ($p = 0.007$) (c) Geometric model ($r = 0.1$)



(d) Barabási-Albert model $p_s = 1$ (e) Watts-Strogatz model ($p_r = 0.07$) (f) k-regular model ($k = 3$)

Figura 3.1: *Estimativas da densidade espectral para grafos com 500 vértices gerados pelos modelos Erdõs-Rényi (a), Gilbert (b), Barabási-Albert (d), Watts-Strogatz (e) e k-regular(f). Em (a) o número de arestas é igual a* $0.007N = o.0007(5002)$. *Em (b), a probabilidade p de conectar dois vértices é igual a* 0.0007. *Em (c) o raio* $r = 0.1$. *Em (d) o expoente de escala* $p_S = 1$. *Em (e) a probabilidade de reconectar um vértice* p_r *é 0.07. Em (f) o grau de cada vértice é* $k = 3$. *Fonte: (de Siqueira Santos et al., 2016)*

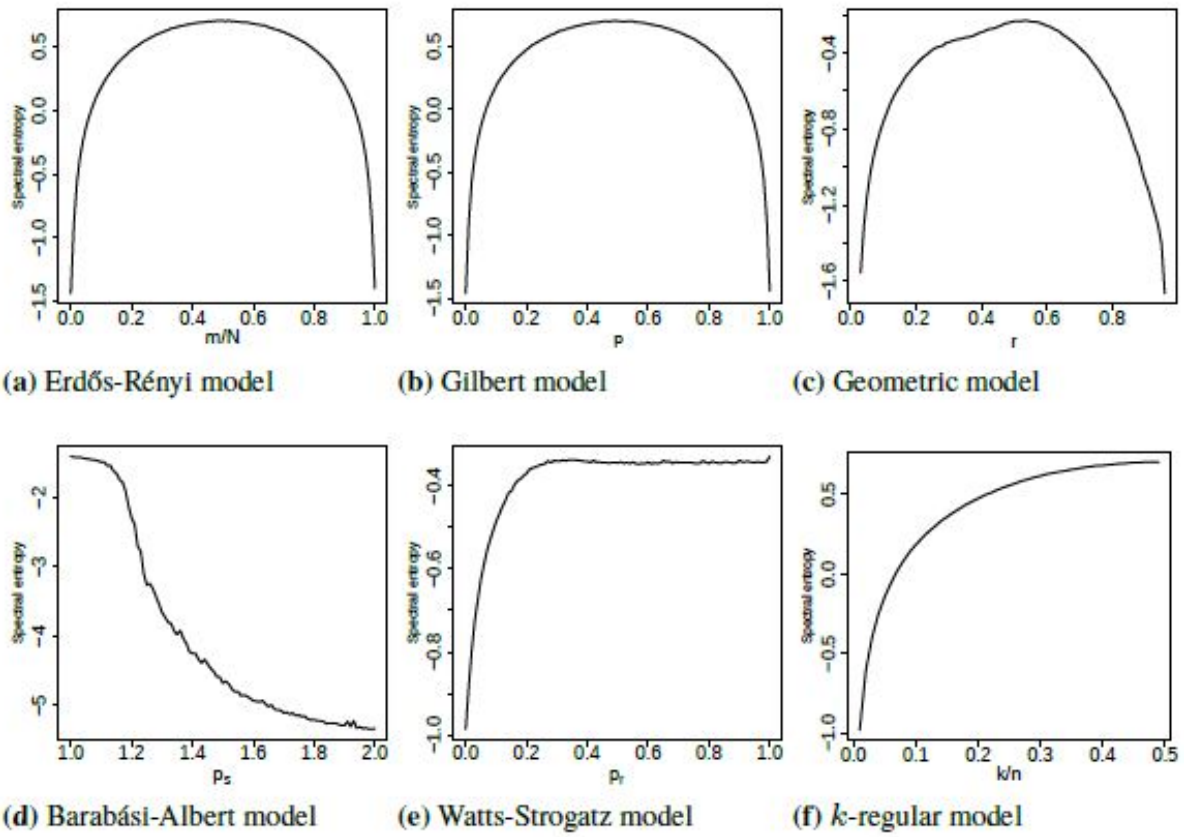


Figura 3.2: *Entropia espectral de grafos. A entropia espectral empírica (eixo y) para os modelos Erdõs-Rényi (a), Gilbert (b), Geometric (c), Barabási-Albert (d), Watts-Strogatz (e) e k-regular(f). Em (a), (b), (c) e (e) os valores no eixo x variam entre 0 e 1. Em (d), os valores variam entre 1 e 2. Em (f), os valores variam entre 0 e 0,5. Em (b), (c), (d) e (e) o eixo x corresponde aos parâmetros p, r, p_s e p_r, respectivamente. Em (a), o parâmetro m é obtido pela multiplicação dos valores no eixo x por $\binom{N=n}{2ln(\rho)}$. Em (f), k é obtido pela multiplicação do valor no eixo x por n. Para cada modelo, a entropia espectral foi obtida com 50 grafos de tamanho $n = 500$. Fonte: (de Siqueira Santos et al., 2016)*

Capítulo 4

Métodos estatísticos para grafos

Uma vez que estão definidos grafos aleatórios e divergências entre grafos aleatórios, várias questões tornam-se pertinentes. Dado um grafo g e um modelo aleatório $\mathcal{M}(\theta)$ com parâmetro θ, é possível medir o quão bem $\mathcal{M}(\theta)$ descreve g? Dado um conjunto de modelos de grafos aleatórios $S = \{M_1, M_2, ..., M_k\}$, qual destes modelos melhor descreve g? Como foi visto no capítulo 3, a densidade espectral descreve várias propriedades estruturais de grafos aleatórios. Então, para medir quão bem um modelo de grafo aleatório $M(\theta)$ descreve um grafo g podemos medir as diferenças entre a densidade espectral do grafo aleatório g e a densidade espectral do grafo aleatório gerado pelo modelo M com parâmetro θ. E se quiséssemos comparar grafos de diferentes populações?

As abordagens tradicionais baseadas em algoritmos para determinar isomorfismo não são adequadas, pois além da complexidade computacional, tais abordagens não representam a variação intrínseca presente em redes do mundo real. Uma solução para este problema é assumir que os grafos obtidos no mundo real são gerados por processos probabilísticos (modelos de grafos aleatórios) e então compará-los para saber se foram gerados pelo mesmo modelo. Entretanto, na prática, raramente sabemos qual modelo gerou o grafo. Então, o primeiro passo é selecionar o modelo.

Nas seções seguintes métodos estatísticos de seleção de modelo e de comparação entre conjuntos de grafos serão apresentados.

4.1 Seleção do modelo e estimativa dos parâmetros

Dado um grafo g e seu espectro ρ_g, alguns modelos de grafo são candidatos a ser ranqueados de acordo com a divergência KL e os modelos com menor KL devem ser considerados bons candidatos para explicar o dado. Então, a divergência KL fornece um critério objetivo de comparação entre modelos, i. e., uma ferramenta para seleção do modelo

Visando a comparação entre grafos de diferentes populações e considerando que para a maioria dos grafos obtidos no mundo real o modelo que os gerou é desconhecido, Takahashi *et al.* (2012) apresentaram uma abordagem que pode ser descrita em dois passos.

Seja $S = \{M_1, M_2, ..., M_k\}$ um conjunto de modelos de grafos aleatórios e g um grafo aleatório. Para escolher qual $M_i \in S$ melhor descreve g:

1. Para um dado modelo estima-se o parâmetro. Seja ρ_g a densidade espectral de um grafo aleatório g. Dado um modelo de grafo aleatório M, seja θ um vetor contendo os valores para cada parâmetro de M. Se todas as possíveis escolhas de θ forem consideradas, então o modelo M gera uma família paramétrica de densidades espectrais $\{\rho_\theta\}$.

Assumindo que existe um valor de θ que minimiza $KL(\rho_g|\rho_\theta)$, denotado por θ^* então

$$\theta^* = arg\min_\theta KL(\rho_g|\rho_\theta$$

Entretando, nas aplicações reais, a densidade espectral (ρ_g) é desconhecida. Portanto, na prática, um estimador $\hat{\rho_g}$ para ρ_g é usado. Então o estimador $\hat{\theta}$ de θ^*

$$\hat{\theta} = arg\min_\theta KL(\hat{\rho_g}|\rho_\theta$$

2. Dado o estimador do parâmetro ($\hat{\theta}$), o parâmetro de cada modelo é estimado e então o modelo que minimiza a divergência de KL é selecionado. Sejam $\{\rho_{\theta_1}\}, \{\rho_{\theta_2}\}, ..., \{\rho_{\theta_k}\}$ famílias paramétricas da densidade espectral, θ_i, para $i = 1, 2, ..., k$ os parâmetros estimados usando o estimador $\hat{\theta}$ de θ_i e $\#(\theta_i)$ a dimensão de θ_i. Então, o melhor candidato $\hat{\theta_j}$ é selecionado por

$$j = arg\min_i 2KL(\hat{\rho_g}|\rho_{\hat{\theta_i}}) + 2\#(\hat{\theta_i})$$

A motivação para esse critério é AIC (Akaike Information Criterion) (**?**), A penalização $2\#(\hat{\theta_i})$ para evitar 'overfitting'. Os três modelos aleatórios analisados tem o mesmo número de parâmetros; portanto, o termo de penalidade não é necessário aqui, mas pode ser em situações mais genéricas. Nos trabalhos analisados, todos os modelos são gerados pelo mesmo n úmero de parâmetros.

Dado um grafo g e seu espectro ρ_g, alguns modelos de grafo são candidatos a ser ranqueados de acordo com a divergência KL e os modelos com menor KL devem ser considerados bons candidatos para explicar o dado. Então, a divergência KL fornece um critério objetivo de comparação entre modelos, i. e., uma ferramenta para seleção do modelo. Especificamente, seja $\hat{\rho_g}$ a distribuição espectral empírica e $\{\rho_{theta_1}\}, ..., \{\rho_{\theta_m}\}$ as m famílias diferentes de distribuição espectral.

4.2 Teste de hipótese entre coleções de grafos

Sejam T_1 e T_2 duas coleções de grafos. Assumindo que todos os grafos de T_1 foram gerados pelo mesmo modelo de grafo aleatório M_1 com parâmetro θ_1 e que os grafos de T_2 foram gerados por um modelo M_2 com parâmetro θ_2, é possível testar se $M_1 = M_2$ e $\theta_1 = \theta_2$?

Uma abordagem para tratar esta questão é medir as diferenças entre as duas coleções de grafos. Neste problema a distribuição referência não é conhecida, e como a divergência de KL é assimétrica, não é possível realizar um teste de hipóteses nesta condição.

Como a divergência de Jensen-Shannon é a versão simétrica da divergência de KL, esta divergência se torna um candidato natural para o teste de hipóteses entre coleções de grafos com densidades espectrais ρ_1 e ρ_2, respectivamente. As hipóteses a serem testadas são:

$$H_0 : JS(\rho_1, \rho_2) = 0$$

versus

$$H_1 : JS(\rho_1, \rho_2) > 0$$

Dada uma amostra de g_1, denotada por T_1 e uma amostra de g_2, denotada por T_2, o teste estatístico é $JS(\hat{\rho_1}, \hat{\rho_2})$, onde $\hat{\rho_1}$ e $\hat{\rho_2}$ são as estimativas das densidades espectrais médias obtidas de T_1 e T_2, respectivamente.

Após escolher a estatística a ser testada, e as hipóteses nula e alternativa, é preciso obter o p-valor, que representa a probabilidade da estatística testada ser pelo menos tão extrema quanto o valor observado nos dados, assumindo que a hipótese nula é verdadeira. Para obter o p-valor para o teste de hipótese, é preciso obter a distribuição da estatística sob a hipótese nula. Isto é feito a partir de uma distribuição assintótica ou através de reamostragens dos dados. Como não são conhecidas fórmulas analíticas não é possível usar esta estratégia, e que ao adotar a abordagem baseada em reamostragens, geralmente é feito usando o método de Monte Carlo, e, portanto, é preciso assumir que as variáveis são independentes e identicamente distribuídas, o que não pode ser garantido neste caso.

Takahashi *et al.* (2012) propuseram um teste para a divergência de Jensen-Shannon entre densidades espectrais que usa *bootstrap* para reamostrar os dados. O *bootstrap* foi proposto porEfron (1979) para estimar a distribuição amostral de uma estatística a partir de reamostragem aleatória dos dados.

Sejam n_1 e n_2 o número de grafos em T_1 e T_2, respectivamente, B o número desejado de replicações via *bootstrap*, e $T = T_1 \cup T_2$. O procedimento para o teste de hipótese entre T_1 e T_2 é descrito a seguir:

1. Calcule $JS(\hat{\rho}_1, \hat{\rho}_1)$

2. Reamostre com substituição n_1 grafos de T e construa um novo conjunto (conjunto *bootstrap*) $\tilde{T}_1$. Obtenha um estimador para a densidade espectral média de $\tilde{T}_1$, que é denotado por $\rho\mathcal{T}_1$.

3. Reamostre com substituição n_2 grafos de T e construa um novo conjunto (conjunto *bootstrap*) $\tilde{T}_2$. Obtenha um estimador para a densidade espectral média de $\tilde{T}_2$, que é denotado por $\rho\mathcal{T}_2$.

4. Calcule $JS(\rho\mathcal{T}_1, \rho\mathcal{T}_2)$.

5. Repita os passos 2-4 B vezes.

6. O p-valor é a proporção de replicações do *bootstrap* nas quais $JS(\rho\mathcal{T}_1, \rho\mathcal{T}_2) > JS(\hat{\rho}_1, \hat{\rho}_1)$

Com o valor da a estatística a ser testada calculada, e as hipóteses nula e alternativa, é preciso obter o p-valor, que representa a probabilidade da estatística testada ser pelo menos tão extrema quanto o valor observado nos dados, assumindo que a hipótese nula é verdadeira. Para obter o p-valor para o teste de hipótese, é preciso obter a distribuição da estatística sob a hipótese nula. Isto é feito a partir de uma distribuição assintótica ou através de reamostragens dos dados. Como não são conhecidas fórmulas analíticas não é possível usar esta estratégia, e que ao adotar a abordagem baseada em reamostragens, geralmente é feito usando o método de Monte Carlo, e, portanto, é preciso assumir que as variáveis são independentes e identicamente distribuídas, o que não pode ser garantido neste caso.

4.3 Análise da variabilidade da Estrutura do Grafo - ANOGVA

A Análise da variabilidade da Estrutura do Grafo - ANOGVA (*Analysis of Graph Structure Variability*) foi proposta por Fujita *et al.* (2017) e pode ser descrita como: dadas k populações de grafos $g_1, g_2, ..., g_k$, o teste consisiste em verificar se todas as populações de

grafos foram geradas pelo mesmo modelo aleatório. Para isso, todas as distribuições espectrais são testadas para verificar igualdade.

Sejam $\hat{\rho_{g_1}}, \hat{\rho_{g_2}}, ..., \hat{\rho_{g_k}}$ a densidade espectral estimada das populações de grafos $g_1, g_2, ..., g_k$, respectivamente, onde $\hat{\rho_{g_i}} (i = 1, ..., k)$ é a média do espectro dos grafos na população g_i. Seja também $\hat{\rho_{g_M}} = \frac{\sum_{i=1}^{k} \hat{\rho_{g_i}}}{k}$. O suporte de $\hat{\rho_{g_M}}$ inclui o suporte de $\hat{\rho_{g_i}}$, para qualquer i. Formalmente, é testado:

$H_0 : KL(\rho_{g_1}, \rho_{g_M}) = KL(\rho_{g_2}, \rho_{g_M}) = ... = KL(\rho_{g_k}, \rho_{g_M}) = 0$, i.e., os grafos de $g_1, g_2, ..., g_k$ são gerados pelo mesmo modelo de grafo aleatório (distribuições espectrais são iguais).

H_1 : Pelo menos uma população de grafos é gerada de outra maneira

Para este teste, foi usada a generalização da divergênciade Jensen-Shannon (Shannon , 1948; **?**) para $k > 2$, que é a estatística:

$$\Delta = \sum_{i=1}^{k} KL(\hat{\rho_{g_i}}, \hat{\rho_{g_M}}).$$

Sob a hipótese nula é esperado que Δ seja pequeno, quando Δ é alto, a hipótese nula deve ser rejeitada. A distribuição assintótica de Δ não é conhecida, portanto foi usado um procedimento computacional baseado no teste de permutação para construir uma distribuição empírica. Os passos deste teste de permutação são:

1. Construa amostras permutadas g_i^* para $i = 1, ..., k$ reamostrando (sem substituição) $|g_i|$ grafos do conjunto completo $\{g_1 \cup g_2 \cup ... \cup g_k\}$
2. Calcule $\hat{\rho_{g_i^*}}$ para cada g_i^* $(i = 1, ..., k)$.
3. Calcule $\hat{\Delta}^* = \sum_{i=1}^{k} KL(\hat{\rho_{g_i^*}}, \hat{\rho_{g_M}})$
4. Repita os passos 1 a 4 até que o número de replicações tenha sido obtido
5. O p-valor da estatística observada $\hat{\Delta}$ é a quantidade de vezes que $\hat{\Delta}^*$, obtido no conjunto permutado, é pelo menos tão grande quando $\hat{\Delta}$ estimado no conjunto original.

A figura 4.1 ilustra a ideia do teste ANOGVA: a distribuição espectral de cada população é comparada com a distribuição referência (a média das distribuições espectrais, ρ_{gM}). Se a soma das distâncias (divergência KL) é grande, significa que pelo menos uma das distribuições espectrais é diferente da referência, ou seja, pelo menos uma das populações foi gerada por um modelo e/ou conjunto de parâmetros diferente.

As abordagens apresentadas para tratar esta questão é medir as diferenças entre as duas coleções de grafos. Neste problema a distribuição referência não é conhecida, e como a divergência de KL é assimétrica, não é possível realizar um teste de hipóteses nesta condição. Como a divergência de Jensen-Shannon é a versão simétrica da divergência de KL, esta divergência se torna um candidato natural para o teste de hipóteses entre coleções de grafos com densidades espectrais ρ_1 e ρ_2, respectivamente.

A limitação principal dessas abordagens é o fato de só serem aplicáveis a conjuntos de grafos, i. e., é necessário observar replicações dos grafos em cada população. De fato, as estratégias mostradas neste capítulo foram aplicadas a testes em redes cerebrais funcionais para verificar se os grafos dos indivíduos controle e diagnosticados com uma desordem cerebral são gerados pelo mesmo modelo de grafo (busca semelhanças entre os grafos de indivíduos

dos dois grupos). Neste caso particular, centenas de grafos (um por indivíduos) estavam disponíveis para o teste. Portanto, o teste baseado em *bootstrap* consistiu em reamostrar os indivíduos independentemente. Entretanto, este não é o caso para a maioria das aplicações, onde geralmente apenas uma observação (grafo) está disponível. Por exemplo, nós geralmente observamos apenas uma rede social, uma rede de interação proteína-proteína por espécie, uma rede www, etc. Portanto, o problema consiste em, dadas duas ou mais condições representadas cada uma por um grafo (sem réplicas), testar se os grafos foram gerados pelo mesmo modelo e parâmetros. No próximo capítulo, descreveremos um método que, se os grafos forem grandes o suficiente, é de fato possível testar a igualdade deles.

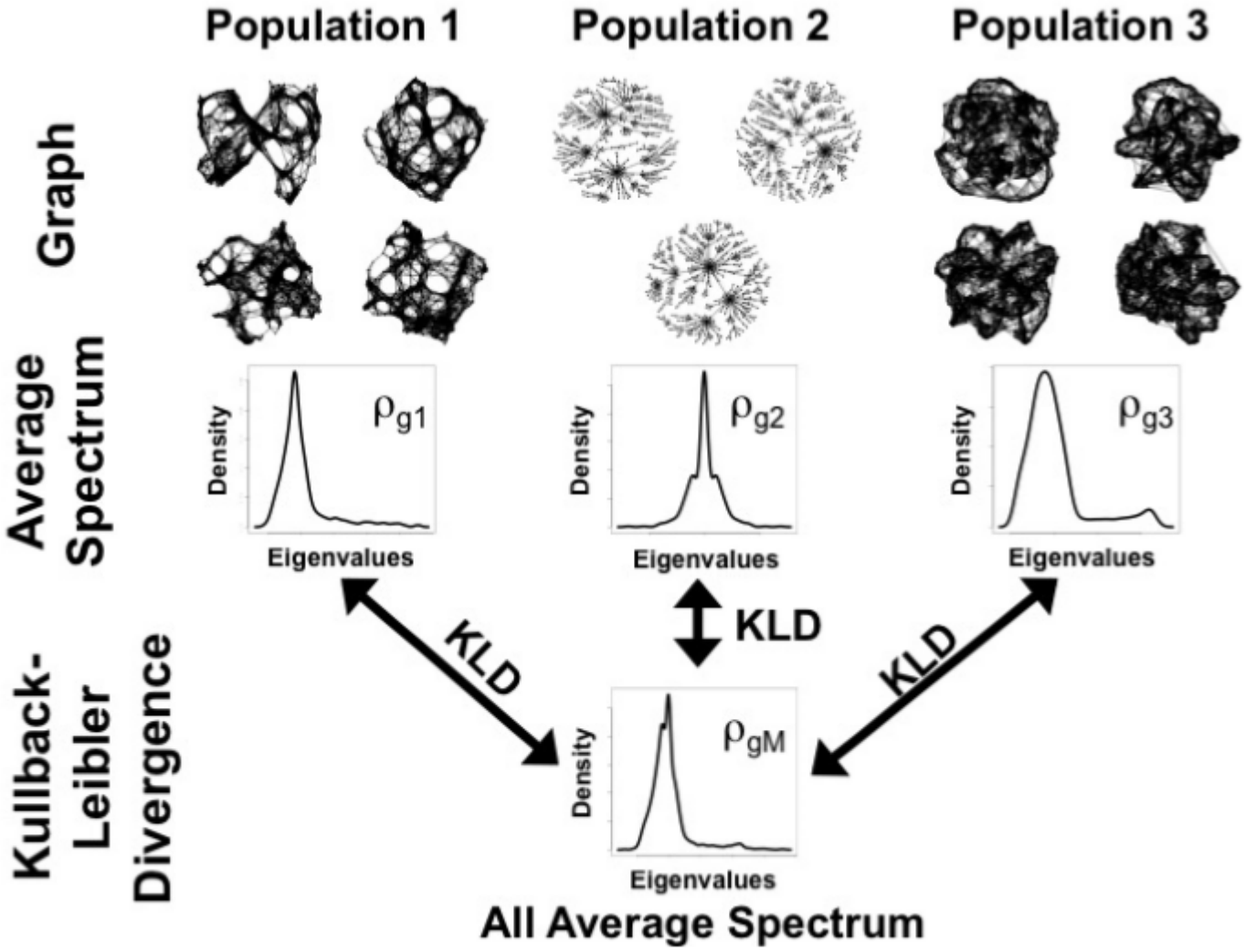


Figura 4.1: *Esquema do ANOGVA. Neste exemplo, $k = 3$ populações de grafos são testadas para verificar se elas foram geradas pelo mesmo modelo aleatório de grafo. Primeiro, a distribuição espectral de cada grafo é estimada e então, a média da distribuição espectral de cada população é estimada ($\rho_{gi}(i = 1, ..., k)$). Então a distribuição espectral média de todas as distribuições espectrais (ρ_{gM} é estimada (média da distribuição média)) Finalmente, a soma da divergência de Kullback-Leibler (KLD) entre $\rho_{gi}(i = 1, ..., k)$ e ρ_{gM} é calculada. Sob a hipótese nula, i. e., quando todos as $k = 3$ populações de grafos são geradas pelo mesmo modelo de grafo aleatório, é esperado Δ pequeno. Fonte: (Fujita et al., 2017)*

Capítulo 5

Testes semiparamétricos para grafos

No capítulo anterior, vimos trabalhos nos quais conjuntos de grafos são comparados para identificar se os grafos que compõem cada conjunto foram gerados pelo mesmo modelo e parâmetros. Tal estratégia apresenta a limitação de só ser aplicável a conjuntos de grafos, ou seja, é necessário observar replicações do grafo em cada população. Para as situações em que centenas de grafos estão disponíveis para o teste, o procedimento de *bootstrap* consiste em reamostrar os grafos independentemente. No entanto, a maior parte das aplicações conta com apenas uma amostra de grafo. Exemplos de aplicações nas quais não encontramos replicações dos grafos são: redes sociais, redes de interação proteína-proteína, a rede internacional de computadores (www - world wide web), etc. O problema, então, consiste em como testar se dois grafos, com apenas uma ocorrência de cada um deles, foram gerados pelo mesmo modelo?

Neste capítulo, será mostrado que o teste de igualdade de grafos pode ser feito sem necessidade de replicar os grafos testados, bastando apenas que o grafo seja grande o suficiente. Para tanto, será definido formalmente o problema e apresentado um teste baseado na razão de verossimilhança (LRT). A seguir, métodos para estimativa dos parâmetros que serão utilizados na generalização do teste LRT para grafos de modelos diferentes do ER, como grafos geométricos, WS e BA. Então, um teste baseado no tradicional teste t de Student é apresentado para testar dois grafos de um modelo. Finalmente, apresentamos o teste para testar simultaneamente dois ou mais grafos.

5.1 Definição do teste paramétrico

Uma vez que o modelo que gerou os grafos é conhecido, o teste paramétrico se resume em testar a igualdade entre seus parâmetros. Podemos então definir o teste paramétrico como: dados dois grafos g_1 e g_2 e o modelo que os gerou (o caso em que os grafos foram gerados por modelos distintos não precisa ser testado), vamos verificar, através de testes paramétricos se os parâmetros usados para gerar tais grafos são iguais. Formalmente: Sejam θ_1 e θ_2 os parâmetros usados para gerar g_1 e g_2, respectivamente, vamos testar

$$H_0 : \theta_1 = \theta_2$$

versus

$$H_1 : \theta_1 \neq \theta_s.$$

5.2 Parâmetros

Para o teste paramétrico entre grafos, partimos da suposição que o modelo é conhecido, mas os parâmetros não, então é preciso estimar os parâmetros para compará-los. Nesta seção, serão apresentados algoritmos para obtenção dos parâmetros dos modelos analisados nas simulações. Inicialmente, o caso do modelo ER, que é direto e intuitivo, permitindo sua utilização em testes como o de razão de verossimilhança. Na sequência, a obtenção dos parâmetros será generalizada utilizando a divergência de Kullback-Leiber para os outros modelos estudados.

5.2.1 Estimativa do parâmetro p do modelo ER

No modelo ER cada aresta é adicionada com probabilidade $p \in (0, 1)$ independente da existência de outras arestas. Sejam dois grafos g_1 e g_2 gerados pelo modelo ER com parâmetros p_1 e p_2, respectivamente. Os grafos g_1 e g_2 são gerados por um Processo de Bernoulli, onde é definido se uma aresta conectando dois vértices i e j existe ou não com probabilidade p_1 e p_2, respectivamente. A probabilidade p_1 e p_2 é igual para todos os pares de vértices dos grafos g_1 e g_2.

Devido ao processo de geração de um grafo ER, é fácil ver que p_1 e p_2 seguem distribuições de Bernoulli. Então, o problema consiste em testar a igualdade dos parâmetros de duas distribuições de Bernoulli.

Sejam $l_{ij} \in \{0, 1\}$ uma variável aleatória de Bernoulli indicando a presença de uma aresta i, j. Para o modelo ER, as variáveis l_{ij} são independentes e assumem os valores:

$$l_{ij} = \{1\text{com probabilidade p}, 0\text{com probabilidade 1-p}$$

O número esperado de arestas de um grafo gerado pelo modelo ER é

$$E[\text{número esperado de arestas}] = E[\sum l_{ij}] = \frac{n(n-1)}{2}p$$

Portanto, um estimador natural para $\hat{p}_i$ $(i = 1, 2)$ é simplesmente dividir o número de arestas em um grafo g_i pelo número total de possíveis arestas $(n_i \times n_i - n_i)/2$.

5.2.2 Estimativa do parâmetro $\hat{\theta}$ do modelo $\mathcal{M}$

Para o caso em que o grafo tenha sido gerado por um modelo diferente do ER, a estimativa do parâmetro não é direta e intuitiva. Neste trabalho, foram analisados três modelos além do ER: Geométrico (GRG), Wattz-Strogatz (WS), Barabási-Albert (BA). Todos eles possuem apenas um parâmetro a ser estimado, portanto, para estimar o parâmetro $\hat{\theta}$ de cada modelo de grafo $\mathcal{M}$ analisado, foi adotado o procedimento baseado na distribuição espectral do grafo proposto por (Takahashi *et al.*, 2012). Este procedimento é descrito a seguir.

Seja δ a função delta de Dirac (que pode ser usada para generalizar fórmulas para variáveis aleatórias discretas e contínuas) que satisfaz:

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

Seja g um grafo gerado pelo modelo $\mathcal{M}$, λ o conjunto de autovalores da matriz de adjacência do grafo g, e os "<>" indicam o valor esperado com respeito a lei de probabilidade do grafo aleatório, então a densidade espectral ρ_g de um grafo aleatório g é

$$\rho_g(\lambda) = \lim_{n\to\infty} < \frac{1}{n}\sum_{j=1}^{n} \delta(\lambda - \lambda_j/\sqrt{n}) > .$$

Entretanto, em aplicações reais, a densidade espectral ρ_g é desconhecida. Portanto, um estimador $\hat{\rho_g}$ de ρ_g é usado. Para obter $\hat{\rho_g}$, primeiro são computados os autovalores (λ) da matriz de adjacência do grafo g e então aplica a regressão *kernel* Gaussiana com o estimador Nadaraya-Watson (Sain e Scott, 1996) para regularizar o estimador. Finalmente, normalize a densidade para obter a integral abaixo da curva igual a um. O intervalo do *kernel* pode ser calculado pela diferença do maior autovalor com o menor autovalor dividido pelo número de classes (Nips 15), onde o número de classes pode ser selecionado usando um critério objetivo como Sturges Sturges (1926) ou SilmermansmSilverman (1986). Assim, um estimador $\hat{\theta}$ de θ^* é

$$\hat{\theta} = arg\min_{\theta} KL(\hat{\rho_g}|\rho_\theta)$$

onde KL é a divergência de Kullback-Leiber.

O procedimento para estimar $\hat{\theta}$ usando a divergência KL é descrito no algoritmo 1.

Algoritmo 1: Estimador de θ pela divergência Kullback-Leibler
Entrada: grafo g e modelo $\mathcal{M}$
Saída: O valor estimado para o parâmetro θ

1. Construa um grid com possíveis valores para θ.
2. Para um valor de θ no grid, construa B grafos usando o modelo $\mathcal{M}$, e estime a distribuição espectral para cada um deles $(\hat{\rho}_{\theta_b}, b = 1, ..., B)$. Então, calcule a média deles $\hat{\rho}_\theta = \frac{\sum_{b=1}^{B} \hat{\rho}_{\theta_b}}{B}$. Aqui, nós consideramos que a distribuição espectral analítica do grafo g gerada pelo modelo $\mathcal{M}$ e parâmetro θ é desconhecida, e, portanto, nós estimamos pela simulação e considere que para grandes valores de B, $\overline{\rho}_\theta \to \rho_\theta(\hat{\rho}_\theta = \overline{\rho_\theta})$.
3. Estime a divergência de Kullback-Leibler entre $\hat{\rho_g}$ e $\hat{\rho}_\theta$
4. Repita os passos 2 e 3 para todos os valores de θ no grid.
5. Selecione o argumento θ que minimiza a divergência KL entre $\hat{\rho_g}$ e $\hat{\rho}_\theta$.

5.3 Estimador da variância

A distribuição assintótica de θ é desconhecida, portanto, é difícil (ou impossível) estimar a sua variância analiticamente (σ_θ^2). Para isso, nós sugerimos um bootstrap paramétrico descrito no algoritmo 2

Algoritmo 2: Estimar σ_θ^2 por bootstrap Entrada: modelo $\mathcal{M}$ e o parâmetro estimado $\hat{\theta}$ de g Saída: A variância estimada de θ ($\hat{\sigma_\theta^2}$)

1. Simule B grafos $(g^{1*}, g^{2*}, ..., g^{B*})$ usando o modelo $\mathcal{M}$ e o parâmetro estimado $\hat{\theta}$ de g.
2. Estime os parâmetros $\hat{\theta_1^*}, \hat{\theta_2^*}, ...\hat{\theta_B^*}$ de $g^{1*}, g^{2*}, ..., g^{B*}$ usando o estimador baseado na divergência de Kullback-Leiber.
3. Estime a média das amostras do bootstrap $\hat{\theta}^*(.) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta_b^*}$.
4. Estime a variância de θ: $\hat{\sigma}_\theta^2 = \frac{\sum_{b=1}^{B}(\hat{\theta_b^*} - \hat{\theta}^*(.))^2}{B-1}$.

5.4 Teste baseado em razão de verossimilhança (LRT)

O teste baseado em razão de verossimilhança (*likelihood ratio test - LRT*) é um método estatístico usado para comparar o quão bem dois modelos estatísticos (um modelo nulo e o outro, um caso especial do modelo nulo, chamado de modelo alternativo) são ajustados. O teste é baseado na razão de verossimilhança, que expressa quantas vezes mais provável os dados estão em um modelo em relação ao outro. O resultado desta razão é usado para obter o p-valor e então decidir se o modelo nulo é rejeitado.

5.4.1 Teste baseado em razão de verossimilhança para grafos gerados pelo modelo ER

Seja n_1 e n_2 os números de vértices e p_1 e p_2 os parâmetros dos grafos g_1 e g_2 gerados pelo modelo ER, respectivamente. O teste de hipótese consiste em testar

$$H_0 : p_1 = p_2$$

versus

$$H_1 : p_1 \neq p_2$$

Como foi visto na seção 5.2.1, um estimador para os parâmetros p_1 e p_2 do modelo ER é intuitivo e segue distribuição de Bernoulli. Então, o problema é testar a igualdade dos parâmetros de duas Bernoullis.

Um estimador natural para $\hat{p}_i$ $(i = 1, 2)$ é simplesmente dividir o número de arestas em um grafo g_i pelo número total de possíveis arestas $(n_i \times n_i - n_i)/2$. Baseado nas $(n_1 \times n_1 - n_1)/2$ amostras de Bernoulli do grafo g_1 e nas $(n_2 \times n_2 - n_2)/2$ amostras de g_2, é possível construir

um teste de razão de verossimilhança (LRT - *likelihood ratio test*). Seja $\hat{p}_0$ a proporção combinada $\hat{p}_0 = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n2}$, então o teste rejeita H_0 com nível de significância α se

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}_0(1-\hat{p}_0)(\frac{1}{n_1}+\frac{1}{n_2})} > z_{\alpha/2}$$

onde $z_{\alpha/2}$ é o $(100\alpha/2)^o$ maior percentil da distribuição normal padrão.

5.5 Teste t

O teste t de Student ou somente teste t é um teste de hipótese que avalia uma certa estatística t para rejeitar ou não a hipótese nula. A aplicação do teste t ocorre quando a estatística de teste segue uma distribuição normal, mas a variância (σ^2) é desconhecida.

O Teste t consiste em formular uma hipótese nula e consequentemente uma hipótese alternativa, calcular o valor de da estatística t e aplicá-lo à função densidade de probabilidade da distribuição t de Student medindo o tamanho da área abaixo dessa função para valores maiores ou iguais a t.

5.5.1 Teste t para grafos

Seguindo a mesma lógica do teste de verossimilhança, podemos desenvolver um teste similar ao tradicional Teste t de Student para o modelo aleatório ER. Entretanto, para grafos que não são gerados pelo modelo ER, não existe um estimador simples para o parâmetro do grafo (θ) ou para a variância (σ^2_θ). Portanto, para grafos que não são do modelo ER, para estimar o paramêtro e sua variância é usado o método proposto por Takahashi *et al.* (2012) e um procedimento de bootstrap Efron e Tibshirani (1993). Tais métodos serão descritos na sequência.

Para estimar o parâmetro $\hat{\theta}$, seja ρ_g a distribuição espectral de um grafo aleatório g. Dado um modelo de grafo aleatório $\mathcal{M}$, seja θ um vetor real contendo cada parâmetro de $\mathcal{M}$. Se considerarmos todas as escolhas possíveis para θ, então o modelo $\mathcal{M}$ gera uma família paramétrica de densidades espectrais ρ_θ. Assumindo que existe um valor θ que minimize a divergência de Kullback-Leiber entre ρ_g e ρ_θ ($\mathrm{KL}(\rho_g|\rho_\theta) = \int_{-\infty}^{+\infty} \rho_g(\lambda) log\frac{\rho_g(\lambda)}{\rho_\theta(\lambda)} d\lambda$), que é denotado por θ^*, nós então temos de Siqueira Santos *et al.* (2016); Takahashi *et al.* (2012)

$$\theta^* = argminKL(\rho_g|\rho_\theta)$$

.

Entretanto, em aplicações reais, a densidade espectral ρ_g é desconhecida. Portanto, na prática, um estimador $\hat{\rho_g}$ de ρ_g é usado de Siqueira Santos *et al.* (2016); Takahashi *et al.* (2012).

$$\hat{\theta} = argminKL(\hat{\rho_g}|\rho_\theta)$$

.

Para estimar as distribuições espectrais, primeiro calcule os autovalores da matriz de adjacência do grafo g e então aplique (Gaussian kernel regression) com o estimador (Nadaraya-Watson) Nadaraya (1964) para regularização do estimador. Finalmente, normalize a densidade para obter a integral abaixo da curva igual a um. O intervalo do kernel pode ser escolhido pela diferença entre o maior autovalor e menor autovalor dividido pelo número

de classes Sain e Scott (1996), onde o número de classes pode ser selecionado usando um critério objetivo, como Sturges Sturges (1926) ou SilmermansmSilverman (1986).

Na sequência, usamos o Algoritmo 2 para estimar a variância de $\theta(\sigma^2_\theta)$.

Após estimar os parâmetros $\hat{\theta_1}$ e $\hat{\theta_2}$ e suas variâncias $\sigma^2_{\theta_1}$ e $\sigma^2_{\theta_2}$ dos grafos g_1 e g_2, respectivamente, o teste rejeita $H_0 : \theta_1 = \theta_2$ com nível α se

$$z = \frac{\hat{\theta_1} - \hat{\theta_2}}{\sqrt{\sigma^2_{\theta_1} + \sigma^2_{\theta_2}}} > z_{alpha/2},$$

onde $z_{alpha/2}$ é o $(100\alpha/2)^o$ percentil da distribuição normal padrão.

5.6 Análise de variância - ANOVA

Análise de variância (ANOVA - *Analysis of variance*) é uma coleção de modelos estatísticos usados para analisar as diferenças entre médias de grupos e seus procedimentos associados (como a variação entre e dentro de grupos).

Seguindo as ideias descritas nas seções anteriores, é possível generalizar o teste para $m \geq 2$ grafos baseando-se na análise de variância.

5.6.1 Análise de variância para grafos

Dado o estimador para $\hat{\theta}$ e sua variância $\hat{\sigma^2_\theta}$, finalmente podemos descrever um teste estatístico para comparar dois ou mais grafos, nomedamente, a análise de variância para grafos.

Sejam g_i $(i = 1, ..., n)$ m grafos com n_i vértices e parâmetros θ_i, respectivamente. Gostariamos de testar

$$H_0 : \theta_1 = \theta_2 = ... = \theta_m$$

versus

$$H_1 : \text{pelo menos um dos parâmetros } (\theta_i, i = 1, 2, ..., m) \text{ é diferente.}$$

Similar à anova padrão, os parâmetros dos grafos podem ser decompostos em:

- variação Total (TV);
- variação entre grafos (variação condicional - CV) e
- variação interna (variação residual - RV).

De maneira que:

$$TV = CV + RV$$

O problema principal consiste em estimar CV e RV. O algoritmo 3 mostra como estimá-los e simultanemante testa m grafos de maneira análoga ao teste anova:

Algoritmo 3: ANOVA para grafos
Entrada: o modelo $\mathcal{M}$ e grafos g_1, g_2, ..., g_m
Saída: A significância do teste (p-value)

1. Estime os parâmetros $\hat{\theta}_1, \hat{\theta}_2, ...\hat{\theta_m}$ de $g_1, g_2, ..., g_m$ usando o estimador baseado na divergência de Kullback-Leiber descrito no algoritmo 1.
2. Estime as variâncias $(\hat{\sigma}^2_{\theta_i})$ para cada parâmetro $(\hat{\theta}_i,\ i = 1, 2, ..., m)$ usando o procedimento de bootstrap descrito no algoritmo 2 com B bootstrap replicates
3. Calcule a variação residual $RV = \sum_{i=1}^{m}(B-1) \times \hat{\sigma}^2_{\theta_i}$
4. Seja $\overline{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$ a média da amostragem do bootstrap, então calcule a variação condicional $CV = \sum_{i=1}^{m}(\hat{\theta}_i - \overline{\theta})^2$.
5. O teste F rejeita H_0 em nível α se

$$F = \frac{CV/(m-1)}{RV/(m \times B - m} > F_{m-1, m\times B-m}(\alpha)$$

onde $F_{m-1,m\times B-m}(\alpha)$ é o $(100\alpha)^o$ maior percentil da distribuição F com $(m-1)$ e $(m \times B - m)$ graus de liberdade.

Capítulo 6

Simulações e resultados

Visando avaliar o poder dos testes propostos foram realizadas simulações computacionais com quatro modelos de grafos aleatórios: Erdos-Rényi Erdös e Rényi (1959), Geométrico Penrose (2003), Watts-Strotgatz (Watts e Strogatz, 1998) e Barabási-Albert (Barabási e Albert , 1999). Neste capítulo o método de avaliação do desempenho do teste será explicado, assim como cada simulação e seus respectivos resultados.

6.1 Curva ROC

Quando um teste estatístico é proposto, pelo menos duas propriedades precisam ser demonstradas: o poder do teste sob a hipótese alternativa (H_1) e a taxa de controle de falsos positivos sob a hipótese nula (H_0). Para checar o poder do teste estatístico, i. e., se o método baseado na distribuição espectral de fato identifica entre dois conjuntos de grafos caracterizados por uma leve diferença nos parâmetros, curvas ROC foram construídas e comparadas ao teste baseado no grau da distribuição.

A curva ROC (*Receiver Operating Characteristics*) é um método gráfico que foi originalmente utilizado na avaliação da qualidade de transmissão de na detecção de sinais Egan (1975). Gráficos ROC são utilizados para avaliação, organização e predição de quaisquer classificadores binários, em várias áreas: mineração de dados, aprendizagem de máquina (*machine learning*) Bradley (1997); **?**, psicologia Green e Swets (1989), medicina Silva *et al.* (2004); **?**, economia Gastwirth (1971),previsão do tempo Mylne (2002) construção e refinamento de modelos **??**. Neste trabalho, utilizaremos curvas ROC para avaliar o desempenho nos testes de hipótese propostos.

Ao realizar um teste de hipóteses, temos os seguintes tipos de erros:

- Falso positivo: ocorre quando rejeitamos a hipótese nula quando ela é verdadeira. Este tipo de erro ocorre quando o teste de hipótese apresenta significância estatística, sendo que na verdade ele ocorreu por acaso. Este erro é chamado de erro tipo I.
- Falso negativo: também chamado de erro tipo II ocorre quando a hipótese nula não é rejeitada quando na verdade ela é falsa.
- Verdadeiro positivo: quando a hipótese nula é rejeitada sendo falsa.
- Verdadeiro negativo: a hipótese nula não é rejeitada e ela é de fato verdadeira.

O gráfico ROC é baseado na probabilidade de detecção, ou taxa de verdadeiros positivos e na probabilidade de falsos alarmes, ou taxa de falsos positivos. Para obter uma curva ROC usual, é preciso calcular a especificidade que pode ser obtida por (numero de verdadeiros

negativos)/(numero de verdadeiros negativos + numero de falsos positivos) e a sensibilidade definida como (número de verdadeiros positivos)/(número de verdadeiros positivos + número de falsos negativos). Então, a curva ROC de um classificador binário que rotula positivo e negativo é construída com um menos a especificidade no eixo x, ou seja, a taxa de falsos positivos e com a sensibilidade (taxas de verdadeiros positivos) no eixo y.

Para checar o poder do teste estatístico, i. e., se o método baseado na distribuição espectral de fato identifica entre dois conjuntos de grafos caracterizados por uma leve diferença nos parâmetros, curvas ROC foram construídas e comparadas ao teste baseado no grau da distribuição. A área abaixo da curva ROC é um resumo quantitativo do poder do teste. Em outras palavras, uma área com valor próximo a um (a curva acima da linha diagonal) denota alto poder enquanto uma área próxima a 0,5 (a curva próxima à linha diagnonal) é equivalente a decisões aleatórias.

Neste trabalho, realizamos testes estatísticos entre dois ou mais grafos gerados pelo mesmo modelo de grafos aleatórios para identificar se podem se foram gerados pelo mesmo conjunto de parâmetros. A hipótese nula de que os grafos dados foram gerados com o mesmo valor de parâmetro e a hipótese alternativa de que pelo menos um dos parâmetros é diferente. Para avaliar o poder estatístico e o controle da taxa de falsos positivos, a curva ROC foi adaptada para ter no eixo x o nível de significância (α) dos testes e no eixo y a proporção de rejeições da hipótese nula de que os grafos foram gerados com parâmetros distintos.

Desta forma, o nível de significância do teste (α) representa o limiar do p-valor para a rejeição da hipótese nula, ou seja, se o p-valor do teste for menor que α a hipótese nula é rejeitada. o poder empírico do teste estatístico é a proporção de rejeições da hipótese nula, que é quantificada pela área sob a curva ROC, com valor entre 0 e 1 e representa a probabilidade de rejeitar a hipótese nula quando ela é falsa.

Assim, quanto mais próxima de um for a área sob a curva ROC, maior será o poder estatístico do teste. Quando o valor da área está próximo de $0,50$, é equivalente a ter decisões aleatórias. Para dados gerados sob a hipótese nula, o valor esperado para a área da ROC é $0,5$, ou seja, é esperado que a curva ROC fique na diagonal. isto ocorre devido a definição de α (probabilidade de ocorrer um falso positivo), que é a probabilidade do p-valor do teste ser menor que α quando H_0 é verdadeira. Ou seja, sob H_0, o p-valor do teste segue distribuição uniforme. Na Figura 6.1, a linha tracejada na diagonal ilustra a curva ROC esperada sob a hipótese nula, a linha vermelha é a curva ROC de um teste com alto poder estatístico (método 1) e a linha verde é a curva de um teste com baixo poder estatístico (método 2).

6.2 Simulações

As simulações foram realizadas com quatro modelos aleatórios de grafos: ER, geométrico, WS e BA. Os grafos foram construídos a partir do uso do pacote *igraph* (http://igraph.org/r/) para *R* (https://www.r-project.org/). As funções utilizadas para gerar os grafos e seus respectivos parâmetros foram:

- ER: erdos.renyi.game - parametro p;
- GRG: grg.game - parâmetro raio r;
- WS: watts.strogatz.game - parâmetro p (p^w);
- BA: barabasi.game - parâmetro potência (p^s);

Durante as simulações, foi possível perceber que para os grafos gerados pelos modelos WS e BA, os grafos deveriam ser maiores que para os grafos gerados pelos modelos e ER

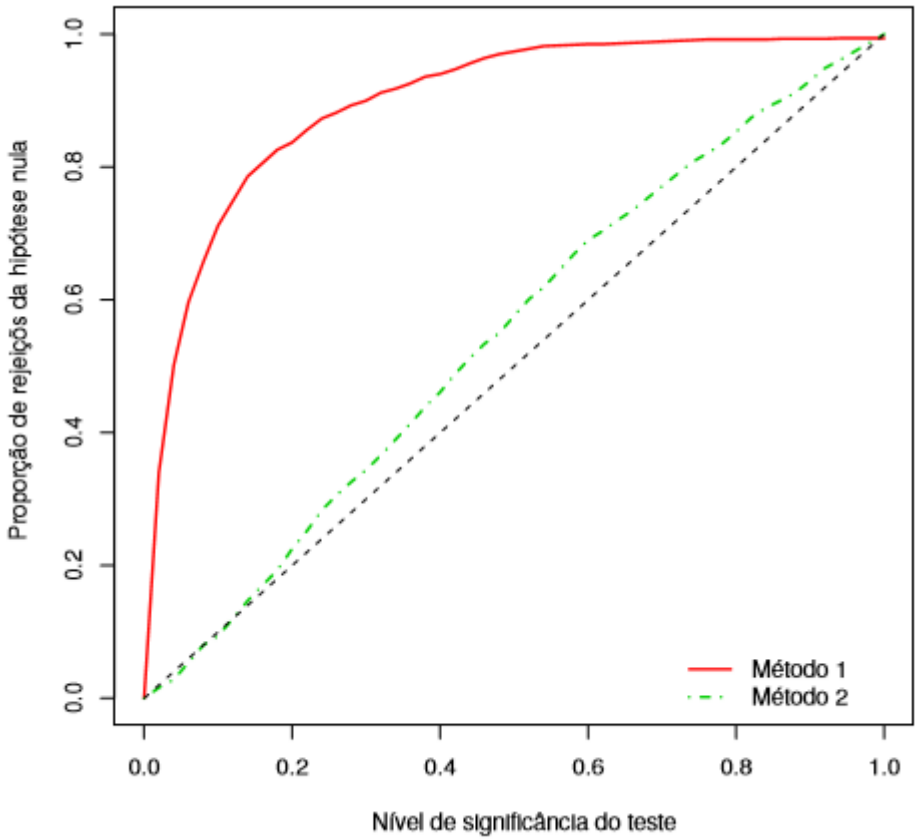


Figura 6.1: *Curvas ROC construídas a partir de dois métodos. A linha tracejada na diagonal ilustra a curva ROC esperada sob a hipótese nula. A linha vermelha é a curva ROC de um teste com alto poder estatístico (método 1) e a linha verde é a curva de um teste com pouco poder estatístico (método 2). Fonte: (**?**)*

e GRG para que o teste proposto fosse capaz de controlar efetivamente a taxa de falsos positivos. Isto se deve, provavelmente, à precisão do parâmetro estimado uma vez que Nips 7 mostrou que a taxa de convergência para o parâmetro estimado para WS e BA é mais lenta que para ER e GRG de Siqueira Santos *et al.* (2016).

Para cada simulação e cada tamanho de grafo n, o experimento foi repetido 500 vezes.

6.3 Simulação: comparativo entre os métodos

Esta simulação foi realizada para analisar o desempenho dos três testes desenvolvidos neste trabalho: teste baseado no teste *t Student*, teste de razão de verossimilhança e o teste ANOVA para grafos. Para verificar o controle do erro tipo I e para avaliar o poder de cada um dos testes, foram gerados dois grafos com o número de vértices $n_1 = n_2$ variando em 50, 75, 100 e 125 e o conjunto de parâmetros como se segue:

1. ER: $p_1 = p_2 = 0,5$ (sob a hipótese nula) e $p_1 = 0,50$ e $p_2 = 0,52$ (sob a hipótese alternativa)
2. Geometric: $r_1 = r_2 = 0,5$ (sob a hipótese nula) e $r_1 = 0,3$, $r_2 = 0,51$ (sob a hipótese alternativa)
3. WS $p_1^w = p_2^w = 0,5$ (sob a hipótese nula) e $p_1^w = XX$, $p_2^w = 0,51$ (sob a hipótese alternativa)

4. BA $p_1^s = p_2^s = 0,5$ (sob a hipótese nula) e $p_1^s = XX$, $p_2^s = 0,51$ (sob a hipótese alternativa)

6.3.1 Resultados

Pela figura 6.2 é possível perceber que os três testes apresentam desempenho semelhante quando aplicados à dois grafos. O que motiva e justifica a generalização do ANOVA para grafos para 2 ou mais grafos.

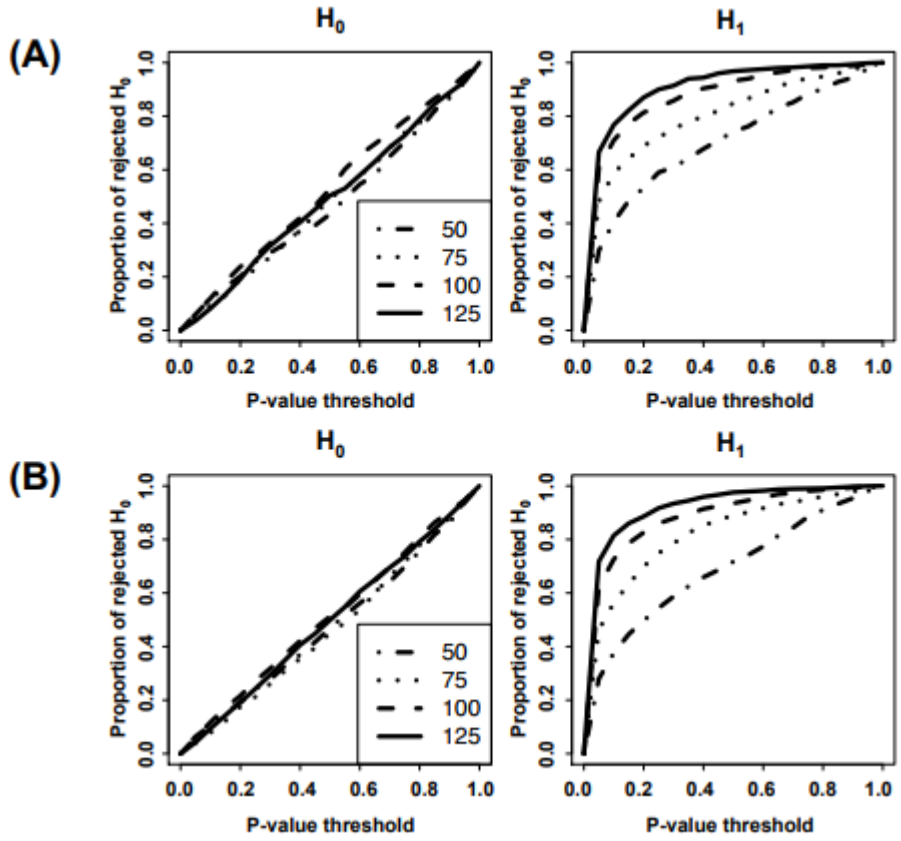


Figura 6.2: *Curvas ROC dos três testes propostos. teste ANOVA para grafos desbalanceados (diferentes números de vértices). O eixo x representa o limiar do p-valor, enquanto o eixo y representa a proporção de rejeições da hipótese nula. Os tipos diferentes de linha (sólido e tracejado) representam as curvas ROC obtidas para cada tamanho de grafo ($n_1 = n_2 =$ variando em 50, 75, 100, 125) em cada um dos testes propostos . A coluna esquerda representa as curvas ROC sob H_0 (quando os parâmetros θ_1 e θ_2 são iguais). A coluna da direita representa a curva ROC sob H_1 ($\theta_1 \neq \theta_2$).*

6.4 Simulação estratégia ANOVA com $m = 3$ grafos

Nesta simulação foram comparados três grafos gerados pelo mesmo modelo e mesmos parâmetros, uma condição na qual o teste baseado na variância (estratégia ANOVA, seção **??**) é útil, i. e., quando mais de dois grafos são comparados. O propósito desta simulação foi avaliar tanto a taxa de controle de falsos positivos quanto o poder do teste proposto. A configuração das simulações realizadas pode ser vista abaixo.

1. ER: $p_1 = p_2 = p_3 = 0,5$ (sob a hipótese nula) e $p_1 = p_3 = 0,5$ e $p2 = 0,52$ (sob a

hipótese alternativa). O número de vértices em cada grafo variou em: $n_1 = n_2 = n_3 = 50, 75, 100, 125$.

2. Geométrico: $r_1 = r_2 = r_3 = 0,3$ (sob a hipótese nula) e $r_1 = r_3 = 0,3$ e $r2 = 0,34$ (sob a hipótese alternativa). O número de vértices em cada grafo variou em: $n_1 = n_2 = n_3 = 50, 75, 100, 125$.

3. WS: $p_1^w = p_2^w = p_3^w = 0,3$ (sob a hipótese nula) e $p_1^w = p_3^w = 0,3$ e $p2^w = 0,35$ (sob a hipótese alternativa). O número de vértices em cada grafo variou em: $n_1 = n_2 = n_3 = 300, 400, 500, 600$.

4. BA: $p_1^s = p_2^s = p_3^s = 1,3$ (sob a hipótese nula) e $p_1^s = p_3^s = 1,3$ e $p2^s = 1,42$ (sob a hipótese alternativa). O número de vértices em cada grafo variou em: $n_1 = n_2 = n_3 = 700, 800, 900, 1000$.

6.4.1 Resultados

A figura 6.3 apresenta as curvas ROC da simulação da estratégia ANOVA com três grafos ($m = 3$) de cada um dos quatro modelos analisados (Erdös-Rényi, geometric, WattsStrogatz, and Barabási-Albert). O eixo x representa o limiar do p-valor, enquanto o eixo y representa a proporção de rejeições da hipótese nula. É possível reparar que, sob a hipótese nula (quando os três grafos g_1, g_2 e g_3 são gerados com os mesmos parâmetros), o teste proposto controla os erros do tipo I efetivamente. Com relação à hipótese alternativa (quando os grafosg_1, g_2 e g_3 têm pelo menos um dos grafos com parâmetro diferente), é possível observar que quanto maior for o grafo, maior será o poder do teste.

6.5 Simulação com grafos com número de vértices (n) diferentes

O propósito dessa simulação foi avaliar tanto a taxa de controle de falsos positivos e o poder do teste proposto em um conjunto de grafos com tamanho diferente, i.e diferentes números de vértices. Para isso, foram gerados três grafos do modelo ER (g_1, g_2 e g_3) com tamanhos $n_1 = n_3 = 125$ e n_2 variando em $n_2 = 50, 75, 100, 125$. Para simular grafos sob a hipótese nula, os parâmetros foram configurados com $p_1 = p_2 = p_3 = 0,5$. Para simular os grafos sob a hipótese alternativa, os parâmetros foram configurados em $p_1 = p_3 = 0,5$ e $p_2 = 0,52$.

6.5.1 Resultados

Na Figura 6.4 é possível observar as curvas ROC para o teste baseado no ANOVA com grafos desbalanceados (tamanhos diferentes). Em (A), o estimador utilizado foi baseado na divergência de KL. Em (B), foi utilizado o estimador baseado no estimador de máxima verossimilhança. Observando a coluna da esquerda é possível notar que o teste ANOVA para grafos, de fato controla a taxa de falsos positivos quando os grafos possuem tamanhos diferentes. A coluna da direita indica que quanto mais balanceado forem os grafos (quanto mais semelhantes os valores de n), maior será o poder do teste. Além disso, é possível perceber que os dois estimadores apresentam valores parecidos. Tal comparação só é possível para o modelo ER, uma vez que os outros grafos não possuem um estimador de máxima verossimilhança conhecido.

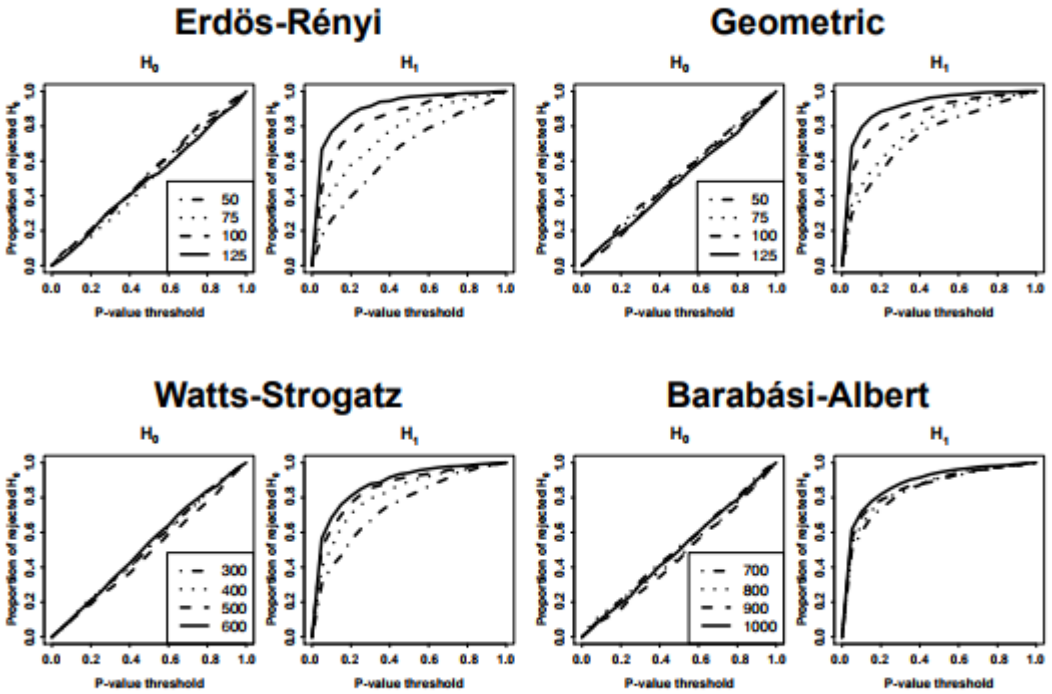


Figura 6.3: *Curvas ROC do teste ANOVA com $m = 3$ grafos para os modelos Erdös-Rényi, geometric, WattsStrogatz, and Barabási-Albert. O eixo x representa o limiar do p-valor, enquanto o eixo y representa a proporção de rejeições da hipótese nula. Traços sólidos representam as curvas ROC para grafos de tamanho diferente. A coluna esquerda representa as curvas ROC sob H_0 (quando os parâmetros θ_1, θ_2 e θ_3 são iguais). A coluna da direita representa a curva ROC sob H_1 ($\theta_1 = \theta_2 \neq \theta_3$). Quanto maior os grafos, maior é o poder do teste na discriminação de grafos gerados por parâmetros diferentes.*

6.6 Comparação entre os estimadores

Para o caso especial em que os grafos são gerados pelo modelo ER, existe um valor máximo (*maximum likelihood - ML*) para θ, que é simplesmente contar o número de arestas e dividir pelo número total de arestas possíveis (n 2). Então, nós comparamos a performance do teste proposto usando o estimador baseado na divergência de KL versus o classico estimador ML para $m = 2$ (dois grafos) e $m = 3$ (três grafos). A área abaixo da curva ROC representa o poder deste teste. Então, as cada área abaixo da curva e seus respectivos intervalos de confiança (CI) com 95% foram estimados e podem ser vistos na Figura 6.5.

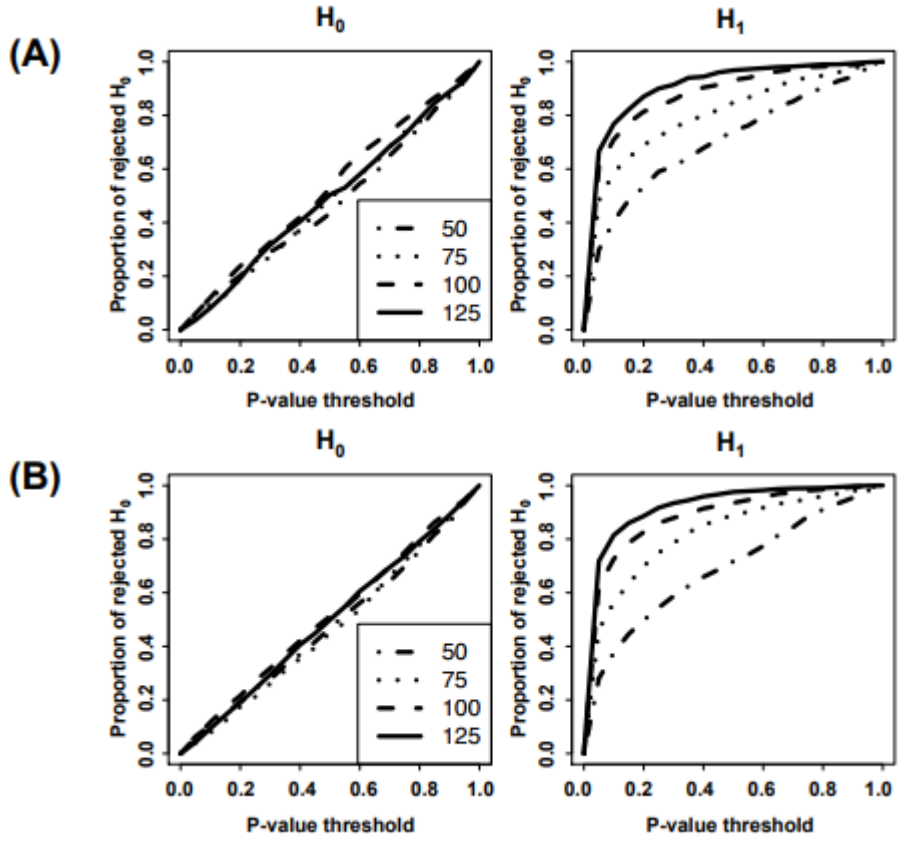


Figura 6.4: *Curvas ROC do teste ANOVA para grafos desbalanceados (diferentes números de vértices). O eixo x representa o limiar do p-valor, enquanto o eixo y representa a proporção de rejeições da hipótese nula. Os tipos diferentes de linha (sólido e tracejado) representam as curvas ROC obtidas para cada tamanho de grafo ($n_1 = n_3 = 125$ e n_2 variando em 50, 75, 100, 125). A coluna da esquerda representa as curvas ROC quando três grafos são gerados com os mesmos parâmetros $p_1 = p_2 = p_3 = 0,50$ (H_0). A coluna da direita representa as curvas ROC quando dois grafos são gerados com o mesmo parâmetro $p_1 = p_2 = 0,50$ e $p_3 = 0,52$ (H_1). (a) mostra curvas ROC obtidas usando o estimador baseado na divergência KL. (b) mostra as curvas ROC do ANOVA para grafos com o estimador baseado na máxima verossimilhança. Quanto mais próximo os tamanhos dos grafos, maior é o poder do teste.*

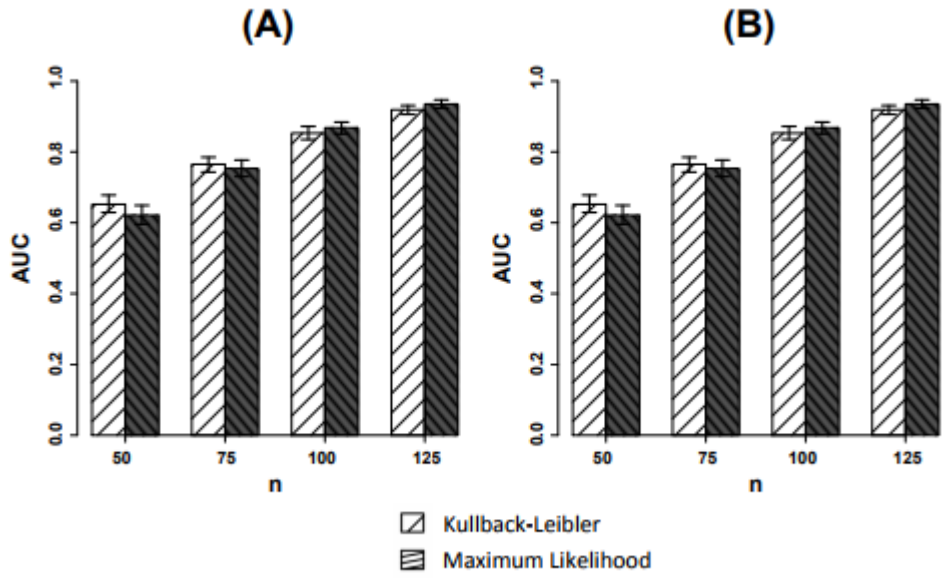


Figura 6.5: *Comparação da performance do teste com os diferentes estimadores para o modelo ER: Kullback-Leibler (KL) e o estimador baseado em máxima verossimilhança. As áreas abaixo da curva das curvas ROC e os intervalosde confiança* 95% *usando a divergência de KL ou o estimador baseado em máxima verossimilhança para os diferentes tamanhos de grafo (*$n050, 75, 100, 125$*). O modelo utilizado é o ER. O poder dos testes usando ambos estimadores é equivalente*

Capítulo 7

Aplicações

No capítulo 6 vimos que o teste proposto, ANOVA para grafos, apresenta bons resultados controlando as taxas de falsos positivos. Nas redes do mundo real, que possuem propriedades como comportamento de mundo pequeno e anexação preferencial, não faz sentido comparar dois grafos usando isomorfismo entre grafos, uma vez que o isomorfismo discriminará grafos que tenham tamanho diferente, ou que se diferem por um vértice. Nas aplicações reais é praticamente impossível encontrar grafos isomorfos e, portanto, para compará-las um teste como o ANOVA para grafos é de suma importância.

Neste capítulo serão apresentadas aplicações reais para o uso do teste proposto. Na seção 7.1 explicamos a aplicação em redes de interação proteína-proteína.

7.1 Redes de interação proteína-proteína

As redes biológicas podem ser representadas como grafos, nos quais os componentes (de uma célula, por exemplo) interagem com outros componentes através de interações par a par. Cada componente pode ser representado por vértices e suas interações são as arestas (Barabási e Oltvai, 2004). Dessa forma, é possível representar diversas redes biológicas, como redes que represesentam interações entre genes e fatores de transcrição que os regulam, redes metabólicas, redes de interação de proteínas.

Nas redes de interação proteína de proteínas, as proteínas interagem umas com as outras, de maneira específica. Entender como essas redes funcionam é um dos principais objetivos da genômica funcional para organismos plenamente sequenciados (**?**). As interações proteína-proteína desempenham um papel fundamental em muitos processos celulares, e no caso de distorção de interfaces de proteínas podem ocorrer o desenvolvimento de várias doenças. A figura 7.1 mostra o mapa de interação para Campylobacter jejuni.

Algumas redes biológicas podem ser redes direcionais, ou seja, a informação representada possui direção. No entanto, o teste proposto não considera grafos dirigidos. As redes de interação proteína-proteína são redes que representam as interações físicas entre as proteínas e não possuem direção. Muitas características descritas no capítulo 2 são observadas em redes biológicas e, particularmente nas redes de interação proteína-proteína. Algumas delas são descritas abaixo.

- Distribuição de probabilidade que obedece a lei de potência, que é uma característica das redes livres de escala vista na seção **??** (Barabási e Albert, 1999) e significa que um pequeno número de nós concentra boa parte das interações. Isso se deve à agregação preferencial (*Preferential attachment*), mecanismo que simula a preferência de novos vértices serem conectados a nós que já possuem muitas ligações. Nas redes de interação

Figura 7.1: *Mapa de interação proteína-proteína para Campylobacter jejuni. (Baudot et al., 2012)*

proteína-proteína já foram mapeados estes mecanismos, que estaria ligado ao fenômeno de duplicação gênica, o qual produz duas proteínas iguais inicialmente e que interagem com o mesmo vértice. Assim, cada proteína que está em contato com uma proteína duplicada recebe uma ligação extra, obtendo uma vantagem natural e tornando-se mais aptas a receber novas ligações (Pastor-Satorras *et al.*, 2003).

- O coeficiente de agrupamento (*Clustering Coefficient*), que é a capacidade de cada vértice ser agrupado. Esta característica representa o conceico de modularidade presente em redes biológicas, que significa que as funcionalidades celulares podem ser particionadas em uma coleção de módulos, na qual cada módulo é uma entidade discreta de muitos de muitos componentes elementares que realiza uma tarefa separável de outras funções de outros módulos (Ravasz *et al.*, 2002). (Barabási e Oltvai, 2004) definiu esta arquitetura hierárquica como uma rede em que nós são conectados de maneira esparsa mas pertencentes a áreas altamente agrupadas, com comunicação entre os grupos sendo mantidas pelos *hubs*, vértices com alto número de interações.

- Efeito de pequeno mundo (*Small-world effect*), no qual dois nós quaisquer podem ser conectados por um caminho curto (Barabási e Oltvai, 2004). Este efeito indica que perturbações locais podem alcançar níveis superiores rapidamente.

- Além disso, (Maslov e Sneppen, 2002) descreve uma propriedade sobre nós altamente

conectados, que evitam se ligarem uns aos outros (*disassortative*).

- Todas as características anteriores levam à capacidade que as redes biológicas celulares possuem para responder mudanças na organização interna enquanto mantém um comportamento relativamente normal. Essa característica recebe o nome de robustez e garante que se 80% dos vértices selecionados aleatoriamente falharem, os vértices restantes ainda formam um grupo compacto com caminhos conectado quaisquer dois nós (Barabási e Oltvai, 2004).

Diferentes técnicas experimentais têm sido desenvolvidas no entendimento de reconhecimento das interações entre proteínas. Algumas abordagens caracterizam as interações individuais das proteínas, enquanto outras visam o rastreamento de interações em escala genômica (**?**). Várias destas abordagens experimentais geraram grande quantidade de informações sobre interações proteína-proteína (**????**), o que não garante a qualidade quantidade de dados nos bancos de dados, nem baixos níveis de falsos positivos na predição de intereção de redes de proteína (**??**).

As aplicações do estudo das redes biológicas celulares são diversas, e vão desde o entendimento do seu funcionamento até a busca por alvos terapêuticos para drogas e vacinas. Por exemplo, **?** identificaram novos genes associados com risco maior de câncer de mama, (**?**) identificaram marcadores de mestátase de câncer de mama, através da extração de propriedades funcionais de proteínas diretamente do estudo topológico de redes de interação. Na parasitologia os estudos podem ajudar no desenvolvimento de vacinas (Andrés F Flórez e Muskus, 2010; LaCount DJ, 2005).

Os testes propostos neste trabalho, podem ser usados na comparação de redes de interação proteína-proteína, levando à generalização de resultados obtidos para alguma delas. Na próxima seção o teste realizado com os dados de algumas redes de interação de proteína serão apresentados.

7.2 Aplicação em redes de interação proteína-proteína

Para a aplicação do teste em dados reais, foram selecionadas redes de interação proteína-proteína. As cepas selecionadas pertencem à *Enterobacteriaceae* e são comumente associadas à ocorrência de gastrointerite. Os seis patógenos entéricos representativos selecionados foram:

1. Campylobacter jejuni NCTC11168
2. Escherichia coli O157:H7
3. Listeria monocytogenes EGDe
4. Salmonella enterica LT2
5. Shigella flexneri 2a str.301
6. Yersinia enterocolitica 8081

7.2.1 Dados utilizados

Os dados foram obtidos em *string-db.org* e estão organizados de maneira que os vértices representam as proteínas, enquanto as arestas representam as interações entre as proteínas. Neste trabalho foram consideradas apenas interações nas quais tenham sido observadas evidências experimentais, i. e., *score* > 0 e foram descartadas interações putativas. Na tabela 7.2 os dados foram resumidos em vértices e arestas por espécie.

Tabela 7.1: *Síntese do número de vértices (proteínas e interações para cada espécie.)*

Espécies (*strain*) significância	Vértices (n)	Arestas (interações)
C. jejuni (NCTC11168)	1398	16135
E. coli (O157:H7)	3295	50838
L. monocytogenes (EGDe)	1760	27180
S. enterica (LT2)	2908	42035
S. flexneri (2a str 301)	2980	38597
Y. enterocolitica (8081)	2592	34759

7.2.2 Resultados

Essas cepas estão largamente associadas na literatura ao modelo Barabási-Albert, portanto, nesta aplicação, assumimos este modelo e testamos o seu parâmetro p (expoente escalar). Pelos resultados a hipótese nula deve ser rejeitada, i. e., existe pelo menos uma rede de interação proteína-proteína entre as analisadas que é gerada por um parâmetro diferente ($p < 0,001$). Então, para identificar quais redes são diferentes, os testes foram realizados comparando as redes duas a duas.

Os p-valores foram corrigidos usando *False Discovery Rate* (FDR) (**?**) e estão resumidos na tabela 7.1. Com esses valores podemos concluir que não existem evidências estatísticas para discriminar as redes de E. coli e S. flexneri ($p = 0,924$). De fato, essas suas espécies são conhecidas por compartilhar características bioquímicas e fenotípicas e portanto, foram consideradas próximas geneticamente (**?**). Além dessas redes C. jejuni e S.enterica também não mostraram evidência estatística para afirmar que são diferentes ($p = 0,791$), que são encontradas comumente em aves domésticas (**?**).

Por outro lado, a rede de interação proteína-proteína de Y. enterocolitica parece ser estatísticamente distinta das redes C. jejuni, L. monocytogenes e S. enterica ($p < 0,001$). durante coinfecções patogênicas de Y. enterocolitica, S. enterica e L. monocytogenes, Y. enterocolitica inibe as outras duas nas células hospedeiras através da diminuição da indução de fagocitose para a absorção bacteriana, garantindo a sobrevivência bacteriana (**?**).

Tabela 7.2: *P-valores obtidos na aplicação das redes de interação proteína-proteína comparadas duas a duas. P-valores foram corrigidos usando a abordagem FDR. P-valores significativos no limiar de 5% estão em negrito)*

	E. Coli	L. monocytogenes	S. enterica	S. flexneri	Y. enterocolitica
C. jenjuni	$0,141$	$0,286$	$0,791$	$0,162$	$< 0,001$
E. coli		$0,319$	$0,174$	$0,924$	$0,064$
L. monocytogenes			$0,372$	$0,372$	$< 0,001$
S. enterica				$0,204$	$< 0,001$
S. flexneri					$0,064$

Capítulo 8

Conclusões

Neste capítulos serão feitas considerações baseadas nos resultados obtidos. O objetivo é avaliar a qualidade da estratégia proposta e enumerar trabalhos que podem ser feitos para melhorar e facilitar a realização de testes para grafos.

8.1 Considerações finais

Os testes propostos foram submetidos à várias simulações: entre os testes propostos, comparação do poder do teste para os diferentes modelos de grafos aleatórios, simulação com grafos desbalanceados, e o ANOVA para grafos com 3 grafos. As simulações realizadas apresentaram resultados satisfatórios, com os testes propostos controlando erros do tipo I com alto poder.

A abordagem proposta apresenta algumas limitações, dentre elas:

- Para os testes, assumimos que o modelo que gerou os grafos é conhecido. Entretanto, em aplicações reais, o modelo do grafo raramente é conhecido. Neste caso, nós propomos que primeiro seja identificado o modelo que melhor descreve os dados usando a abordagem de seleção de modelos descrito no capítulo 5. Se os modelos forem iguais, então use o nosso método proposto para testar os parâmetros. Caso contrário, os grafos são diferentes.

- A estimação do parâmetro baseado na divergência de KL pode ser usado na maioria dos grafos, entretanto aprsenta alto custo computacional (tanto no cálculo do espectro do grafo, como a aotimização do procedimento demandam tempo e memória). Além disso, tal abordagem só está definida para grafos direcionados. Portanto, desenvolver estimadores para grafos dirigidos e métodos hábeis para calcular em grandes redes em tempo e espaço viáveis levará à melhoria dos testes.

Para alguns grafos aleatórios específicos, como o k-regular (Nips 11), o teste proposto não é necessário. Um grafo k-regular é um grafo no qual todos os vértices tem o mesmo grau k, consequentemente, a variância dos graus é zero. Então, o teste consiste em determinar o grau de um vértice de cada grafo e diretamente, verificar a igualdade entre eles.

As abordagens descritas no capítulo 5 funcionam para famílias de grafos, mas, poucas são as aplicações reais em que é possível replicar os grafos disponíveis a fim de poder testá-los. Por exemplo, é encontramos redes de interação proteína-proteína, porém não é simples obter replicações destas redes. O mesmo é válido para redes sociais: podemos querer comparar se duas redes sociais são iguais, no entanto, não encontraremos um número satisfatório de replicações das redes.

Na aplicação apresentada no trabalho, os testes conseguiram discriminar redes que são conhecidamente concorrentes de acordo com a literatura. Além disso, não detectou evidência estatística para discriminar redes que são reconhecidamente semelhantes.

8.2 Trabalhos futuros

Como já foi dito na seção anterior, a principal limitação dos testes propostos está na estimação dos parâmetros. Portanto, desenvolver estimadores para grafos dirigidos e que sejam computacionalmente viáveis será de grande ajuda no teste de discriminação entre grafos.

Outra possibilidade é generalizar o teste proposto para grafos com vários parâmetros. Neste caso, um MANOVa (ANOVA multivariada) para grafos deve ser mais útil. Baseado nas ideias descritas aqui nós acreditamos que a generalização da ANOVA para grafos para uma versão multivariada é consequência direta.

Referências Bibliográficas

Andrés F Flórez e Muskus (2010) Jong Bhak Byoung-Chul Kim Allan Kuchinsky John H Morris Jairo Espinosa Andrés F Flórez, Daeui Park e Carlos Muskus. Protein network prediction and topological analysis in leishmania major as a tool for drug target selection. *BMC Bioinformatics*. Citado na pág. 2, 46

Bapat e Raghavan (1997) R. B. Bapat e T. E. S. Raghavan. *Nonnegative Matrices and Applications*. Encyclopedia of Mathematics and its Applications. Cambridge University Press. doi: 10.1017/CBO9780511529979. Citado na pág. 12

Barabási e Oltvai (2004) A.-L. Barabási e Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113. Citado na pág. 1, 44, 45, 46

Barabási e Albert (1999) Albert-László Barabási e Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512. Citado na pág. 1, 7, 8, 10, 36, 44

Baudot *et al.* (2012) Anaïs Baudot, Oussema Souiai e Christine Brun. Network analysis and protein function prediction with the prodistin web site. 804:313–26. Citado na pág. 45

Bollobás *et al.* (2001) Béla Bollobás, Oliver Riordan, Joel Spencer e Gábor Tusnády. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18:279–290. Citado na pág. 10

Borgatti *et al.* (2009) Stephen P Borgatti, Ajay Mehra, Daniel J Brass e Giuseppe Labianca. Network analysis in the social sciences. *science*, 323(5916):892–895. Citado na pág. 1, 2

Bradley (1997) Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159. Citado na pág. 36

Brouwer e Haemers (2012) Andries E. Brouwer e Willem H. Haemers. *Spectra of Graphs*. Springer-Verlag New York. Citado na pág. 12

Bullmore e Sporns (2009) Ed Bullmore e Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198. ISSN 1471-003X. doi: 10.1038/nrn2575. URL http://dx.doi.org/10.1038/nrn2575. Citado na pág. 1

Chung e Lu (2006) Fan Chung e Linyuan Lu. *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics)*. American Mathematical Society, Boston, MA, USA. ISBN 0821836579. Citado na pág. 1

Chung (1994) Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society. Citado na pág. 6

Cvetković *et al.* (1980) D.M. Cvetković, M. Doob e H. Sachs. *Spectra of graphs: theory and application*. Pure and applied mathematics. Academic Press. ISBN 9780121951504. Citado na pág. 11, 12

de Siqueira Santos *et al.* (2016) Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, João Ricardo Sato, Carlos Eduardo Ferreira e André Fujita. *Mathematical Foundations and Applications of Graph Entropy*, chapter Statistical methods in graphs: parameter estimation, model selection, and hypothesis test, páginas 98–120. John Q Public. Citado na pág. 9, 17, 21, 22, 33, 38

Efron (1979) B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26. ISSN 00905364. doi: 10.2307/2958830. URL http://dx.doi.org/10.2307/2958830. Citado na pág. 25

Efron e Tibshirani (1993) Bradley Efron e Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall. Citado na pág. 33

Egan (1975) J. P. Egan. *Signal detection theory and ROC analysis*. Series in Cognition and Perception. Academic Press, New York, NY. Citado na pág. 36

Erdös e Rényi (1959) Paul Erdös e Alfréd Rényi. On random graphs. *Publicationes Mathematicae (Debrecen)*, 6:290–297. Citado na pág. 1, 6, 7, 8, 36

Freeman (1978) Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, página 215. Citado na pág. 13

Fujita *et al.* (2017) Andre Fujita, Maciel Calebe Vidal e Daniel Yasumasa Takahashi. A statistical method to distinguish functional brain networks. *Frontiers in Neuroscience*. Citado na pág. 2, 3, 25, 28

Gastwirth (1971) Joseph L. Gastwirth. A general definition of the lorenz curve. *Econometrica*. Citado na pág. 36

Gilbert (1959) E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144. doi: 10.1214/aoms/1177706098. URL https://doi.org/10.1214/aoms/1177706098. Citado na pág. 1, 6

Green e Swets (1989) David Marvin Green e John A. Swets. *Signal Detection Theory and Psychophysics*. Robert E Kringer Publishing. Citado na pág. 36

Khinchin (1957) A. I. Khinchin. *Mathematical foundations of information theory*. Dover Publications. ISBN 0486604349. URL http://www.worldcat.org/isbn/0486604349. Citado na pág. 16

LaCount DJ (2005) Chettier R Phansalkar A-Bell R Hesselberth JR Schoenfeld LW Ota I Sahasrabudhe S Kurschner C Fields S Hughes RE. LaCount DJ, Vignali M. A protein interaction network of the malaria parasite plasmodium falciparum. *Nature*. Citado na pág. 2, 46

Lan V Zhang e Roth (2004) Oliver D King Lan V Zhang, Sharyl L Wong e Frederick P Roth. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*. Citado na pág. 2

Lovász (2007) László Lovász. Eigenvalues of graphs, November 2007. Lecture notes. Citado na pág. 12

Mani *et al.* (2008) Kartig M Mani, Celine Lefebvre, Kai Wang, wei Keat Lim, Katia Basso, Riccardo Dalla-Favera e Andrea Califano. A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas. *Molecular Systems Biology*. Citado na pág. 2

Maslov e Sneppen (2002) S. Maslov e K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910. Citado na pág. 45

Milgram (1963) Stanley Milgram. Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371–378. Citado na pág. 7

Mylne (2002) Kenneth R. Mylne. Decision-making from probability forecasts based on forecast value. *Metereological Applications*. Citado na pág. 36

Nadaraya (1964) E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142. doi: 10.1137/1109020. Citado na pág. 33

Pastor-Satorras *et al.* (2003) R Pastor-Satorras, E Smith e RV Solé. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*. Citado na pág. 45

Penrose (2003) M. Penrose. *Random Geometric Graphs*. Oxford studies in probability. Oxford University Press. ISBN 9780198506263. Citado na pág. 2, 8, 36

Ravasz *et al.* (2002) E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai e A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555. Citado na pág. 45

Rogers (2010) Timothy Rogers. *New results on the spectral density of random matrices*. Tese de Doutorado, King's College London. Citado na pág. 15

Sain e Scott (1996) Stephan R. Sain e David W. Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 91(436):1525–1534. ISSN 01621459. URL http://www.jstor.org/stable/2291578. Citado na pág. 31, 34

Scott (2012) J. Scott. *Social Network Analysis*. SAGE Publications. ISBN 9781446259450. Citado na pág. 1

Shannon (1948) C. E. Shannon. A Mathematical Theory of Communication. *Bell system technical journal*, 27. Citado na pág. 16, 26

Silva *et al.* (2004) A.C. Silva, P.C. Carvalho e M. Gattass. Diagnosis of solitary lung nodule using texture e geometry in computerized tomography images: Preliminary results. *IEEE Latin America Transactions*. Citado na pág. 36

Silverman (1986) Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London. Citado na pág. 31, 34

Spielman (2012) Daniel Spielman. *Combinatorial Scientific Computing*, chapter Spectral Graph Theory, páginas 495–517. Chapman and Hall. Citado na pág. 12

Sturges (1926) Herbert A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66. doi: 10.1080/01621459.1926.10502161. URL http://dx.doi.org/10.1080/01621459.1926.10502161. Citado na pág. 31, 34

Takahashi *et al.* **(2012)** Daniel Yasumasa Takahashi, João Ricardo Sato, Carlos Eduardo Ferreira e André Fujita. Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PLos One*, 7:e49949. Citado na pág. 2, 3, 15, 17, 19, 23, 25, 30, 33

Wasserman e Faust (1994) Stanley Wasserman e Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press. Citado na pág. 12

Watts e Strogatz (1998) Duncan J. Watts e Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442. Citado na pág. 1, 7, 8, 9, 13, 36